CS294 Partition Functions **Fall 2020**

Lecture 6: September 15

Instructor: Alistair Sinclair

6.1 Markov chain Monte Carlo

In this lecture we introduce the Markov chain Monte Carlo (MCMC) paradigm, which remains the most widely applicable approach to the design of approximate counting algorithms. The method actually dates back to applications in statistical physics in the 1960s, though it is only relatively recently that rigorous performance guarantees have been proved for these algorithms, resulting from the development of various quantitative tools for the analysis of Markov chains.

MCMC is actually a method for randomly sampling from a given distribution; however, as we have seen, random sampling (in addition to its inherent interest) leads directly to approximate counting via various types of reduction. We will discuss the MCMC paradigm in the following framework, which captures all the counting and partition function scenarios introduced in previous lectures. Assume we are given a (very large but) finite state space Ω and a weight function $w : \Omega \to \mathbb{R}^+$. Our goal is to design an algorithm that samples every element $x \in \Omega$ with probability $\pi(x) = \frac{w(x)}{Z}$, where $Z = \sum_{x\in\Omega} w(x)$ is the normalizing factor (partition function). For vanilla counting problems, $w(x) = 1$ for all $x \in \Omega$; in the context of generating functions, $w(x) = \lambda^{|x|}$; for statistical physics problems or Markov random fields, $w(x) = \exp(-\beta H(x))$. Note that, although we can readily compute the weight $w(x)$ of any given element x, we have no a priori knowledge of the normalizing factor Z.

The idea is to design an ergodic Markov chain with state space Ω and *stationary distribution* π, and appeal to the fact that, asymptotically, the distribution of the state of the chain approaches π regardless of the initial state. We can then get a random sampling algorithm by simply simulating the chain for sufficiently many steps and outputting the final state. The efficiency of this algorithm will depend crucially on the rate of convergence (usually called the *mixing time*) of the chain.

6.2 Markov chains

Definition 6.1. *A Markov chain on Ω is a stochastic process $\{X_0, X_1, \ldots, X_t, \ldots\}$ with each $X_i \in \Omega$ such that*

$$\Pr(X_{t+1} = y \mid X_t = x, X_{t-1}, \ldots, X_0) = \Pr(X_{t+1} = y \mid X_t = x) =: P(x, y).$$

Clearly, the Markov chain above can be described by the $|\Omega| \times |\Omega|$ matrix P whose (x, y) entry is $P(x, y)$, and we often blur the distinction between the Markov chain and its transition matrix P. Note that P is a *stochastic* matrix, i.e.:

- P is non-negative, i.e., $\forall\ x, y \in \Omega,\ \ P(x, y) \geq 0$;
- all row sums of P are 1, i.e., $\forall x \in \Omega,\ \ \sum_{y\in\Omega} P(x, y) = 1$.

Let $p_x^{(t)} \in \mathbb{R}^{1\times|\Omega|}$ be the row vector corresponding to the distribution of X_t when the Markov chain starts at x, i.e. $X_0 = x$. Then the evolution of the Markov chain can be defined in terms of iterated applications of P:

$$p_x^{(t)} = p_x^{(t-1)}P = p_x^{(0)}P^t.$$

Note that while we assumed that the starting distribution $p_x^{(0)}$ is a point distribution—i.e., the starting vertex is x with probability 1—the same equations hold even if we start with a general distribution.

A Markov chain can also be represented as a weighted directed graph $G = (V, E)$, where $V = \Omega$ and $E = \{(x, y) \in \Omega \times \Omega \mid P(x, y) > 0\}$, and edge (x, y) has weight $P(x, y) > 0$. Note that an edge is present between x and y if and only if the transition probability between x and y is non-zero, and that self-loops are allowed since we can have $P(x, x) > 0$.

Many natural Markov chains have the property that $P(x, y) > 0$ if and only if $P(y, x) > 0$. In this case the graph G is essentially undirected (except for the values of the edge weights). A very important special case is when the Markov chain is *reversible*:

Definition 6.2. *Let $\pi > 0$ be a probability distribution over Ω. A Markov chain P is said to be* reversible *with respect to π if $\forall\ x, y \in \Omega$, $\pi(x)P(x, y) = \pi(y)P(y, x)$.*

Note that any *symmetric* matrix P is trivially reversible w.r.t. the uniform distribution π.

A reversible Markov chain can be completely represented by an undirected graph with weight $Q(x, y) := \pi(x)P(x, y) = \pi(y)P(y, x)$ on edge $\{x, y\}$ (without specifying P or π explicitly; and any fixed multiple of Q will do as well). To see this, note that the transition probability $P(x, y)$ can be computed from $P(x, y) = \frac{Q(x,y)}{\sum_z Q(x,z)}$. In fact, as we shall see below, for a reversible Markov chain π must in fact be its stationary distribution, and this can be computed from the $Q(x, y)$ also (using the fact that $\frac{\pi(x)}{\pi(y)} = \frac{P(y,x)}{P(x,y)}$). Thus a reversible Markov chain can be viewed as a standard random walk on the undirected graph with edge weights $Q(x, y)$.

Remark: For any Markov chain with stationary distribution π, the quantity $\pi(x)P(x, y)$ is called the *ergodic flow* from x to y, i.e., the amount of probability mass flowing from x to y in stationarity. Reversibility says that the ergodic flows from x to y and from y to x are equal; for this reason, the condition in Definition 6.2 is known as the "detailed balance" condition. Of course, by conservation of mass we always have $\pi(S)P(S, \overline{S}) = \pi(\overline{S})P(\overline{S}, S)$ for any subset of states $S \subseteq \Omega$ (where $\overline{S} = \Omega \setminus S$). Detailed balance says that this also holds *locally*, for every pair of states.

6.2.1 Convergence to stationarity

Under mild conditions, any finite Markov chain converges asymptotically to a unique stationary (or equilibrium) distribution, regardless of the initial state.

Definition 6.3. *A probability distribution π over Ω is a* stationary distribution *for P if $\pi = \pi P$.*

Definition 6.4. *A Markov chain P is* irreducible *if for all x, y, there exists some t such that $P^t(x, y) > 0$. Equivalently, the graph corresponding to P is strongly connected (or just connected in case the graph is undirected).*

Definition 6.5. *A Markov chain P is* aperiodic *if for all x, y we have* $\gcd\{t : P^t(x, y) > 0\} = 1$.

We now state a theorem which gives a necessary and sufficient condition for convergence of a Markov chain to its stationary distribution regardless of the initial state.

Theorem 6.6 (Fundamental Theorem of Markov Chains)**.** *If a Markov chain P is irreducible and aperiodic then it has a unique stationary distribution π. This is the unique left eigenvector of P (normalized so that its entries sum to 1) with eigenvalue 1. Moreover, $P^t(x,y) \to \pi(y)$ as $t \to \infty$ for all $x, y \in \Omega$.*

In light of this theorem, we shall sometimes refer to an irreducible, aperiodic Markov chain as *ergodic*.

We shall give an elementary probabilistic proof of the above theorem shortly. However, at this point we will sketch a more traditional algebraic proof for the special case where the Markov chain is reversible.

In preparation for this, let us verify that, if P is reversible w.r.t. π, then π is a stationary distribution.

Claim 6.7. *If a Markov chain P is reversible w.r.t. π, then π is a stationary distribution for P.*

Proof.

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

□

Mathematically, the main reason for the tractability of reversible Markov chains is the following:

Claim 6.8. *Let P be reversible w.r.t. π. Then P is similar to a symmetric matrix under transformation by the diagonal matrix* $\mathrm{diag}(\sqrt{\pi(x)})$*. Thus in particular the eigenvalues and eigenvectors of P are real and P is diagonalizable.*

Proof. Let $D = \mathrm{diag}(\sqrt{(\pi(x))})$. Then

$$DPD^{-1}(x,y) = \frac{\sqrt{(\pi(x))}}{\sqrt{(\pi(y))}}P(x,y) = \frac{\sqrt{(\pi(y))}}{\sqrt{(\pi(x))}}P(y,x) = D^{-1}PD(y,x).$$

□

Our proof of Theorem 6.6 for reversible chains will make use of the following classical theorem:

Theorem 6.9 (Perron-Frobenius)**.** *Any irreducible, aperiodic stochastic matrix P has an eigenvalue $\lambda_1 = 1$ with unique associated left eigenvector $e_1 > 0$. Moreover, all other eigenvalues λ_i of P satisfy $|\lambda_i| < 1$.*

Proof of Theorem 6.6 (sketch for reversible case). By Claim 6.8, P has real eigenvalues and we can choose a basis for $\mathbb{R}^{|\Omega|}$ among its eigenvectors. Let the eigenvalues be $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{|\Omega|}$, and the corresponding eigenvectors be $e_1, e_2, \ldots, e_{|\Omega|}$. By Theorem 6.9, $\lambda_1 = 1$ and $|\lambda_i| < 1$ for all $i \geq 2$; also, by Claim 6.7, $e_1 = \pi$ is the unique stationary distribution. To see the convergence property, note that we can write the initial distribution $p^{(0)}$ as a linear combination of eigenvectors, i.e., $p^{(0)} = \sum_i \alpha_i e_i$. But then $p^{(t)} = p^{(0)}P^t = \sum_i \alpha_i \lambda_i^t e_i$. Since $|\lambda_i| < 1$ for all $i \geq 2$, this implies that $p^{(t)} \to \alpha_1 e_1$, which is the same as e_1 up to a scalar factor. However, by conservation of mass we must have $\alpha_1 = 1$, so the distribution converges to π. □

We can see from the above proof that the rate of convergence is determined by the eigenvalues λ_i for $i \geq 2$, and in particular by the *second eigenvalue* λ_2, which is closest to 1. (Actually the smallest eigenvalue $\lambda_{|\Omega|}$ may also be relevant, but in our applications we may always assume by adding self-loops—see Observation 6.10 below—that $\lambda_{|\Omega|}$ is bounded away from -1 so it does not determine the rate of convergence.) The quantity $1 - \lambda_2$ is known as the *spectral gap*, and for rapid convergence we want it to be not too small. We will discuss this in more detail in due course.

If P is not reversible then the Perron-Frobenius theorem still applies but the linear algebra proof of Theorem 6.6 is a bit more complicated because P is no longer guaranteed to be diagonalizable and we cannot assert the existence of a basis of eigenvectors; see, e.g., the classic text [Sen06] for details.

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 $\lambda_1 = 1$ and $\lambda_2 = -1$, 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 $|\lambda_i| \leq 1$.

In this course we will not spend much time worrying about periodicity, because of the following simple observation (proof: **exercise!**).

Observation 6.10. *Let P be an irreducible (but not necessarily aperiodic) stochastic matrix. For any $0 < \alpha < 1$, the matrix $P' = \alpha P + (1 - \alpha)I$ is stochastic, irreducible and aperiodic, and has the same stationary distribution as P.*

The transformation from P to P' corresponds to introducing a *self-loop* at every state with probability $1 - \alpha$. The value of α is often set to $1/2$. P' is usually called a "lazy" version of P. In the design of MCMC algorithms, we can always eliminate periodicity by passing to a lazy version of our chain. This just has the effect of slowing down the rate of convergence by at most a factor of 2. Moreover, algorithmically we don't even pay this factor of 2 as we can independently simulate a delay at each state using a geometric r.v. of mean 2.

6.3 Examples of Markov Chains

6.3.1 Random Walks on Undirected Graphs

Definition 6.11. *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 6.12. *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 $\pi(x) = \deg(x)/(2|E|)$.*

Proof. **Exercise!** □

As we noted earlier, this is a special case of random walk on a weighted undirected graph with edge weights $Q(x, y)$, which corresponds to an arbitrary reversible Markov chain.

6.3.2 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 sample from the distribution given by the uniform weight $w(x) = 1\ \forall x \in \Omega$, i.e., to sample a permutation of the cards u.a.r. Thus, in the Markov chain setting, we want the stationary distribution π be uniform. Here are three different shuffling schemes:

Random Transpositions

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

This is a pretty slow way of shuffling. The chain 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 random transpositions are invertible, we have $P(x, y) = P(y, x)$ for all x, y, so P is symmetric. This implies immediately that its stationary distribution is uniform (since it is reversible w.r.t. the uniform distribution).

Top-to-random

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

This shuffle is again irreducible **[exercise]** and aperiodic (due to self-loops). However, note that it is not symmetric (or even reversible). 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). Since every non-zero transition probability is $\frac{1}{n}$, this implies that $\sum_x P(x, y) = 1$; thus the matrix P is *doubly stochastic* (i.e., 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 [Gil55, Ree81])

- *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 (resp., right hand R) with probability $\frac{|L|}{|L|+|R|}$ (resp., $\frac{|R|}{|L|+|R|}$).*

Note that the second step of the shuffle is equivalent to choosing an interleaving of the two parts uniformly at random **[exercise]**.

This chain is a reasonable model of how real dealers shuffle cards. The chain is irreducible **[exercise]**, aperiodic (due to self-loops), and doubly stochastic, and hence its stationary distribution is uniform. Because of its real-life importance and mathematical elegance, this process has been heavily analyzed; as an example of this, see [BD92] for rigorous evidence for the folklore claim that "seven shuffles are enough for a 52-card deck", and the references there for other work on shuffling.

Note: This shuffle is quite different from the "perfect shuffle" performed by professional magicians, who split the deck exactly in half and then perfectly interleave the two halves. The perfect shuffle has no randomness, so the configuration of the deck after any given number of shuffles is known exactly—the basis of several card tricks.

6.3.3 Glauber dynamics

Glauber dynamics (also known as the "Gibbs sampler", or "heat-bath" dynamics) is a general term for local Markov chains defined on the very general class of spin systems (or Markov random fields). Recall that configurations of a spin system are assignments $\sigma : V \to [q]$ of q spin values to the vertices of a connected graph $G = (V, E)$. Configuration σ has weight $w(\sigma) = \exp(-\beta H(\sigma))$, where $H(\sigma)$ is a local Hamiltonian and β is an inverse temperature parameter. We will assume for simplicity, as usual, that H includes only nearest neighbor interactions, i.e., it is a sum of vertex and edge potentials. Because of the locality of H, the conditional distribution of the spin σ_v at any vertex v depends only on the spins of its neighbors (the spatial Markov property).

This suggests the following Markov chain for sampling from the Gibbs distribution $\pi(\sigma) = \frac{w(\sigma)}{Z}$:

If the current state is $\sigma \in \Omega$:

- Pick a vertex $v \in V$ uniformly at random.
- Replace the spin σ_v at v by a spin chosen according to the distribution π conditioned on the spins at the neighbors of v.

Note that the required conditional distribution can easily be computed just from the *weights* $w(\sigma')$ of configurations σ' that differ from σ at v (i.e., no knowledge of the normalizing factor Z is required). Moreover, we only need relative weights, which can be computed just from local information around v.

As a concrete example, consider the Ising model from the first lecture. Recall that, in the case of zero field, we can write the weight of a configuration as $w(\sigma) = \lambda^k$, where k is the number of disagreeing edges (i.e., edges connecting $+1/-1$ spins) and $\lambda = \exp(-2\beta)$. Suppose we pick a vertex v, and that the number of neighbors of v with spin $+1$ (resp., spin -1) is d^+ (resp., d^-). Then we set the spin at v to be $+1$ with probability $\frac{\lambda^{d^-}}{\lambda^{d^+}+\lambda^{d^-}}$, and -1 with the complementary probability. [**Exercise:** Check this!]

It is an easy **exercise** to check that the above chain is aperiodic (because of the presence of self-loops), and reversible w.r.t. π. For systems with soft constraints (in which every possible spin assignment is allowed), it is also obviously irreducible (every configuration is reachable from every other via single spin flips)—and indeed this property is often true with hard constraints as well, though it needs to be checked in that case. These properties ensure convergence to the Gibbs distribution, as required.

To see the issue with hard constraints, consider the example of graph colorings with q colors. Here the Glauber dynamics takes the following very simple form: in any current proper coloring, pick a vertex v u.a.r. and flip its color to a randomly chosen *legal* color given the colors of its neighbors. Now if q is too small (say, $q \leq \Delta + 1$, where Δ is the maximum degree of G) then it is easy to construct a coloring that is "frozen" under the above dynamics, i.e., no legal move is possible, although there are other legal colorings of the graph (e.g., just let G be a clique on $\Delta + 1$ vertices). On the other hand, it's not too hard to check [**exercise!**] that if $q \geq \Delta + 2$ then the Glauber dynamics is in fact irreducible.

The above version of Glauber dynamics is often called the "heat-bath" dynamics, and is the most widely used. Other variants (in which, e.g., larger connected sets of spins are flipped together) are also often studied.

6.3.4 The Metropolis Process

The Metropolis process [MRR$^+$53] gives a general recipe for constructing a Markov chain with *any* desired stationary distribution. Recall that we are given a set of configurations Ω and a strictly positive weight

function $w : \Omega \mapsto \mathbb{R}^+$, and our goal is to sample from the distribution $\pi(x) = \frac{w(x)}{Z}$ over Ω, where Z is an unknown normalization constant.

The Metropolis process consists of two ingredients:

Neighborhood structure: 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 (e.g., in the case of a spin system, by a single spin flip). We use the notation $x \sim y$ to denote that x, y are neighbors.

Proposal distribution: For each $x \in \Omega$, the "proposal distribution" is a probability distribution $\kappa(x, \cdot)$ on the set $N(x) \cup \{x\}$, where $N(x)$ is the set of neighbors of x in the graph, such that $\kappa(x, y) > 0$ for all $y \in N(x)$. (A common choice is to let $\kappa(x, y) = \kappa(y, x) = \frac{1}{\max\{d(x), d(y)\}}$, where $d(x), d(y)$ are the degrees of x, y in the graph, with any remaining probability allocated to the self-loop.)

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 $\min\left\{1, \frac{w(y)\kappa(y,x)}{w(x)\kappa(x,y)}\right\}$, else stay at x.

The reason for the term "proposal distribution" for κ is now clear. We "propose" a new, neighboring state according to the distribution induced by κ on the current state, and then move to this state with a probability that depends on the stationary distribution to which we want the process to converge. Notice that the actual transition probabilities are $P(x, y) = \kappa(x, y) \cdot \min\left\{1, \frac{\pi(y)\kappa(y,x)}{\pi(x)\kappa(x,y)}\right\}$. 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 $\pi(y)/\pi(x)$).

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. By the Fundamental Theorem, this implies that π is its unique stationary distribution.

Claim 6.13. *The Metropolis process 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 all pairs of neighbors x, y. Assume without loss of generality that $w(y)\kappa(y, x) \leq w(x)\kappa(x, y)$. Then

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

□

The Metropolis process defines a huge family of reversible Markov chains for any given distribution π. In any given application, there is still much work to be done in defining a suitable neighborhood structure (and, usually less important, the associated proposal distributions).

6.3.5 The Swendsen-Wang dynamics

Our last example is special to the ferromagnetic Ising and Potts models, and is included as an example of a *non-local* Markov chain, i.e., one that changes large portions of the configuration in one step. This chain is widely used and studied in the statistical physics community. Recall that a configuration of the Potts model on a graph $G = (V, E)$ is an assignment $\sigma : V \to [q]$, and the weight of a configuration is λ^k, where $k := k(\sigma)$ is the number of bichromatic (disagreeing) edges in σ and $\lambda = \exp(-2\beta) \in (0, 1]$. The Ising model is just the case $q = 2$.

In configuration σ, the Swendsen-Wang dynamics makes the following transition:

1. Delete all edges that are bichromatic in σ.
2. For each monochromatic edge e in σ, retain e with probability $p := 1 - \lambda$ and delete it otherwise.
3. Assign to each connected component in the remaining graph a spin chosen independently and u.a.r. from $[q]$.

This chain is irreducible (e.g., we could remove all the edges in step 2 and then make any spin assignment in step 3) and aperiodic (self-loops exist). It's slightly non-trivial to show that it's also reversible w.r.t. the Potts Gibbs measure $\mu(\sigma) = \frac{\lambda^{k(\sigma)}}{Z}$. (This is a good **exercise**; you should enumerate over the possible sets of edges remaining after step 2 in going from σ to σ', and vice versa. Alternatively, see [ES88].)

References

[BD92] D. Bayer and P. Diaconis. Trailing the dovetail shuffle to its lair. *Annals of Applied Probability*, 2:294–313, 1992.

[ES88] R.G. Edwards and A.D. Sokal. Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Review D*, 38:2009–2012, 1988.

[Gil55] E. Gilbert. *Theory of shuffling.* Technical Memorandum, Bell Laboratories, 1955.

[MRR+53] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[Ree81] J. Reeds. Unpublished manuscript. 1981.

[Sen06] E. Seneta. *Non-negative matrices and Markov chains.* Springer-Verlag, New York, 2006.