CS880: Approximations Algorithms

Scribe: Dave Andrzejewski **Lecturer:** Shuchi Chawla
Topic: Approx counting/sampling, MCMC methods **Date:** 4/24/07

The previous lecture showed that, for self-reducible problems, the problem of estimating the size of the set of feasible solutions is equivalent to the problem of sampling nearly uniformly from that set. This lecture explores the applications of that result by developing techniques for sampling from a uniform distribution. Specifically, this lecture introduces the concept of Markov Chain Monte Carlo (MCMC) sampling approaches.

25.1 Markov Chain Monte Carlo (MCMC)

25.1.1 Problem motivation

There are many situations where we wish to sample from a given distribution, but it is not immediately clear how to do so. For example, we may have a function that gives the probability of an event within a normalization factor, but no way to calculate that normalization factor. Or the sample space of possible outcomes may be exponentially large, or even infinite.

25.1.2 Random walks and their properties

We approach this problem by considering our state space to be a graph, with individual events as nodes on this graph. If we can define this graph in a way such that the stationary distribution of a random walk over this graph is equal to our target distribution that we wish to draw samples from, then samples from a random walk on this graph can be used to approximate samples from the target distribution. We introduce the following definitions:

- $\Omega =$ the state space
- $n = |\Omega|$
- $P =$ the transition matrix, $P_{ij} = Pr[i \rightarrow j]$
- $\pi =$ a distribution on the nodes in Ω

Then, if we start from a node chosen from the distribution π and take a single step according to our transition matrix P, we get the distribution πP. Note that a random walk obeying these definitions is memoryless. That is, the next step in the walk depends only on the current state, and not on any history beyond that. This is also known as the *Markov property*, and our random walk is an example of a *Markov chain*.

For our random walk, there are two quantities of particular interest. The first is the *stationary distribution* π^*. This is a distribution over all states with the special property that $\pi^* P = \pi^*$. If we follow a random walk on a Markov chain, after a while we expect our position to be distributed according to π^*. This is the definition of a stationary distribution. It is the limit distribution of the location of a random walk as the number of steps taken goes to infinity. There are special properties of the chain which are required to guarantee the existence and uniqueness of π^*, and these will be introduced shortly.

The second quantity of interest is the *mixing time* τ_ϵ, which is a measure of how long a random walk on the graph will take to converge to π^*. This will be defined more formally.

To illustrate these concepts, we will study the simple example of a d-regular undirected graph (all vertices have degree d). We define the transition probabilities from vertex to be uniform over all outgoing edges. That is, each edge is taken with probability $1/d$. We then make the following claims.

Claim 25.1.1 *For a uniform random walk over our d-regular graph, π^* is uniform.*

Claim 25.1.2 *For the directed version of our graph with $d(in) = d(out) = d$, π^* is uniform.*

Claim 25.1.3 *If G is undirected, $\pi^*(v) = \frac{d(v)}{2m}$.*

To see this last claim, consider that we can also define our random walk as a distribution over all edges. Let $Q(u, v)$ be the probability of taking the edge (u, v). Then:

$$\begin{aligned} Q(u,v) &= \pi^*(u)P_{uv} && (25.1.1)\\ &= \frac{d(u)}{2m}\frac{1}{d(u)} && (25.1.2)\\ &= \frac{1}{2m} && (25.1.3)\end{aligned}$$

Summing this over all d of v's neighbors then shows us that this π^* satisfies $\pi^*P = \pi^*$ for any all v and is therefore a valid stationary distribution. The first two claims follow by a similar argument.

But when can we know that a unique π^* exists? Consider the 2 example graphs shown in Figure 25.1.2. For the first graph, if we say that you start in the left node with probability 1, then your probability of being in that node is 1 at the start of iteration 1,3,5,... and 0 at the start of iteration 2,4,6,... Although the uniform distribution satisfies $\pi P = \pi$, it is not guaranteed that all random walks on the graph will converge to this distribution, as shown by this example. Therefore this graph can not have a stationary distribution, because it is periodic. For the second graph, you can reach 2 different stationary distributions starting from the center node, depending on which direction is taken on the first step. In this case there is no unique stationary distribution. These ideas are formalized in the following theorem:

Theorem 25.1.4 *An aperiodic irreducible finite Markov chain is ergodic and has a unique stationary distribution.*

A chain is aperiodic if for every state, there is no number > 1 which can divide the index of every future step which has non-zero probability of returning to that state. That is, given that you are in the state, there is no periodic pattern to when you can return (every second step, third step etc). This can be a somewhat tricky notion to prove about a graph, but adding self-edges to all nodes

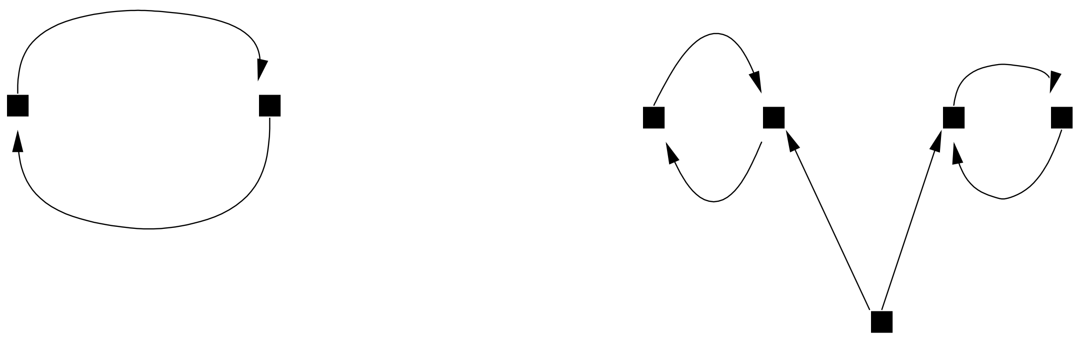

Figure 25.1.1: Two Markov chains which do not have unique stationary distributions.

automatically makes the chain aperiodic, so it is often easiest to simply define the graph in this way.

A chain is irreducible if it is possible to get from any state to any other state. This notion of ’reachability’ can be considered an equivalence relation, with a chain being irreducible if all states are in the same equivalence class.

Now we formally define the mixing time τ_ϵ, first defining the mixing time of a random walk starting from state x.

$$\tau_\epsilon(x) = \min\{t \text{ s.t.} |\pi_x P^{t'} - \pi^*|_1 < \epsilon \ \forall t' \geq t\} \tag{25.1.4}$$

Where $\pi_x = 1$ for state x, and 0 for all other states. The mixing time for the chain itself is then:

$$\tau_\epsilon = \max_x \tau_\epsilon(x) \tag{25.1.5}$$

Different chains can have widely varying mixing times. For example, a walk over a regular expander can never get ’trapped’ in any subset of the graph, because of the expander property. In the ’lollipop’ graph (Figure 25.1.2) which consists of a clique with $n/2$ nodes and a long ’stem’ with $n/2$ nodes, a random walk will take $O(n^3)$ steps to ever reach the end of the stem starting from a point inside the clique.

25.1.3 Mixing time analysis

For a given chain then, we would like to bound τ_ϵ by some polynomial in n. How can we do this? There are 3 general analysis approaches.

1. Conductance
2. Canonical paths
3. Coupling

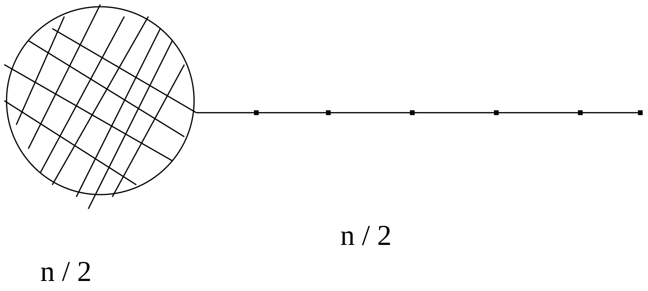

Figure 25.1.2: A graph with a slow mixing time.

The key property of the transition matrix which determines mixing time is the eigenvalue gap γ between the principal and second eigenvalues. Assume that all eigenvalues λ are real (which is the case for undirected graphs anyways). Then order the eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. Then call the eigenvalue gap $\gamma = \lambda_1 - \lambda_2$.

We know that at least one of the λ_i's is equal to 1, because by definition π^* is an eigenvector or P whenever a stationary distribution exists ($\pi^* P = \pi^*$). Since P is a stochastic matrix, it is also easy to see that $|\lambda_i| \leq 1 \forall i$, therefore $\lambda_1 = 1$ and π^* is the first eigenvector. It is also interesting to note than if $|\lambda_i| < 1$, that means that v_i must have mixed-sign components, and cannot be normalized to sum to 1. Also, λ_2 cannot be 1, unless the graph has a more than one strongly connected component, and the stationary distribution is therefore non-unique.

Also, the eigenvectors of P are guaranteed to be orthogonal if P is real and symmetric, which corresponds to a time-reversible Markov chain, where time-reversible means that $\pi^*(u)P_{uv} = \pi^*(v)P_{vu} \forall u, v$.

Theorem 25.1.5 *$\tau_\epsilon \leq O(\frac{1}{\gamma}\log(\frac{n}{\epsilon})$ for a time-reversible Markov chain.*

Proof: Consider the representation of a distribution over states π in the basis of the eigenvectors $\pi = \sum_i c_i v_i$.

$$\pi P = (\sum_i c_i v_i)P = \sum_i c_i \lambda_i v_i \tag{25.1.6}$$

$$\pi P^2 = \sum_i c_i \lambda_i^2 v_i \tag{25.1.7}$$

$$\pi P^t = \sum_i c_i \lambda_i^t v_i \tag{25.1.8}$$

Observe that for $|\lambda_i| < 1$, $\lambda_i^t \to 0$ as $t \to \infty$. The lone exception if for $i = 1$, since $|\lambda_1| = 1$. Thus we can see that $\pi P^t \to \pi^* = c_1 v_1$ as $t \to \infty$, where $c_1 = 1$ as $\pi^* = v_1$. We can express the distribution at time t in terms of the stationary distribution and an error term.

$$\begin{aligned}
\pi P^t &= \pi^* + \sum_{i>1} c_i \lambda_i^t v_i && (25.1.9)\\
||\pi P^t - \pi^*||_2^2 &= ||\sum_i c_i \lambda_i^t v_i||_2^2 && (25.1.10)\\
&= \sum_i c_i^2 \lambda_i^{2t} ||v_i||_2^2 && (25.1.11)
\end{aligned}$$

(the previous step uses the orthogonality of the v_i's)

$$\begin{aligned}
&\leq \lambda_2^{2t} \sum_{i>1} c_i^2 ||v_i||_2^2 && (25.1.12)\\
&\leq \lambda_2^{2t} && (25.1.13)
\end{aligned}$$

The final step can be seen by noting that $\sum_{i>1} c_i^2||v_i||_2^2 \leq \sum_i c_i^2||v_i||_2^2 = ||\pi||_2^2 \leq 1$, because our original π is a probability distribution.

Now we wish to use this ℓ_2 bound to get an ℓ_1 bound. It is a general result that, in n-dimensional space $||x||_1 \leq \sqrt{n}||x||_2$ (Figure 25.1.3). This result is essentially a restatement of the Cauchy-Schwarz inequality. Plugging this result into our bound from above, we get:

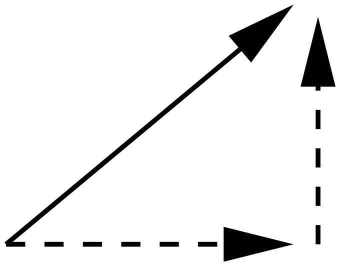

Figure 25.1.3: Visual representation of bounding the ℓ_1-norm in terms of the ℓ_2-norm.

$$||\pi P^t - \pi^*||_1 \leq \lambda_2^t \sqrt{n} \qquad (25.1.14)$$

We want to pick t such that $\lambda_2^t\sqrt{n} \leq \epsilon$.

$$\begin{aligned}
(1-\gamma)^t\sqrt{n} &\leq \epsilon && (25.1.15)\\
e^{-t\gamma}\sqrt{n} &\leq \epsilon && (25.1.16)\\
\frac{\sqrt{n}}{\epsilon} &\leq e^{t\gamma} && (25.1.17)\\
t &\geq \frac{1}{\gamma}\log(\frac{\sqrt{n}}{\epsilon}) && (25.1.18)
\end{aligned}$$

■

25.1.4 Conductance

Definition 25.1.6 *The* conductance $\phi(G)$ *of a distribution* π *over graph* G *is defined as*

$$\phi(G) = \min_{S \subset V, \pi(S) \leq 1/2} \left[\frac{\sum_{u \in S, v \notin S} \pi(u) P_{uv}}{\pi(S)} \right] \tag{25.1.19}$$

The quantity in the brackets can be roughly understood as the probability of 'escaping' set S. The conductance $\phi(G)$ then finds the 'stickiest' set in G, and is closely related to the notion of graph sparsity discussed earlier in the course. We can bound the conductance in terms of the eigenvalue gap γ.

Theorem 25.1.7

$$\frac{\phi^2(G)}{2} \leq \gamma \leq 2\phi(G) \tag{25.1.20}$$

This inequality can be used to bound the sparsity, but since $\phi(G)$ can be small, $\phi^2(G)$ can be very small, giving a loose and unhelpful bound. If we can show that $\phi(G)$ is large however, this bound will be relatively tight. A tight bound on the eigenvalue gap γ can then be plugged into the previous theorem to bound τ_ϵ. This is the basic approach of the conductance method.

Corollary 25.1.8

$$\tau_\epsilon \leq O(\frac{1}{\phi^2(G)} \log(\frac{n}{\epsilon})) \tag{25.1.21}$$

25.1.5 Canonical paths

The basic concept of the canonical paths technique is to find paths between all pairs of points (canonical paths), and then argue that no edge is "overloaded" with respect to its probability, which plays a role analogous to its capacity.

If the canonical paths can be shown to have low congestion, this can be shown to be equivalent to saying the graph has high conductance.

Let T_{xy} be the canonical path between x and y. Then the congestion ρ of a set of canonical paths T be defined as:

$$\rho(T) = \max_e \sum_{e \in T_{xy}, e=(u,v)} \frac{\pi^*(x)\pi^*(y)}{Q(u,v)} \tag{25.1.22}$$

The congestion $\rho(T)$ can then be used to bound the mixing time τ_ϵ.

Theorem 25.1.9

$$\tau_\epsilon \leq \rho(T) \log(\frac{n}{\epsilon}) \tag{25.1.23}$$

25.1.6 Coupling

The coupling technique works by running two Markov chains X and Y in parallel. The chain X is started from some initial distribution π, while the chain Y is started from the stationary distribution π^*. The evolution of the chains im time is then coupled or linked. If it can be shown that for some t we get $X^t = Y^t$, then by the Markov property the chain X will have reached the stationary distribution π^*.

Consider the example of the random walk on the graph shown in Figure 25.1.6. From our earlier claims, we can immediately see the stationary distribution π^* for this walk is uniform. So now consider our two random walks X and Y, where their evolution is defined such that for each step:

- X chooses a neighbor uniformly
- Y go in the same 'direction' as X

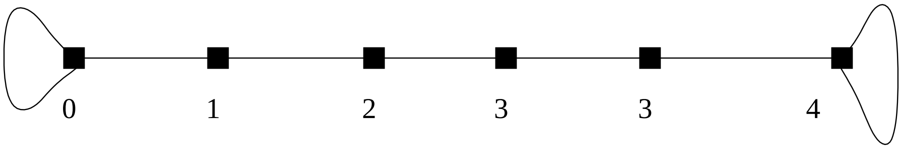

Figure 25.1.4: Coupling method example.

The key consequence of this scheme is that whenever one of the random walks is at an endpoint, the distance between the two points will be reduced by 1 if that walk takes the self-loop edge of that endpoint, which will happend with probability 1/2. This means that, no matter where the two chains started, we can guarantee that $X = Y$ after a self-loop edge has been taken n times. The following lemma formalizes the relationship between mixing (small $||X^t - \pi^t||_1$) and coupling ($X^t = Y^t$).

Lemma 25.1.10 $||X^t - \pi^*||_1 \leq 2Pr[X^t \neq Y^t]$

Proof:

$$\begin{aligned}
Pr[X^t = Y^t] &\leq \sum_i \min\{X_i^t, Y_i^t\} && (25.1.24)\\
||X^t - Y^t||_1 &= \sum_i \max\{X_i^t, Y_i^t\} - \min\{X_i^t, Y_i^t\} && (25.1.25)\\
&= \sum_i (\max\{X_i^t, Y_i^t\} + \min\{X_i^t, Y_i^t\}) - 2\min\{X_i^t, Y_i^t\} && (25.1.26)\\
&= 2 - 2\sum_i \min\{X_i^t, Y_i^t\} && (25.1.27)\\
&\leq 2(1 - Pr[X^t = Y^t]) && (25.1.28)\\
&\leq 2(Pr[X^t \neq Y^t]) && (25.1.29)
\end{aligned}$$

■

In order to complete the argument, we need to determine a t such that $Pr[X^t \neq Y^t] \leq \epsilon$. The crux of this method is determining the hitting time $h(u,v)$ for each pair of points (u,v), where $h(u,v)$ is the expected time to reach v, starting from u. This will give us the time it takes to go to an end point on the line starting from an arbitrary point in the middle.

For this example, the maximum $h(u,v)$ occurs when u and v are the two endpoints on the opposite sides of the graph, A and B. This can be shown to be n^2 by solving the following set of equations.

$$h(i,0) = 1 + \frac{1}{2}h(i-1,0) + \frac{1}{2}h(i+1,0) \ \forall 1 < i < n \tag{25.1.30}$$

$$h(1,0) = 1 + \frac{1}{2}h(2,0) \tag{25.1.31}$$

This $O(n^2)$ hitting time in turn implies an $O(1/\epsilon, n^3)$ mixing time for this random walk. We omit the details of this step, which involves the use of Markov's inequality.

The coupling analysis technique requires typically requires the underlying graph to have a nice, known structure. It can be applied to a number of Markov chains, such as electrical networks, card shuffling, and random graph colorings.

References

[1] V. Vazirani. Approximation Algorithms. Springer, 2001.