CS294 Markov Chain Monte Carlo: Foundations & Applications **Fall 2009**

Lecture 3: September 10

Lecturer: Prof. Alistair Sinclair *Scribes: Andrew H. Chan, Piyush Srivastava*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In the previous lecture we looked at some algorithmic applications of Markov Chains. In this lecture, we will study the Metropolis framework, which gives a recipe for designing Markov Chains that converge to a given stationary distribution, and look at its application to the problem of sampling from the Gibbs distribution on the Ising model. We will also look at another solution to the Ising model called the *Heat Bath Dynamics*, which is less general and works only for spin systems. We end today's lecture with an introduction to mixing times.

3.1 Examples of Markov Chains (continued)

3.1.1 The Metropolis Process

The Metropolis framework addresses the most general version of our sampling problem as follows: Given a set of configurations Ω and a strictly positive weight function $w : \Omega \mapsto \mathbb{R}^+$, we want to sample from Ω with respect to the distribution $\pi(x) = \frac{w(x)}{Z}$, where Z is a normalization constant given by $Z = \sum_{x\in\Omega} w(x)$. Specifically, we want to create a Markov chain whose stationary distribution is the distribution π. In a typical application, Ω may be a set of combinatorial structures, such as the set of satisfying assignments of a boolean formula.

In order to construct a Metropolis process in this setting, we require two ingredients:

Neighborhood Structure Our first requirement is a "neighborhood structure," which is a connected undirected graph with the elements of Ω as its vertices. Typically, two elements are connected by an edge iff they differ by some local change. (For example, in the case of a spin system, we may consider two configurations to be adjacent iff they differ in the spin at exactly one site. Note that we assume the graph is *undirected* and *connected* because the Markov chain must be both irreducible and reversible. Also, we allow self-loops, which may be needed to ensure aperiodicity.

We use the notation $x \sim y$ to denote that x and y are neighbors.

Proposal Distribution Let $\Gamma(\Omega, \mathcal{E})$ be the underlying graph. For each $x \in \Omega$ we define a "proposal distribution," which is a function $\kappa(x, \cdot)$ with the following properties:

- $\kappa(x, y) = \kappa(y, x) > 0$ for all x, y such that $x \sim y, x \neq y$.
- For all $x \in \Omega$, $\sum_{y\neq x} \kappa(x, y) \leq 1$ and $\kappa(x, x) = 1 - \sum_y \kappa(x, y)$.
- For notational convenience, we define $\kappa(x, y) = 0$ if x and y are not neighbors.

The transitions of the Markov chain are now specified as follows. From any state $x \in \Omega$:

- Pick a neighbor $y \sim x$ with probability $\kappa(x, y)$.
- "Accept" the move to y with probability $\min\left\{1, \frac{w(y)}{w(x)}\right\} = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$, else stay at x.

> The reason for the term “proposal distribution” for κ is now clear. We “propose” a new vertex according to the distribution induced by κ on the current vertex, and then move to this vertex with a probability that depends on the stationary distribution to which we want the process to converge. Notice that the actual transition probabilities are $P(x,y) = \kappa(x,y) \cdot \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$. Crucially, though, we can implement this knowing only the weights w (i.e., we don’t need to know the normalizing factor Z, which cancels out in the ratio $\pi(x)/\pi(y)$).

Note that the Metropolis process is always irreducible, as the neighborhood structure is connected and can be made aperiodic by the usual trick of introducing self-loops.

We now show the reversibility of the Metropolis process with respect to π. From the Fundamental Theorem, this implies that π is its unique stationary distribution.

Claim 3.1 *The Metropolis Markov chain defined above is reversible with respect to* $\pi(x) = \frac{w(x)}{Z}$.

Proof: We need to check that $\pi(x)P(x,y) = \pi(y)P(y,x)$ for each x and y. Assume without loss of generality that $w(y) \leq w(x)$. Then

$$\begin{aligned} \pi(x)P(x,y) &= \frac{w(x)}{Z}\kappa(x,y)\frac{w(y)}{w(x)} \\ &= \frac{w(y)}{Z}\kappa(y,x) \\ &= \pi(y)P(y,x), \end{aligned}$$

where the second equality follows from the symmetry of κ. ■

As a technical point, we note that the condition $\kappa(x,y) = \kappa(y,x)$ is not necessary. If κ is not symmetric, we can simply change the acceptance probability to $\min\left\{1, \frac{w(y)\kappa(y,x)}{w(x)\kappa(x,y)}\right\}$. It is easy to verify (**exercise!**) that with this definition reversibility w.r.t. π still holds.

3.1.2 The Ising Model

As an example of the above formalism, we will now consider the Ising model. We recall that the model consists of a graph $G = (V,E)$, as shown in Figure 3.1, where each of the vertices corresponds to a *site* and can be in one of the states $\{+,-\}$. Denoting the set of sites by V, the set of configurations Ω is therefore $\{+,-\}^V$. For a configuration $x \in \Omega$, we denote by $a(x)$ the number of edges in E whose endpoints have agreeing spins, and by $d(x)$ the number of edges whose endpoints have disagreeing spins. Also, for a site $s \in V$, we denote its spin in configuration $x \in \Omega$ by $x(s)$. As discussed in Lecture 1, the Gibbs distribution for the Ising model (with ferromagnetic interaction) is given by

$$\begin{aligned} \pi(x) &\propto \exp\left\{\beta \sum_{\{s,t\}\in E} x(s)x(t)\right\} \\ &= \exp\{\beta(a(x) - d(x))\} \\ &\propto \exp\{2\beta a(x)\} \\ &= \lambda^{a(x)}, \end{aligned}$$

where $\lambda = \exp(2\beta)$. The second proportionality here follows from the facts that for any $x \in \Omega$, $a(x) + d(x) = 2|E|$, and that the number of edges $|E|$ is fixed. The parameter β is the inverse temperature. In the case of ferromagnetic interaction, the lower energy configurations are those in which the neighboring spins are fully aligned, and hence the distributions with more spins aligned have higher weight in the Gibbs distribution. We now show how the Metropolis process can be used to sample from the Gibbs distribution.

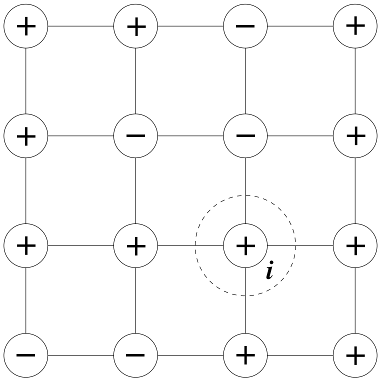

Figure 3.1: Example configuration of a 2-dimensional ferromagnetic Ising model. The dotted circle indicates the vertex i.

We must first define the neighborhood structure on $\Omega = \{+,-\}^V$. A natural choice is to let $x \sim y$ if and only if x and y differ in exactly one spin. Obviously this neighborhood structure is connected. For the proposal distribution, a natural choice is $\kappa(x,y) = \frac{1}{n}$ for all $x \sim y$, where $n = |V|$, the number of sites. Thus, at each step of the Markov chain, we pick a site uniformly at random and "flip" its spin with the appropriate probability given by the Metropolis rule.

As an example, the circled vertex i in Figure 3.1 has three agreeing $+$ neighbors and one disagreeing $-$ neighbor. Let the initial configuration be denoted by x and the configuration resulting from flipping i to $-$ be denoted by y. Then we flip i with probability $\min\left\{1, \frac{w(y)}{w(x)}\right\} = \min\left\{1, \lambda^{a(y)-a(x)}\right\} = \min\left\{1, \frac{1}{\lambda^2}\right\} = \frac{1}{\lambda^2}$, as $\lambda > 1$ for the ferromagnetic model.

3.1.3 The Heat-Bath

Above we saw how the Metropolis process can be used to construct a reversible Markov chain for the ferromagnetic Ising model, or indeed for any spin system. We now mention an alternative approach for spin systems, known as the *heat-bath dynamics*. However, whereas the Metropolis process can be used for an arbitrary probability distribution π, the heat-bath dynamics is suitable only for spin systems.

As an example, let us again consider the ferromagnetic Ising model. The transitions of the heat-bath Markov chain from a state $x \in \Omega$ are as follows:

- Pick a site i uniformly at random.
- Replace the spin $x(i)$ at i by a spin chosen according to the distribution π, conditioned on the spins at the neighbors of i.

So, if d^+ neighbors of i have spin $+$ and d^- have spin $-$, then the conditional probabilities are given by $\Pr_\pi(x(i) = +|d^+, d^-) = \frac{\lambda^{d^+}}{\lambda^{d^+}+\lambda^{d^-}}$ and $\Pr_\pi(x(i) = -|d^+, d^-) = \frac{\lambda^{d^-}}{\lambda^{d^+}+\lambda^{d^-}}$. Returning to Figure 3.1, if vertex i is chosen then its new spin is set to $-$ with probability $\frac{\lambda}{\lambda^3+\lambda}$ and to $+$ with probability $\frac{\lambda^3}{\lambda^3+\lambda}$.

It is an easy **exercise** to check that the heat-bath Markov Chain is aperiodic (because of the presence of self-loops), irreducible (all possible configurations are connected by single spin-flips), and reversible w.r.t. π. This ensures convergence to the Gibbs distribution, as required.

Both the heat-bath dynamics and the Metropolis process for spin systems are special cases of the general framework known as *Glauber Dynamics*. In this framework, the graph G can be arbitrary and a configuration is an assignment of spins from some set $\{1, 2, \ldots, q\}$ to the vertices. In each transition, a vertex is chosen uniformly at random, and its spin is updated according to a local distribution, which depends only on the spins at the vertex and its neighbors. (We will discuss spin systems in more detail later in the course.)

3.2 The Fundamental Theorem and Mixing Time

In this section we present a partial proof of the Fundamental Theorem of Markov Chains. The proof will proceed via estimates of mixing times. We first restate the Fundamental Theorem.

Theorem 3.2 *Let P be the transition matrix of an irreducible and aperiodic Markov chain on a finite set Ω. Then there exists a unique probability distribution π over Ω such that $\pi(x) > 0$ for all $x \in \Omega$, $\pi P = \pi$, and for any initial state $x \in \Omega$*

$$\lim_{t\to\infty} p_x^{(t)} = \pi.$$

Strictly speaking, we will *not* be proving the above theorem completely. We assume that there exists a distribution π such that $\pi P = \pi$ (that is, we assume a stationary distribution exists), and then show that the rest of the theorem holds. We first note that irreducibility of the Markov chain implies that $\pi(x) > 0$ for every x in Ω. We also note that *uniqueness* of π follows from the fact that π satisfies $\pi P = \pi$ and $\lim_{t\to\infty} p_x^{(t)} = \pi$ for every x.

We begin by defining a notion of distance between probability distributions.

Definition 3.3 *For two probability distributions μ and ν on Ω, the* total variation distance *is*

$$\|\mu - \eta\| \equiv \frac{1}{2} \sum_{x\in\Omega} |\mu(x) - \eta(x)| = \max_{A\subseteq\Omega} |\mu(A) - \eta(A)|\,.$$

We note that this is just the ℓ_1 metric, scaled by a factor of $\frac{1}{2}$. This scaling ensures that $\|\mu - \eta\|$ lies in $[0, 1]$.

Exercise: Verify that $\frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \eta(x)| = \max_{A\subseteq\Omega} |\mu(A) - \eta(A)|$.

As an aside, we note that the total variation distance considers the difference between probabilities of all possible events, and hence can be affected severely even by apparently minor changes in the distribution. Let us consider the following example:

Example: Consider n cards. Let μ be the uniform distribution over all permutations of the n cards, and η be the same distribution except that the bottom card is fixed. Then $\|\mu - \eta\| = 1 - \frac{1}{n}$ because the probability of the bottom card of the first deck being the same as the fixed card of the second deck is $\frac{1}{n}$, but for the second deck it is 1.

We will show that the total variation distance between π and $p_x^{(t)}$ decreases exponentially with t/τ_{mix}, where τ_{mix} is a parameter depending on the Markov chain. This gives us an estimate on the time required to get within a given distance of π.

The proof uses the method of *coupled distributions.* We first define the coupling of two distributions and then state the "Coupling Lemma."

Definition 3.4 (Coupling). *Let μ and η be any two probability distributions over Ω. A probability distribution ω over $\Omega \times \Omega$ is said to be a coupling of μ and η if its marginals are μ and η; that is,*

$$\begin{aligned}\mu(x) &= \sum_{y\in\Omega} \omega(x,y),\\ \eta(x) &= \sum_{y\in\Omega} \omega(y,x).\end{aligned}$$

Lemma 3.5 (Coupling Lemma). *Let μ and η be probability distributions on Ω, and let X and Y be random variables with distributions μ and η, respectively. Then*

1. $\Pr[X \neq Y] \geq \|\mu - \eta\|$.
2. *There exists a coupling of (μ, η) such that* $\Pr[X \neq Y] = \|\mu - \eta\|$.

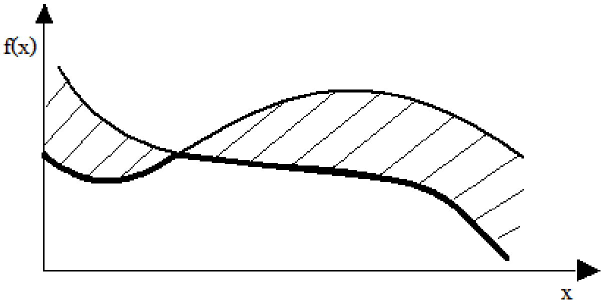

Figure 3.2: Two probability distributions. The bold line indicates the "lower envelope" of the two distributions.

Proof: [Informal sketch] Consider the two probability distributions shown in Figure 3.2. Suppose we try to construct a joint distribution for the two of them that maximizes the probability that they are equal. Clearly, the best we can do is to make $X = Y = z$ with probability $\min\{\Pr(X = z), \Pr(Y = z)\}$ for each value $z \in \Omega$. This is indicated by the bold line in the figure (the "lower envelope" of the two distributions). In this case, the probability that $X \neq Y$ is given by half the area of the shaded region. We then have

$$\frac{1}{2}(\text{Area of Shaded Region}) = \Pr[X \neq Y] = \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \eta(x)| = \|\mu - \eta\|.$$

■

In order to discuss the convergence of $p_x^{(t)}$ to π, we will need the following definitions:

Definition 3.6

1. *For any $x \in \Omega$, we define $\Delta_x(t) = \|p_x^{(t)} - \pi\|$.*
2. $\Delta(t) = \max_{x\in\Omega} \Delta_x(t)$ *is the maximum possible distance from π after t time steps.*
3. $\tau_x(\epsilon) = \min\{t : \Delta_x(t) \leq \epsilon\}$ *is the first time step t at which the distance $\|p_x^{(t)} - \pi\|$ drops to ϵ.*
4. $\tau(\epsilon) = \max_{x\in\Omega} \tau_x(\epsilon)$.

We can now give the key definition of the *mixing time*:

Definition 3.7 (Mixing Time). *The mixing time τ_{mix} of a Markov chain is $\tau(1/2e)$.*

In other words, the mixing time is the time until the variation distance, starting from the worst possible initial state $x \in \Omega$, reaches $1/2e$. This value is chosen for algebraic convenience only, as we shall see below. We now prove some basic facts about the time-dependent behavior of the Markov chain, which will also justify our definition of mixing time. The first fact says that $\Delta(t)$ is non-increasing, so that once the variation distance reaches $1/2e$, it never exceeds it again.

Claim 3.8 *$\Delta_x(t)$ is non-increasing in t.*

Proof: Let $X_0 = x$ and Y_0 have the stationary distribution π. We fix t and couple the distributions of the random variables X_t and Y_t such that $\Pr[X_t \neq Y_t] = \|p_x^{(t)} - \pi\| = \Delta_x(t)$, which is possible because of the Coupling Lemma. We now use this coupling to define a coupling of the distributions of X_{t+1} and Y_{t+1} as follows:

- If $X_t = Y_t$, then set $X_{t+1} = Y_{t+1}$,
- Otherwise, let $X_t \to X_{t+1}$ and $Y_t \to Y_{t+1}$ independently.

Then we have

$$\Delta_x(t+1) \equiv \|p_x^{(t+1)} - \pi\| \leq \Pr[X_{t+1} \neq Y_{t+1}] \leq \Pr[X_t \neq Y_t] = \Delta_x(t).$$

The first inequality holds because of the Coupling Lemma, and the second inequality is true by the construction of the coupling. ■

We now define more general quantities which capture the evolution of distance between corresponding distributions for arbitrary initial configurations.

Definition 3.9

1. *$D_{xy}(t) = \|p_x^{(t)} - p_y^{(t)}\|$.*
2. *$D(t) = \max_{x,y\in\Omega} D_{xy}(t)$.*

The following simple relationship between $D(t)$ and $\Delta(t)$ is left as an **exercise**:

Claim 3.10 *$\Delta(t) \leq D(t) \leq 2\Delta(t)$.*

We now prove that the maximum variation distance, $\Delta(t)$, decays exponentially with time constant τ_{mix}. This will follow from the fact that $D(t)$ is submultiplicative.

Claim 3.11 *$\Delta(t) \leq \exp\left(-\left\lfloor \frac{t}{\tau_{\text{mix}}} \right\rfloor\right)$.*

Proof: Let $X_0 = x$ and $Y_0 = y$. We use the Coupling Lemma to couple the distributions of X_t and Y_t so that

$$D_{xy}(t) \equiv \|p_x^{(t)} - p_y^{(t)}\| = \Pr[X_t \neq Y_t].$$

We then construct a coupling of X_{t+s} and Y_{t+s} as follows:

- If $X_t = Y_t$ then set $X_{t+i} = Y_{t+i}$ for $i = 1, 2, \ldots s$,
- Otherwise, let $X_t = x'$ and $Y_t = y' \neq x'$. Use the Coupling Lemma to couple the distributions of X_{t+s} and Y_{t+s}, conditioned on $X_t = x'$ and $Y_t = y'$, such that

$$\Pr[X_{t+s} \neq Y_{t+s} | X_t = x', Y_t = y'] = \|p_{x'}^{(s)} - p_{y'}^{(s)}\| = D_{x'y'}(s) \leq D(s). \tag{3.1}$$

The last inequality holds by the definition of $D(t)$. We now have

$$\begin{aligned} D_{xy}(t+s) &= \|p_x^{(t+s)} - p_y^{(t+s)}\| \\ &\leq \Pr[X_{t+s} \neq Y_{t+s}], \text{ by the Coupling Lemma} \\ &\leq D(s)D_{xy}(t), \text{ by the construction of the coupling} \\ &\leq D(s)D(t). \end{aligned}$$

Since this holds for all x, y, we get that $D(t+s) \leq D(s)D(t)$. It follows that $D(kt) \leq D(t)^k$ for all positive integers k. Consequently,

$$\Delta(k\tau_{\text{mix}}) \leq D(k\tau_{\text{mix}}) \leq D(\tau_{\text{mix}})^k \leq (2\Delta(\tau_{\text{mix}}))^k \leq e^{-k}.$$

The last inequality follows from the definition of τ_{mix}, and proves the claim. (It is in the last step that we need $\Delta(\tau_{\text{mix}})$ to be *strictly* less than $\frac{1}{2}$; our choice of $\Delta(\tau_{\text{mix}}) = \frac{1}{2e}$ satisfies this and leads to a particularly simple expression for $\tau(\epsilon)$.) ■

Corollary 3.12 $\tau(\epsilon) \leq \tau_{\text{mix}} \left\lceil \log\left(\epsilon^{-1}\right) \right\rceil$.

The corollary follows immediately from Claim 3.11 and justifies our definition of mixing time: the cost of obtaining any desired variation distance ϵ is only a modest factor times τ_{mix}.

Exercise: The above arguments almost constitute an elementary probabilistic proof of the fundamental theorem. The only hole is that we assumed that the stationary distribution π exists. Show that π exists when P is irreducible and aperiodic. [Hint: Let $x \in \Omega$ be arbitrary. Define $q_x(x) = 1$ and $q_x(y)$ to be the expected number of times that the Markov chain, started in state x, visits y before returning to x. Show that $\pi(y) \propto q_x(y)$ is stationary.] Then assemble all the pieces into a proof of the fundamental theorem.

References

[MRRTT53] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Phys*, **21** (1953), pp. 1087–1092.