CS294-2 Markov Chain Monte Carlo: Foundations & Applications **Fall 2006**

Lecture 2: August 31

Lecturer: Alistair Sinclair *Scribes: Omid Etesami, Alexandre Stauffer*

2.1 Applications of Markov Chain Monte Carlo (continued)

2.1.1 Statistical Inference

Consider a statistical model with parameters Θ and a set of observed data X. The aim is to obtain Θ based on the observed data X, that is, to calculate the probability $\Pr(\Theta \mid X)$. Using Bayes' rule, $\Pr(\Theta \mid X)$ translates to

$$\Pr(\Theta \mid X) = \frac{\Pr(X \mid \Theta)\Pr(\Theta)}{\Pr(X)},$$

where $\Pr(\Theta)$ is the *prior* distribution and refers to the information previously known about Θ, $\Pr(X \mid \Theta)$ is the probability that X is obtained with the assumed model, and $\Pr(X)$ is the unconditioned probability that X is observed. $\Pr(\Theta \mid X)$ is commonly called the *posterior* distribution and can be written in the form $\pi(\Theta) = w(\Theta)/Z$, where the weight $w(\Theta) = \Pr(X \mid \Theta)\Pr(\Theta)$ is easy to compute but the normalizing factor $Z = \Pr(X)$ is unknown. MCMC can then be used to sample from $\Pr(\Theta \mid X)$. We can further use the sampling in the following applications:

- Prediction: obtain the probability $\Pr(Y \mid X)$ that some future data Y is observed given X. $\Pr(Y \mid X)$ clearly can be written as $\sum_\Theta \Pr(Y \mid \Theta)\Pr(\Theta \mid X) = \mathrm{E}_\pi \Pr(Y \mid \Theta)$. Therefore we can use sampling to predict $\Pr(Y \mid X)$.
- Model comparison: perform sampling to estimate $Z = \Pr(X)$, using this to compare some models and find which one is the best.

2.2 Markov Chains

Assume a finite state space Ω. A Markov chain on Ω is a random process $\{X_0, X_1, \ldots, X_t, \ldots\} \in \Omega^\infty$, such that

$$\Pr(X_{t+1} = x_{t+1} \mid X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = \Pr(X_{t+1} = x_{t+1} \mid X_t = x_t) = P(x_t, x_{t+1}),$$

where P is a $\Omega \times \Omega$ matrix called the matrix of transition probabilities. Clearly, P is nonnegative, i.e., $P(x,y) \geq 0$ for all x, y, and $\sum_{y\in\Omega} P(x,y) = 1$ for all x. A matrix P with these properties is called a *stochastic matrix.*

Let $p_x^{(t)}$ be the probability distribution of X_t given that $X_0 = x$. We can write

- $p_x^{(t+1)} = p_x^{(t)} P$ (vector-matrix multiplication)

- $p_x^{(t)} = p_x^{(0)} P^t$ (where of course $p_x^{(0)}$ denotes the point mass at x)
- $p_x^{(t)}(y) = P^t(x,y)$

Sometimes we will also allow a general distribution $p^{(0)}$ at time 0, in which case we write $p^{(t)} = p^{(0)} P^t$ etc.

We call a probability distribution π over Ω a *stationary distribution* for P if $\pi = \pi P$.

Definition 2.1 *P is irreducible if for all x, y, there exists some t such that $P^t(x,y) > 0$.*

Definition 2.2 *P is aperiodic if for all x, y we have* $\gcd\{t : P^t(x,y) > 0\} = 1$. *Equivalently (**exercise!**), P is aperiodic if there exists t such that $P^t(x,y) > 0$ for all x, y.*

Note that both definitions do not refer to specific values of the elements of P, but just to whether those values are nonzero. Now, let $G(P)$ be the (directed) graph on vertex set Ω such that (x,y) is an edge iff $P(x,y) > 0$. Then P is irreducible iff $G(P)$ is strongly connected. If $G(P)$ is undirected (i.e., whenever (x,y) is an edge then so is (y,x)), then P is aperiodic iff $G(P)$ is bipartite (**exercise!**). Notice that the existence of a self-loop in $G(P)$ is sufficient to ensure that P is aperiodic (**exercise!**).

Theorem 2.3 (Fundamental Theorem of Markov Chains) *If P is irreducible and aperiodic then it has a unique stationary distribution π (which is the unique—up to normalization—left eigenvector with eigenvalue 1). Moreover, $P^t(x,y) \to \pi(y)$ as $t \to \infty$ for all $x \in \Omega$.*

The classical proof of this theorem proceeds via the Perron-Frobenius theorem for non-negative matrices:

Theorem 2.4 (Perron-Frobenius) *Any irreducible, aperiodic stochastic matrix P has an eigenvalue $\lambda_0 = 1$ with unique associated left eigenvector $e_0 > 0$. Moreover, all other eigenvalues λ_i of P satisfy $|\lambda_i| < 1$.*

Proof: *(of Theorem 2.3)* Here we present a sketch proof for the case where P is reversible (see section 2.2.1 below). In this case the eigenvalues of P are real, and its eigenvectors span $R^{|\Omega|}$.

- Write the initial distribution over the basis of the eigenvectors as $P^{(0)} = \sum_{i \ge 0} \alpha_i e_i$.
- Then we have $p^{(t)} = \sum_{i \ge 0} \alpha_i e_i \lambda_i^t \to \alpha_0 e_0 = \pi$.

When P is not reversible, its eigenvectors do not necessarily form a basis so the above argument fails. However, using a more technical argument one can still deduce Theorem 2.3 from the Perron-Frobenius theorem in this more general setting. For a proof, see, e.g., the book by Seneta [Se80]. In the next lecture, we will see a more elementary probabilistic proof of the fundamental theorem. ∎

If P is irreducible (but not necessarily aperiodic), then π still exists and is unique, but the Markov chain does not necessarily converge to π from every starting state. For example, consider the two-state Markov chain with $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. This has the unique stationary distribution $\pi = (1/2, 1/2)$, but does not converge from either of the two initial states. Notice that in this example $\lambda_0 = 1$ and $\lambda_1 = -1$, so there is another eigenvalue of magnitude 1, contradicting the Perron-Frobenius theorem. However, the Perron-Frobenius theorem does generalize to the periodic setting, with the weaker conclusion that the remaining eigenvalues satisfy $|\lambda_i| \le 1$.

In this course we will not spend much time worrying about periodicity, because of the following simple observation:

Claim 2.5 *For $0 < \alpha < 1$, if P is irreducible then $P' = \alpha P + (1-\alpha)I$ is irreducible and aperiodic, and has the same stationary distribution as P.*

This operation corresponds to introducing a self-loop at all vertices of $G(P)$ with probability $1 - \alpha$. The value of α is usually set to $1/2$.

P' is called a "lazy" version of P. In the design of MCMC algorithms, we mostly do not need to worry about periodicity, since instead of running the Markov chain P, the algorithm can run the lazy P'. This just has the effect of slowing down time by a factor of 2.

2.2.1 Reversible Markov Chains

Definition 2.6 *A Markov chain P is* reversible *with respect to a distribution π if for every x, y, we have*

$$\pi(x)P(x,y) = \pi(y)P(y,x).$$

Proposition 2.7 *If P is irreducible, aperiodic, and reversible with respect to π, then π is the unique stationary distribution of P.*

Proof: For every y, we have

$$[\pi P](y) = \sum_x \pi(x)P(x,y) = \sum_x \pi(y)P(y,x) = \pi(y).$$

Hence π is stationary, and by the Fundamental Theorem it is unique. ■

Notice that the reversibility condition implies *local* balance of flow for the stationary Markov chain: for every pair of states x, y, the probability that we move from x to y in one step is the same as the probability that we move from y to x. Note that *global* balance of flow holds even for irreversible Markov chains: i.e., for any partition of Ω into two sets $(S, \bar{S})$, in stationarity the probability that in one step we move from S to $\bar{S}$ is the same as the probability that we move from $\bar{S}$ to S, or equivalently $\pi(S)P(S,\bar{S}) = \pi(\bar{S})P(\bar{S}, S)$.

Corollary 2.8 *If P is reversible and symmetric, then the stationary distribution is uniform.*

2.3 Examples of Markov Chains

2.3.1 Random Walks on Undirected Graphs

Definition 2.9 *Random walk on an undirected graph $G(V, E)$ is given by the transition matrix*

$$P(x,y) = \begin{cases} 1/deg(x) & \text{if } (x,y) \in E; \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 2.10 *For random walk P on an undirected graph, we have:*

- *P is irreducible iff G is connected;*
- *P is aperiodic iff G is non-bipartite;*
- *P is reversible with respect to $\pi(x) = deg(x)/(2|E|)$.*

2.3.2 Ehrenfest Urn

In the Ehrenfest Urn, we have 2 urns and n balls, where there are j balls in the first urn and $n-j$ balls in the other. At each step of the Markov chain, we pick a ball u.a.r. and move it to the other urn.

The non-negative entries of the transition matrix are given by

$$\begin{aligned} P(j\ , j+1) &= (n-j)/n, \\ P(j\ , j-1) &= j/n. \end{aligned}$$

The Markov chain is irreducible, and it is easy to see (**exercise!**) that $\pi(j) = \binom{n}{j}/2^n$ is the stationary distribution. However, P is periodic with period 2.

2.3.3 Card Shuffling

In card shuffling, we have a deck of n cards, and we consider the space Ω of all permutations of the cards. Thus $|\Omega| = n!$. The aim is to have the stationary distribution π be uniform.

We look at three different shuffling techniques:

Random Transpositions

Pick two cards i and j uniformly at random, and switch card i with card j.

This is a pretty slow way of shuffling, but it is irreducible (any permutation can be expressed as a product of transpositions), and also aperiodic (since we may choose $i = j$ so the chain has self-loops). Since it is symmetric, that is $P(x, y) = P(y, x)$ for every two permutations x and y, the stationary distribution π is uniform.

Top-to-random

Take the top card and insert it at one of the n positions chosen uniformly at random.

This shuffle is again irreducible and aperiodic (**exercise!**). However, note that it is not reversible: If we insert the top card into (say) the middle of the deck, we cannot bring it back to the top in one step.

However, notice that every permutation y can be obtained, in one step, from exactly n different permutations (corresponding to the n possible choices for the identity of the previous top card). Hence $\sum_x P(x, y) = 1$, or in other words, the matrix P is *doubly stochastic* (its column sums, as well as its row sums, are 1). It is easy to show that the uniform distribution is stationary for doubly stochastic matrices; in fact (**exercise!**), π is uniform *if and only if* P is doubly stochastic.

Riffle Shuffle (Gilbert-Shannon-Reeds [Gi55,Re81])

- *Split the deck into two parts according to the binomial distribution $Bin(n, 1/2)$.*
- *Drop cards in sequence, where the next card comes from the left hand L with probability $|L|/(|L| + |R|)$.*

Notice that the second step of the shuffle is equivalent to choosing a *random interleaving* of the two parts (**exercise!**).

As a final **exercise**, show that the riffle shuffle is irreducible, aperiodic and doubly stochastic (and hence its stationary distribution is again uniform).

References

[Se80] E. Seneta, *Non-negative matrices and Markov chains*, 2nd ed., Springer-Verlag, New York, 1980.

[Gi55] E. Gilbert, "Theory of shuffling," Technical Memorandum, Bell Laboratories, 1955.

[Re81] J. Reeds, Unpublished manuscript, 1981.