PRINCETON UNIVERSITY F'02 COS 597D: A THEORIST'S TOOLKIT

Lecture 7: Markov Chains and Random Walks

Lecturer: *Sanjeev Arora* Scribe:*Elena Nabieva*

1 Basics

A *Markov chain* is a discrete-time stochastic process on n states defined in terms of a transition probability matrix (M) with rows i and columns j.

$$\mathbf{M} = \left(P_{ij}\right)$$

A transition probability P_{ij} corresponds to the probability that the state at time step $t+1$ will be j, given that the state at time t is i. Therefore, each row in the matrix $\mathbf{M}$ is a distribution and $\forall i, j \in S P_{ij} \geq 0$ and $\sum_j P_{ij} = 1$.

Let the initial distribution be given by the row vector $\mathbf{x} \in \mathbf{R}^n$, $x_i \geq 0$ and $\sum_i x_i = 1$. After one step, the new distribution is $\mathbf{xM}$. It is easy to see that $\mathbf{xM}$ is again a distribution. Sometimes it is useful to think of x as describing a certain amount fluid sitting at each node, such that the sum of the amounts is 1. After one step, the fluid sitting at node i distributes to its neighbors, such that P_{ij} fraction goes to j.

We stress that the evolution of a Markov chain is *memoryless*: the transition probability P_{ij} depends only on the state i and not on the time t or the sequence of transititions taken before this time.

Suppose we take two steps in this Markov chain. The memoryless property implies that the probability of going from i to j is $\sum_k P_{ik} P_{kj}$, which is just the (i,j)th entry of the matrix M^2. In general taking t steps in the Markov chain corresponds to the matrix M^t.

DEFINITION 1 *A distribution π for the Markov chain $\mathbf{M}$ is a* stationary distribution *if $\pi\mathbf{M} = \pi$.*

Note that an alternative statement is that π is an eigenvector which has all nonnegative coordinates and whose corresponding eigenvalue is 1.

EXAMPLE 1 Consider a Markov chain defined by the following random walk on the nodes of an n-cycle. At each step, stay at the same node with probability 1/2. Go left with probability 1/4 and right with probability 1/4.

The uniform distribution, which assigns probability $1/n$ to each node, is a stationary distribution for this chain, since it is unchanged after applying one step of the chain.

DEFINITION 2 *A Markov chain $\mathbf{M}$ is* ergodic *if there exists a unique stationary distribution π and for every (initial) distribution $\mathbf{x}$ the limit $\lim_{t\to\infty} \mathbf{xM}^t = \pi$.*

THEOREM 1
The following are necessary and sufficient conditions for ergodicity:

1. *connectivity: $\forall i, j : \mathbf{M}^t(i,j) > 0$ for some t.*

2. *aperiodicity:* $\forall i : gcd\{t : \mathbf{M}^t(i,j) > 0\} = 1$.

REMARK 1 Clearly, these conditions are necessary. If the Markov chain is disconnected it cannot have a unique stationary distribution —there is a different stationary distribution for each connected component. Similarly, a bipartite graph does not have a unique distribution: if the initial distribution places all probability on one side of the bipartite graph, then the distribution at time t oscillates between the two sides depending on whether t is odd or even. Note that in a bipartite graph $gcd\{t : \mathbf{M}^t(i,j) > 0\} \geq 2$. The sufficiency of these conditions is proved using eigenvalue techniques (for inspiration see the analysis of mixing time later on).

Both conditions are easily satisfied in practice. In particular, any Markov chain can be made aperiodic by adding self-loops assigned probability 1/2.

DEFINITION 3 *An ergodic Markov chain is* reversible *if the stationary distribution π satisfies for all i, j, $\pi_i \mathbf{P}_{ij} = \pi_j \mathbf{P}_{ji}$.*

Uses of Markov Chains. A Markov Chain is a very convenient way to model many situations where the "memoryless" property makes sense. Examples including communication theory (Markovian sources), linguistics (Markovian models of language production), speech recognition, internet search (Google's Pagerank algorithm is based upon a Markovian model of a random surfer).

2 Mixing Times

Informally, the *mixing time* of a Markov chain is the time it takes to reach "nearly uniform" distribution from any arbitrary starting distribution.

DEFINITION 4 *The mixing time of an ergodic Markov chain M is t if for every starting distribution x, the distribution xM^t satisfies $\left|xM^t - \pi\right|_1 \leq 1/4$. (Here $|\cdot|_1$ denotes the ℓ_1 norm and the constant "1/4" is arbitrary.)*

The next exercise clarifies why we are interested in ℓ_1 norm.

EXERCISE 1 For any distribution π on $\{1, 2, \ldots, N\}$, and $S \subseteq \{1, 2, \ldots, N\}$ let $\pi(S) = \sum_{i \in S} \pi_i$. Show that for any two distributions π, π',

$$\left|\pi - \pi'\right|_1 = 2 \max_{S \subseteq \{1,\ldots,N\}} \left|\pi(S) - \pi'(S)\right|. \tag{1}$$

Here is another way to restate the property in (1). Suppose A is some deterministic algorithm (we place no bounds on its complexity) that, given any number $i \in \{1, 2, \ldots, N\}$, outputs Yes or No. If $|\pi - \pi'|_1 \leq \epsilon$ then the probability that A outputs Yes on a random input drawn according to π cannot be too different from the probability it outputs Yes on an input drawn according to π'. For this reason, ℓ_1 distance is also called *statistical difference*.

We are interested in analysing the mixing time so that we can draw a sample from the stationary distribution.

EXAMPLE 2 (Mixing time of a cycle) Consider an *n-cycle*, i.e., a Markov chain with n states where, at each state, $\mathbf{Pr}(left) = \mathbf{Pr}(right) = \mathbf{Pr}(stay) = 1/3$.

Suppose the initial distribution concentrates all probability at state 0. Then t steps correspond to about $2t/3$ random coin tosses and the index of the final state is

$$(\#(\text{Heads}) - \#(\text{Tails})) \pmod{n}.$$

Clearly, it takes $\Omega(n^2)$ steps for the walk to reach the other half of the circle with any reasonable probability, and the mixing time is $\Omega(n^2)$. We will later see that this lowerbound is fairly tight.

2.1 Approximate Counting and Sampling

Markov chains allow one to sample from very nontrivial sets, provided we know how to find at least *one* element of this set. The idea is to define a Markov chain whose state space is the same as this set. The Markov chain is such that it has a unique stationary distribution, which is uniform. We know how to find one element of the set. We do a walk according to the Markov chain with this as the starting point, and after $T = O(\text{mixing time})$ steps, output the node we are at. This is approximately a random sample from the set. We illustrate this idea later. First we discuss why sampling from large sets is important.

Usually this set has exponential size set and it is only given implicitly. We give a few examples of some interesting sets.

EXAMPLE 3 (Perfect matchings) For some given graph $G = (V, E)$ the set of all perfect matchings in G could be exponentially large compared to the size of the graph, and is only know implicitly. We know how to generate some element of this set, since we can find a perfect matching (if one exists) in polynomial time. But how do we generate a random element?

EXAMPLE 4 ($0-1$ knapsack) Given $a_1 \ldots a_n, b \in \mathbf{Z}^+$, the set of vectors $(x_i, \ldots, x_n)$ s.t. $\sum a_i x_i \leq b$.

In both cases, determining the exact *size* of the set is in $\#\mathbf{P}$ (the complexity class corresponding to counting the number of solutions to an **NP** problem). In fact, we have the following.

THEOREM 2 (VALIANT, LATE 1970S)
If there exist a polynomial-time algorithm for counting the number of perfect matchings or the number of solutions to the $0-1$ *knapsack counting problem, then* $\mathbf{P} = \mathbf{NP}$ *(in fact,* $\mathbf{P} = \mathbf{P}^{\#\mathbf{P}}$*).*

Valiant's Theorem does not rule out finding good approximations to this problem.

DEFINITION 5 *A* Fully Polynomial Randomized Approximation Scheme *(FPRAS) is a randomized algorithm, which for any* ϵ *finds an answer in time polynomial in* $(\frac{n}{\epsilon} \log \frac{1}{\delta})$ *that is correct within a multiplicative factor (1+*ϵ*) with probability (1-*δ*).*

It turns out that approximate counting is equivalent to approximate sampling.

THEOREM 3 (JERRUM, VALIANT, VAZIRANI, 1984)
If we can sample almost uniformly, in polynomial time, from $A = \{(x_1, \ldots, x_n) \colon \sum a_i x_i \leq b\}$, *then we can design an FPRAS for the knapsack counting problem.*

Conversely, given an FPRAS for knapsack counting, we can draw an almost uniform sample from A.

REMARK 2 By "sampling almost uniformly" we mean having a sampling algorithm whose output distribution has ℓ_1 distance $\exp(-n^2)$ (say) from the uniform distribution. For ease of exposition, we think of this as a uniform sample.

PROOF: We first show how to count approximately assuming there is a polynomial time sampling algorithm. The idea is simple though the details require some care (which we suppress here). Suppose we have a sampling algorithm for knapsack. Draw a few samples from A, and observe what fraction feature $x_1 = 0$. Say it is p. Let A_0 be the set of solutions with $x_1 = 0$. Then $p = |A_0| / |A|$. Now since A_0 is the set of $(x_2, \ldots, x_n)$ such that $\sum_{i \geq 2} a_i x_i \leq b - a_1 x_1$, it is also the set of solutions of a knapsack problem, but with one fewer variable. Using the algorithm recursively, assume we can calculate $|A_0|$. Then we can calculate

$$|A| = |A_0| / p.$$

Now if we do not know $|A_0|, p$ accurately but up to some accuracy, say $(1 + \epsilon)$. So we will only know $|A|$ up to accuracy $(1 + \epsilon)^2 \approx 1 + 2\epsilon$.

Actually the above is not accurate, since it ignores the possibility that p is so small that we never see an element of A_0 when we draw $poly(n)$ samples from A. However, in that case, the set $A_1 = A \setminus A_0$ must be at least $1/2$ of A and we can estimate its size. Then we proceed in the rest of the algorithm using A_1.

Therefore, by choosing ϵ appropriately so that $(1 + \epsilon)^n$ is small, and using the Chernoff bound, we can achieve the desired bound on the error in polynomial time.

The converse is similar. To turn a counting algorithm into a sampling algorithm, we need to show how to output a random member of A. We do this bit by bit, first outputting x_1, then x_2, and so on. To output x_1, output 0 with probability p and 1 with probablity $1 - p$, where $p = |A_0| / |A|$ is calculated by calling the counting algorithm twice. Having output x_1 with the correct probability, we are left with a sampling problem on $n - 1$ variables, which we solve recursively. Again, we need some care because we only have an approximate counting algorithm instead of an exact algorithm. Since we need to count the approximate counting algorithm only $2n$ times, an error of $(1 + \epsilon)$ each time could turn into an error of $(1 + \epsilon)^{2n}$, which is about $1 + 2\epsilon$. □

Thus to count approximately, it suffices to sample from the uniform distribution. We define a Markov chain M on A whose stationary distribution is uniform. Then we show that its mixing time is poly(n).

The Markov chain is as follows. If the current node is $(x_1, \ldots, x_n)$ (note $a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \leq b$) then

1. with probability $1/2$ remain at the same node
2. else pick $i \in \{1, \ldots, n\}$.

   Let $\mathbf{y} = (x_1, \ldots, x_{i-1}, 1 - x_i, x_{i+1}, \ldots, x_n)$. If $\mathbf{y} \in A$, go there. Else stay put.

Note that M is

1. aperiodic because of self-loops
2. connected because every sequence can be turned into the zero vector in a finite number of transformations, i.e., every node is connected to $\vec{0}$.

Therefore, M is ergodic, i.e., has a unique stationary distribution. Since the uniform distribution is stationary, it follows that the stationary distribution of M is uniform.

Now the question is: how fast does M converge to the uniform distribution? If M mixes fast, we can get an efficient approximation algorithm for the knapsack counting: we get the solution by running M for the mixing time and sampling from the resulting distribution after the mixing time has elapsed.

THEOREM 4
(Morris-Sinclair, 1999): The mixing time for M is $O(n^8)$.

Fact (see our remark later in our analysis of mixing time): running the M for a bit longer than the mixing time results in a distribution that is extremely close to uniform.

Thus, we get the following sampling algorithm:

1. Start with the zero vector as the initial distribution of M.
2. Run M for $O(n^9)$ time.
3. output the node at which the algorithm stops.

This results in a uniform sampling from A.

Thus Markov chains are useful for sampling from a distribution. Often, we are unable to prove any useful bounds on the mixing time (this is the case for many Markov chains used in simulated annealing and the Metropolis algorithm of statistical physics) but nevertheless in practice the chains are found to mix rapidly. Thus they are useful even though we do not have a proof that they work.

3 Bounding the mixing time

For simplicity we restrict attention to regular graphs.

Let M be a Markov chain on a d-regular undirected graph with an adjacency matrix A. Assume that M is ergodic and that d includes any self-loops.

Then, clearly $M = \frac{1}{d}A$.

Since M is ergodic, and since $\frac{1}{n}\vec{1}$ is a stationary distribution, then $\frac{1}{n}\vec{1}$ is the unique stationary distribution for M.

The question is how fast does M convege to $\frac{1}{n}\vec{1}$? Note that if $\mathbf{x}$ is a distribution, $\mathbf{x}$ can be written as

$$\mathbf{x} = \vec{1}\frac{1}{n} + \sum_{i=2}^{n} \alpha_i \mathbf{e_i}$$

where $\mathbf{e_i}$ are the eigenvectors of M which form an orthogonal basis and $\mathbf{1}$ is the first eigenvector with eigenvalue 1. (Clearly, $\mathbf{x}$ can be written as a combination of the eigenvectors; the observation here is that the coefficient in front of the first eigenvector $\vec{1}$ is $\vec{1} \cdot x / \left|\vec{1}\right|_2^2$ which is $\frac{1}{n}\sum_i x_i = \frac{1}{n}$.)

$$\begin{aligned} M^t\mathbf{x} &= M^{t-1}(M\mathbf{x}) \\ &= M^{t-1}(\frac{1}{n}\vec{1} + \sum_{i=2}^{n} \alpha_i \lambda_i \mathbf{e_i}) \\ &= M^{t-2}(M(\frac{1}{n}\vec{1} + \sum_{i=2}^{n} \alpha_i \lambda_i \mathbf{e_i})) \\ &\cdots \\ &= \frac{1}{n}\vec{1} + \sum_{i=2}^{n} \alpha_i \lambda_i^t \mathbf{e_i} \end{aligned}$$

Also

$$\| \sum_{i=2}^{n} \alpha_i \lambda_i^t \mathbf{e_i} \|_2 \leq \lambda_{max}^t$$

where λ_{max} is the second largest eigenvalue of M. (Note that we are using the fact that the total ℓ_2 norm of any distribution is $\sum_i x_i^2 \leq \sum_i x_i = 1$.)

Thus we have proved $\left|M^t\mathbf{x} - \frac{1}{n}\mathbf{1}\right|_2 \leq \lambda_{max}^t$. Mixing times were defined using ℓ_1 distance, but Cauchy Schwartz inequality relates the ℓ_2 and ℓ_1 distances: $|p|_1 \leq \sqrt{n}\,|p|_2$. So we have proved:

THEOREM 5
The mixing time is at most $O(\frac{\log n}{\lambda_{\max}})$.

Note also that if we let the Markov chain run for $O(k \log n/\lambda_{\max})$ steps then the distance to uniform distribution drops to $\exp(-k)$. This is why we were not very fussy about the constant 1/4 in the definition of the mixing time earlier.

Finally, we recall from the last lecture: for $S \subset V$, $Vol(S) = \sum_{i \in S} d_i$, where d_i is the degree of node i, the *Cheeger Constant* is

$$h_G = \min_{S \subset V, vol(S) \leq \frac{Vol(V)}{2}} \frac{\left|E(S, \overline{S}\right|}{Vol(S)}$$

If μ is the smallest nonzero eigenvalue of the Laplacian L of M, then

$$2h_G \geq \mu \geq \frac{h_G^2}{2}$$

The Laplacian for our graph is

$$L = I - M$$

Therefore,

$$spec(L) = \{0 = \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n\}$$

and

$$spec(M) = \{1 = 1 - \mu_1 \geq 1 - \mu_2 \geq \ldots \geq 1 - \mu_n\}$$

Note that $\lambda_{max} = (1 - \mu_2)^t$.
Therefore,

$$\| \sum_{i=2}^{n} \alpha_i \lambda_i^t \mathbf{e_i} \|_2 \leq (1 - \frac{h_G^2}{2})^t \sqrt{n}$$

.

and we obtain the Jerrum-Sinclair inequality:
$\|M^t \mathbf{x} - \frac{1}{n}\vec{1}\|_2 \leq (1 - \frac{h_G^2}{2})^t \sqrt{n}$.
Examples:

1. For n-cycle: $\lambda_{max} = (1 - \frac{c}{n^2})^t$, mixing time $\approx O(n^2 \log n)$ (c is some constant).
2. For a hypercube on 2^n nodes (with self-loops added), $\lambda_{max} = (1 - \frac{c}{n})$ (this was a homework problem), so mixing time $\approx O(n \log n)$ (c is some constant).

Observe that the mixing time is much smaller than the number of nodes, i.e., the random walk does not visit all nodes.

Finally, we note that random walks also give a randomized way to check $s-t$ connectivity (for undirected graphs) in logarithmic space, a surprising result since the usual method of checking $s - t$ connectivity, namely, breadth-first-search, seems to inherently require linear space.

The main idea is that a random walk on a connected graph on n nodes mixes in $O(n^4)$ time (the Cheeger constant must be at least $1/n^2$) and so a logarithmic space algorithm can just do a random walk for $O(n^2 \log n)$ steps (note that space $O(\log n)$ is required is just to store the current node and a counter for the number of steps) starting from s, and if it never sees t, it can output reject. This answer will be correct with high probability. This application of random walks by Alleliunas, Karp, Karmarkar, Lipton and Lovasz 1979 was probably one of the first in theoretical computer science and it has been the subject of much further work recently.

4 Analysis of Mixing Time for General Markov Chains

Thanks to Satyen Kale for providing this additional note

In the class we only analysed random walks on d-regular graphs and showed that they converge exponentially fast with rate given by the second largest eigenvalue of the transition matrix. Here, we prove the same fact for general ergodic Markov chains. We need a lemma first.

LEMMA 6
Let M be the transition matrix of an ergodic Markov chain with stationary distribution π and eigenvalues $\lambda_1(=1) \geq \lambda_2 \geq \ldots \geq \lambda_n$, corresponding to eigenvectors $v_1(=\pi), v_2, \ldots v_n$. Then for any $k \geq 2$,

$$v_k \vec{1} = 0.$$

PROOF: We have $v_k M = \lambda_k v_k$. Mulitplying by $\vec{1}$ and noting that $M\vec{1} = \vec{1}$, we get

$$v_k \vec{1} = \lambda_k v_k \vec{1}.$$

Since the Markov chain is ergodic, $\lambda_k \neq 1$, so $v_k \vec{1} = 0$ as required. □

We are now ready to prove the main result concerning the exponentially fast convergence of a general ergodic Markov chain:

THEOREM 7
In the setup of the lemma above, let $\lambda = max\{|\lambda_2|, |\lambda_n|\}$. Then for any initial distribution x, we have

$$||xM^t - \pi||_2 \leq \lambda^t ||x||_2.$$

PROOF: Write x in terms of $v_1, v_2, \ldots, v_n$ as

$$x = \alpha_1 \pi + \sum_{i=2}^{n} \alpha_i v_i.$$

Multiplying the above equation by $\vec{1}$, we get $\alpha_1 = 1$ (since $x\vec{1} = \pi\vec{1} = 1$). Therefore $xM^t = \pi + \sum_{i=2}^{n} \alpha_i \lambda_i^t v_i$, and hence

$$\begin{aligned} ||xM^t - \pi||_2 &\leq ||\sum_{i=2}^{n} \alpha_i \lambda_i^t v_i||_2 && (2) \\ &\leq \lambda^t \sqrt{\alpha_2^2 + \cdots + \alpha_n^2} && (3) \\ &\leq \lambda^t ||x||_2, && (4) \end{aligned}$$

as needed. □