Georgia Tech **Fall 2006**

Markov Chain Monte Carlo Methods

Lecture 7: September 12, 2006

Coupling from the Past

Eric Vigoda

7.1 Coupling from the Past

7.1.1 Introduction

We saw in the last lecture how Markov chains can be useful algorithmically. If we have a probability distribution we'd like to generate random samples from, we design an ergodic Markov chain whose unique stationary distribution is the desired distribution. We then run the chain (i.e., start at an arbitrary state and evolve according to the transition matrix), until the process is at (or close to) the stationary distribution.

In the last lecture we saw that the chain eventually reaches the stationary distribution. In order for this Markov chain approach to be efficient we need to know how fast we converge to stationarity. Most of the class will focus on methods for bounding the convergence rate. However there are many chains which we believe converge quickly, but no one has proved it. Today we'll look at a method which "notices" when we reach the stationary distribution. The resulting samples are guaranteed to be from the stationary distribution, and in many cases the algorithm is extremely efficient. The method is known as *Coupling from the Past*, and was introduced by Propp and Wilson in 1996 [1].

Let $\mathcal{M}$ be an ergodic Markov chain defined on a space Ω and transition matrix P. Thus, for $x, y \in \Omega$ and integer $t > 0$, $P(x, y)$ is the probability of going from x to y in t steps. Let π be the stationary distribution of $\mathcal{M}$.

We will use the three state Markov chain in Figure 3.1 as a running example. Consider the following experiment. Create three copies of the Markov chain, each with a distinct starting state A, B, and C. Define a "global" coupling for the three chains. More precisely, a global coupling is a joint evolution for all three chains, such that each chain viewed in isolation is a faithful copy of the chain of interest. Now, as soon as all three chains reach the same state (i.e., they've all coupled), output the resulting state. Is the output from the stationary distribution?

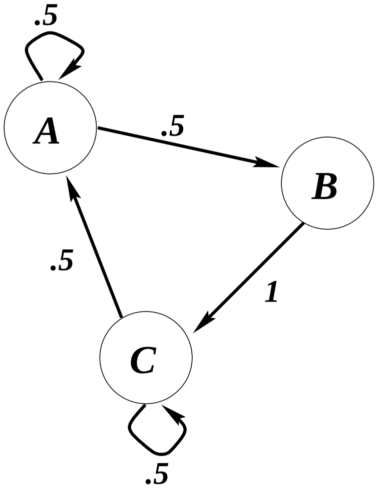

Figure 7.1: A simple three state Markov chain

Our intuition from last class is that once we've coupled from all pairs of initial states, we've reached stationarity. Thus, we might expect that the output is from stationarity. However, this is false. In our example the output will never be state B. This because B is never the first state where all the chains couple, since B has a unique predecessor they would have coupled a step earlier at state A. But B has some positive probability in the stationary distribution, so the output is clearly not from the stationary distribution. However, it turns out that if we run this experiment "backwards", it works. It's a very clever idea. Let's formalize the notion of a global coupling before detailing the general algorithm and analyzing it.

7.1.2 Formal Description

Let $f : \Omega \times [0,1] \rightarrow \Omega$ be a "global" coupling, i.e., for all $x \in \Omega$ and r chosen uniformly at random from $[0,1]$

$$Pr(f(x,r) = y) = P(x,y). \tag{7.1}$$

Thus, the random seed r defines the evolution from all states. In words, from $X_t \in \Omega$, the chain evolves by choosing r uniformly at random from the interval $[0,1]$ and then moving to state $f(X_t, r)$.

Let $r_t,\ t \in \mathbb{Z}$ be independent, each chosen uniformly at random from $[0,1]$. Let $f_t : \Omega \rightarrow \Omega$ be the function $f(\cdot, r_t)$ (i.e. $f_t(x) = f(x, r_t)$). For a chain (X_t), the function f_t will define the evolution at time t, i.e., $X_{t+1} = f(X_t)$. Let $t_1 < t_2$ be integers. Let

$$F_{t_1}^{t_2}(x) = (f_{t_2-1} \circ f_{t_2-2} \circ \cdots \circ f_{t_1})(x) = f_{t_2-1}(f_{t_2-2}(\ldots f_{t_1}(x) \ldots)).$$

Thus, $F_{t_1}^{t_2}$ defines the evolution from times $t_1 \rightarrow t_2$. From (7.1) it follows that for every

$x, y \in \Omega$

$$Pr(F_{t_1}^{t_2}(x) = y) = P^{t_2 - t_1}(x, y).$$

The earlier (wrong!) algorithm (which simulates the chains forward in time) can now be formalized as follows. We will run $|\Omega|$ chains X^σ, $\sigma \in \Omega$ simultaneously. The chain X^σ will have starting state σ, i.e., $X_0^\sigma = \sigma$. We will run all these chains simultaneously, using the coupling f, and outputting the final state when they coalesced. More formally, let T be the first time when $|F_0^T(\Omega)| = 1$. (Observe that F_0^T is a constant function, i.e., $|F_0^T(\Omega)| = 1$, if all the chains have reached the same state.) Then the output value is the unique element of the set $F_0^T(\Omega)$. We asked the following question - will the output value be from the stationary distribution?

Our Markov chain in Figure 3.1 shows that the output value does not have to be from the stationary distribution. Now we can formalize the strange question - what would happen if we "run the chain from the past"? More formally, let M be the first time when $|F_{-M}^0(\Omega)| = 1$. Output the unique element of $F_{-M}^0(\Omega)$. What is the distribution of the output?

Theorem 7.1 ([1]) *$F_{-M}^0(\Omega)$ has the same distribution as π.*

Proof: For any fixed $t > 0$, note that, for all $x, y \in \Omega$,

$$\Pr\left(\ F_0^t(x) = y\ \right) = \Pr\left(\ F_{-t}^0(x) = y\ \right).$$

The probability space for the LHS is over the choices $r_1, \ldots, r_t$, whereas for the RHS it's over $r_{-t+1}, \ldots, r_{-1}, r_0$. Thus, regardless of which order we construct these t random seeds, the distributions of F_0^t and F_{-t}^0 are the same.

Therefore, if we run the simulation infinitely far in the past we reach stationarity. More formally, for all $x, y \in \Omega$,

$$\begin{aligned} \lim_{t\to\infty} \Pr\left(\ F_{-t}^0(x) = y\ \right) &= \lim_{t\to\infty} \Pr\left(\ F_0^t(x) = y\ \right) \\ &= \pi(y). \end{aligned}$$

For $t = t_1 + t_2$, observe

$$F_{-t}^0 = F_{-t_2}^{-t_1-1} \circ F_{-t_1}^0$$

Therefore, if F_{-M}^0 is a constant function, then for all $t > M$, all $x \in \Omega$,

$$F_{-t}^0(x) = (F_{-t}^{-M-1} \circ F_{-M}^0)(x) = F_{-M}^0(x).$$

Outputting $F_{-M}^0(x)$ is the same as outputting $F_{-\infty}^0(x)$. This proves the theorem. ■

Note, the above proof gets at the heart of the difference for going backwards versus going forward. Recall T is the first time t such that F_0^t is a constant function. Note for $t > T$ we know F_0^t is a constant function, but it is not necessarily true that $F_0^t(x) = F_0^T(x)$. Hence, after time T all of the couplings converge, but where they converge might change as the time increases. Whereas, we saw in the last proof, for $t > M$ we have $F_{-t}^0(x) = F_{-M}^0(x)$, thus they continue to converge upon the same state.

7.1.3 Implementation

Theorem 7.1 translates to an algorithm for *perfect* sampling from the stationary distribution of the Markov chain $\mathcal{M}$. However there are several issues we must deal with before we get an efficient sampling algorithm

- Is it possible to implement the algorithm efficiently? More precisely, can we decide $|F^0_{-t}(\Omega)| = 1$ in polynomial time even though the sample space of the Markov chain is exponentially large?
- How do we pick the coupling f? We would like to make $E[M]$ small. Note that for a bad choice of f it can even happen that $P(M = \infty) > 0$.

It turns out that for *monotone* systems we can tackle both of the above problems. We will consider the following example which shows that in the case of the Ising model we can use the natural partial ordering of the model to decide coalescence efficiently by computing F^0_{-t} for just two elements $\top, \bot \in \Omega$. Moreover the expected running time $\exp M$ will be $O(T_{\text{mix}}(1/4) \ln |V|)$. Recall, $T_{\text{mix}}(1/4)$ is the time (from the worst initial state) until the chain is within variation distance at most 1/4 from the stationary distribution.

7.1.4 The Ising Model

Consider the Ising model on an undirected graph $G = (V, E)$. Let $\Omega = \{+1, -1\}^V$. For $\sigma \in \Omega$, recall the Hamiltonian is defined as

$$H(\sigma) = \sum_{(u,v)\in E} \mathbf{1}(\sigma(u) \neq \sigma(v)) = \frac{1}{2} \sum_{(u,v)\in E} (1 - \sigma(u)\sigma(v)).$$

At inverse temperature $\beta > 0$, the weight of a configuration is then

$$w(\sigma) = \exp(-\beta H(\sigma)).$$

The Gibbs distribution is

$$\mu(\sigma) = \frac{w(\sigma)}{Z_G(\beta)},$$

where $Z_G(\beta)$ is the partition function (or normalizing constant) defined as

$$Z_G(\beta) = \sum_{\sigma\in\Omega} w(\sigma).$$

To sample from the Gibbs distribution we will use a simple single-site Markov chain known as the Glauber dynamics (with Metropolis filter). The transitions $X_t \to X_{t+1}$ are defined as follows. From $X_t \in \Omega$,

- Choose $v \in_R V$, $s \in_R \{+1, -1\}$ and $r \in_R [0, 1]$.
- Let $X^{'}(v) = s$ and $X^{'}(w) = X_t(w),\ w \neq v$.
- Set

$$X_{t+1} = \begin{cases} X^{'} & \text{if } r \leq \min\{1, w(X')/w(X_t)\} = \min\{1, e^{-\beta H(X^{'})}/e^{-\beta H(X_t)}\} \\ X_t & \text{otherwise} \end{cases}$$

There is a natural partial order on Ω where $\sigma_1 \preceq \sigma_2$ iff $\sigma_1(v) \leq \sigma_2(v)$ for all $v \in V$. Let $\perp, \top$ be the minimum and maximum elements of $\preceq$.

Let f be the coupling in which v, s, r are the same for all chains. Then $X_t \preceq Y_t$ implies $X_{t+1} \preceq Y_{t+1}$, i.e., the coupling preserves the ordering. To see this note that

$$e^{-\beta H(X^{'})}/e^{-\beta H(X_t)} = \exp\left(\beta(s - X^t(v)) \sum_{\{u,v\} \in E} X^t(u)\right).$$

Thus for $s = -1$ if chain Y_t makes the move then X_t also does, hence $X_{t+1} \preceq Y_{t+1}$. Similarly for $s = +1$.

Since f preserves monotonicity,

$$|F_{t_1}^{t_2}(\Omega)| = 1 \iff F_{t_1}^{t_2}(\perp) = F_{t_1}^{t_2}(\top).$$

The coupling time of f is the smallest T such that $F_0^T(\perp) = F_0^T(\top)$. The coupling from the past time of f is the smallest M such that $F_{-M}^0(\perp) = F_{-M}^0(\top)$. We have that

$$\Pr(\,T > t\,) = \Pr\left(\,F_0^t(\perp) \neq F_0^t(\top)\,\right) = \Pr\left(\,F_{-t}^0(\perp) \neq F_{-t}^0(\top)\,\right) = \Pr(\,M > t\,).$$

Hence M has the same distribution as T.

Let

$$\overline{d}(t) = d_{TV}(P^t(\perp, \cdot), P^t(\top, \cdot)).$$

Note,

$$\overline{d}(t) \leq 2 \max_{z \in \Omega} d_{TV}(P^t(z, \cdot), \pi).$$

Thus, if we show $T_{\text{mix}}(\epsilon) \leq t$, then $\overline{d}(t) \leq 2\epsilon$.

Lemma 7.2 $\Pr(\,M > t\,) = \Pr(\,T > t\,) \leq \overline{d}(t)|V|$.

Proof: For $z \in \Omega$ let $h(z)$ be the length of the longest decreasing chain in $\preceq$ starting with z. If $X_t > Y_t$, then $h(X_t) \geq h(Y_t) + 1$.

Now let $X_0 = \perp, Y_0 = \top$. Then

$$\begin{aligned} \Pr(\ T > t\ ) = \Pr(\ X_t \neq Y_t\ ) &\leq \mathrm{E}(\ h(Y_t) - h(X_t)\ ) \\ &= \mathrm{E}(\ h(Y_t)) - E(h(X_t)\ ) \\ &\leq |V| d_{TV}(P^t(\perp, \cdot), P^t(\top, \cdot)) \\ &= |V| \overline{d}(t). \end{aligned}$$

as desired. ■

Lemma 7.3 $\mathrm{E}(\ M\ ) = \mathrm{E}(\ T\ ) = 2T_{\mathrm{mix}}(1/4) \ln(4|V|)$.

Proof: Let $t = T_{\mathrm{mix}}(1/4) \log(4|V|)$. Note, $T_{\mathrm{mix}}(1/4|V|) \leq t$. Hence, $\overline{d}(t) \leq 1/2|V|$. Therefore, we have $\Pr(\ T > t\ ) \leq 1/2$ and hence $\Pr(\ T > kt\ ) \leq 1/2^k$. Thus

$$\mathrm{E}(\ T\ ) = \sum_{\ell \geq 0} \Pr(\ T \geq \ell\ ) \leq \sum_{k \geq 0} t \Pr(\ T \geq kt\ ) \leq t \sum_{k \geq 0} \frac{1}{2^k} \leq 2t.$$

■

References

[1] James Gary Propp, David Bruce Wilson *Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics,* Random Structures Algorithms 9 (1996), no. 1-2, 223–252.