CSE525: Randomized Algorithms and Probabilistic Analysis June 6, 2013

Lecture 25

Lecturer: Anna Karlin *Scribe: Daryl Hanson & Jerry Li*

1 Introduction

For all $\epsilon, \delta > 0$, we say that a randomized algorithm gives an (ϵ, δ) approximation for a value V if the output X of the algorithm satisfies

$$\Pr(|X - V| > \epsilon|V|) < \delta.$$

Here we will discuss Monte Carlo methods, which are a collection of tools for obtaining such approximations through sampling and estimation. A typical Monte Carlo method runs as follows: we sample i.i.d. random variables $X_1, \ldots, X_m$ whose mean $\mu = E(X_i)$ is the quantity that we desire. If $m \geq 3\log(2/\delta)/(\epsilon^2\mu)$ then by Chernoff bounds,

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geq \epsilon\mu\right) \leq \delta.$$

Notice, however that if μ is small then m might have to be very large–hence when constructing Monte Carlo algorithms we should try to make μ as large as possible. We illustrate with an example.

2 DNF Counting

Given a DNF formula φ with n variables, DNF counting is the problem finding the number of satisfying assignments for φ. Note that in general satisfiability for DNF is easy as we need only satisfy a single clause, but the counting problem is hard. Indeed, if we could do this, then given any 3-CNF formula f with n variables, we could take its negation, count how many satisfying assignments its negation has, and if that was equal to 2^n then $\neg f$ is a tautology so f is unsatisfiable, otherwise, it is satisfiable, so we have solved 3-SAT. The counting problem is hard for a class called $\#P$, which is a very strong form of intractability. We seek to find approximate solutions to this problem.

The obvious approach is the following: sample random assignments uniformly at random, and take the estimate to be $p2^n$ where p was the fraction of assignments that satisfied φ. Let X_i be the random variable which is 1 if the ith assignment we chose satisfied F, zero otherwise, and suppose we sample m times. Then $p = \frac{1}{m}\sum_{i=1}^{m} X_i$. Let μ be true fraction of satisfying assignments. Then for fixed $\epsilon, \delta > 0$ if we want

$$\Pr(|X - \mu| \geq \epsilon\mu) \leq \delta$$

then using only Chernoff bounds as above would require $m = \Omega(1/\mu)$. However, in general, μ can be exponentially small: for instance, if there is only one satisfying assignment then $\mu = 2^{-n}$, so the obvious approach may take exponential time to get nice probabilistic guarantees.

Instead, we will do the following. Write

$$\phi = C_1 \vee C_2 \vee \ldots C_t$$

and let SC_i be the set of assignments that satisfy clause i. We wish to approximate $S = |\cup_i SC_i|$. We do so as follows: let

$$U = \{(i, x) : 1 \leq i \leq t, x \in SC_i\}.$$

It is easy to count $|SC_i|$, as if C_i contains i variables then $|SC_i| = 2^{n-i}$. As $|U| = \sum_{i=1}^{t} |SC_i|$, it is easy to count $|U|$, so to approximate S we need only to find out approximately how much we repeat ourselves in U. More rigorously, let

$$X = \{(i, x) : 1 \leq i \leq t, x \in SC_i, x \notin SC_j, j \leq i - 1\}.$$

Then clearly $|X| = S$, and we will estimate $|X|$ be approximating $|X|/|U|$. To do this, we will sample elements (i, x) from U uniformly at random and determine if $(i, x) \in X$. Given a random sample $(i, x) \in U$, it is not hard to check if $(i, x) \in X$: simply check whether or not x satisfies C_j for some $j < i$, which take polynomial time. Importantly, we claim that $|X|/|U| \geq 1/t$, as trivially, every satisfying assignment must satisfy some clause, so when we do the same Chernoff bound analysis as before, to get an (ϵ, δ) approximation it will suffice to take $3t \log(2/\delta)/\epsilon^2$ samples, so assuming we can sample from U, this whole procedure will take polynomial time.

Thus it suffices to pick samples uniformly at random from U. To do so, for each $1 \leq i \leq t$ we pick i with probability $\frac{|SC_i|}{|U|}$, then we randomly pick an element x of $|SC_i|$. The latter is easy to do because it amounts to flipping $n - j$ coins, where j is the number of variables present in C_i. Then

$$\Pr((i, x) \text{ is selected}) = \frac{|SC_i|}{U} \frac{1}{|SC_i|} = \frac{1}{U}$$

so we have found a way to sample uniformly form U, so we are done.

3 Independent Sets

Suppose we have a graph $G = (V, E)$. Order its edges arbitrarily, and label them $E = \{e_1, \ldots, e_m\}$.
Define a sequence of graphs $G_i = (V, E_i)$ where $E_i = \{e_j | j \leq i\}$.
Notice that $G_0 = (V, \phi)$ and $G_m = G$.
Next, define $\Omega(G_i)$ to be the set of Independent Sets in G_i.

$$|\Omega(G)| = \frac{|\Omega(G_m)|}{|\Omega(G_{m-1})|} \times \frac{|\Omega(G_{m-1})|}{|\Omega(G_{m-2})|} \times \ldots \times \frac{|\Omega(G_1)|}{|\Omega(G_0)|} \times |\Omega(G_0)|$$

To estimate $|\Omega(G)|$, we just need to have a good estimate for

$$r_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}$$

Lemma 1. *For all $i, r_i \geq \frac{3}{4}$.*

Proof Write $e_i = (u, v)$. Then the only way we lost an I.S. I from G_{i-1} to G_i is if both u aand v are in I. However, if $I \in \Omega(G_{i-1})$, then so are $I \setminus \{u\}$, $I \setminus \{v\}$, and $I \setminus \{u, v\}$. All three of these are also members of $\Omega(G_i)$, so for any I.S. lost, there are at least three others that remain, and so $r_i \geq \frac{3}{4}$. ■

Now we present an algorithm for estimating r_i, given that we have an algorithm for generating random samples (which will be presented later):

1. $X = 0$
2. Repeat M times: *(for $M = \frac{3m^2}{\epsilon^2} \ln(\frac{2m}{\delta})$)*
   - generate uniform sample from $\Omega(G_{i-1})$
   - if sample is an independent set in G_i, increment X.
3. return $\tilde{r}_i = \frac{X}{M}$.

By the Chernoff bound discussed previously, since r_i is large, $\tilde{r}_i$ is a $(\frac{\epsilon}{2m}, \frac{\delta}{m})$ approximation of r_i.
Our estimate of $|\Omega(G)|$ is equal to

$$2^n \prod_{i=1}^{m} \tilde{r}_i$$

while the true value is

$$2^n \prod_{i=1}^{m} r_i.$$

We claim the two are close, with high probability.

Claim 2.

$$\Pr\left(|\prod_{i=1}^{m} \frac{\tilde{r}_i}{r_i} - 1| \leq \epsilon \right) \geq 1 - \delta$$

Proof For each individual i, we have

$$Pr\left(|\tilde{r}_i - r_i| \leq \frac{\epsilon}{2m} r_i \right) \geq 1 - \frac{\delta}{m}$$

or equivalently,

$$Pr\left(r_i(1 - \frac{\epsilon}{2m}) \leq \tilde{r}_i \leq r_i(1 + \frac{\epsilon}{2m}) \right) \geq 1 - \frac{\delta}{m}$$

or

$$Pr\left(1 - \frac{\epsilon}{2m} \leq \frac{\tilde{r}_i}{r_i} \leq 1 + \frac{\epsilon}{2m} \right) \geq 1 - \frac{\delta}{m}$$

so by the union bound,

$$Pr\left(1 - \epsilon \leq (1 - \frac{\epsilon}{2m})^m \leq \prod_{i=1}^{m} \frac{\tilde{r}_i}{r_i} \leq (1 + \frac{\epsilon}{2m})^m \leq 1 + \epsilon \right) \geq 1 - \delta$$

and we are done. ■

To estimate $|\Omega(G)|$, all that remains is to figure out how to generate a uniform sample from $\Omega(G_{i-1}) \forall i$. Additionally, an *almost* uniform sample suffices – the extra error can be carried through the previous ϵ/δ proof.

3.1 Sampling

The current goal is to sample elements from a universe Ω according to some distribution π.
One cool way to do so is to design a Markov chain whose state space is Ω that has stationary distribution π. Then, we can simulate this Markov chain until it "mixes", and use the state at that time as a sample. Notice that in the Independent Set case, the Markov chain has $|\Omega|$ states, which is exponential in $|G|$, so we need a chain with a logarithmic mixing time (e.g. an expander graph).

The two key questions related to this are:

1. How do we design a chain with the right distribution?
2. How do we bound the mixing time?

In the case of Independent Sets, construct the Markov chain as follows:
The states are the independent sets of G, and X_t is some independent set.
To transition, choose a vertex v uniformly at random from V:

- if $v \in X_t$ then $X_{t+1} = X_t \setminus v$
- if $v \notin X_t$ and $X_t \cup v$ is an independent set of G, then $X_{t+1} = X_t \cup v$
- otherwise, $X_{t+1} = X_t$

A few things to observe: First, the graph is irreducible, since to get from I to I', it is possible to first remove every vertex in I and then to add in each vertex in I'. If there exists an edge, then it is aperiodic, as it is possible to stay in the same state. Finally, the chain is doubly stochastic, which implies that the stationary distribution is uniform over all independent sets.

Let us now check that P is doubly stochastic. First, notice that the transition from X_t to X_{t+1} is deterministic once we have chosen v. Then, notice that if we have transitioned to state X_{t+1} by choosing vertex v, there is exactly one state X_t that we could have come from: if $v \in X_{t+1}$, then it must have been added to $X_{t+1} \setminus v$. If $v \notin X_{t+1}$ and it could safely have been added, it must have just been removed from $X_{t+1} \cup v$. If it's not in X_{t+1} and adding it would break the constraints, then it was a self loop from X_{t+1}. So if $X_{t+1} = j$, then for each of the n vertices there was a $\frac{1}{n}$ contribution to some P_{ij} and no contribution to any of the others. So $\sum_i P_{ij} = 1$.

Let's consider a general technique to find such a chain. Given a state space Ω and a connected graph on Ω, we need to define transition probabilities so that we will have a stationary distribution matching a target π.

3.2 Metropolis Algorithm

As input, we take a state space Ω, a connected graph $G = (\Omega, E)$, and a π such that $\sum_{i\in\Omega} \pi_i = 1$.

Let Δ be the maximum degree in the graph.

$$P_{xy} = \begin{cases} \frac{1}{2\Delta}\min(1, \frac{\pi_y}{\pi_x}) & x \neq y, y \in N(x) \\ 0 & x \neq y, y \notin N(x) \\ 1 - \sum_{y\neq x} P_{xy} & x = y \end{cases}$$

One nice property is that defining P only depends on the ratios of π, not on π_x or π_y in isolation. In some circumstances, π is only known up to proportionality, and it's not easy to get the normalizing factor, but that does not affect this algorithm.

Claim 3. *For all $x, y, \pi_x P_{xy} = \pi_y P_{yx}$.*

Proof If $x \neq y$, assume without loss of generality that $\pi_x \leq \pi_y$. Then $\pi_x(\frac{1}{2\Delta}) = \pi_y(\frac{1}{2\Delta}\frac{\pi_x}{\pi_y})$. by explicit calculation. ■

This claim implies that π is a stationary distribution for our Markov chain.

As an example, suppose we wanted to sample the independent sets according to the distribution $\pi(I) = \frac{\lambda^{|I|}}{Z}$, where $Z = \sum_{I'} \lambda^{|I'|}$.
Then by the algorithm above, the probability matrix is defined as

$$P_{I,I'} = \frac{1}{2n}\min(1, \lambda).$$

We now wish to show that the metropolis algorithm can be run in polynomial time; that is, we only need to run the simulation for polynomially many steps before we get ϵ-close to the stationary distribution. We do so in the next lecture, using the techniques we develop below.

4 Mixing Times and Couplings

4.1 Motivation

We have seen before that there are many important properties of Markov chains that relate to how fast a starting distribution converges to the stationary distribution, including the spectral gap and various conductance properties. Analysis of these invariants often produced upper bounds on how fast the Markov chain could mix, and vice versa. Often times (like in the Metropolis algorithm), the reason why we are interested

in a Markov chain is because we to approximate the stationary distribution, and so the technique is to simply start the Markov chain somewhere, and let it run for a while. While we know that eventually the distribution will converge to the stationary distribution, we would like some way to know concretely how many steps are needed. In particular, we hope that only polynomially many steps are needed.

As an aside, apparently there are important connections to the hardcore lattice gas model in statistical physics, whatever that means.

4.2 Total Variation and Mixing Times

We introduce here a technique to bound the mixing time of finite, aperiodic, irreducible Markov chains. Intuitively, the mixing time of a Markov chain is how long it takes for the Markov chain to approach the unique stationary distribution.

Definition 4. *Let D_1, D_2 be two probability distributions over some finite sample space Ω. Then the total variation of D_1 and D_2 is*

$$\|D_1 - D_2\|_{TV} = \frac{1}{2}\sum_{x\in\Omega} |D_1(x) - D_2(x)|.$$

The total variation of two distributions is a measure of how far they differ in the worst case. This is made rigorous in the lemma below.

Lemma 5. *For all D_1, D_2,*

$$\|D_1 - D_2\|_{TV} = \max_{A\subset\Omega} |D_1(A) - D_2(A)|.$$

Proof Let $A_1 = \{x \in \Omega : D_1(x) \geq D_2(x)\}$ and $A_2 = A_1^c$. Then

$$\begin{aligned}
\|D_1 - D_2\|_{TV} &= \frac{1}{2}(D_1(A_1) - D_2(A_1) + D_2(A_2) - D_1(A_2)) \\
&= \frac{1}{2}[D_1(A_1) - D_2(A_1) + 1 - D_2(A_1) - (1 - D_1(A_1))] \\
&= D_1(A_1) - D_2(A_1).
\end{aligned}$$

Similarly

$$\|D_1 - D_2\|_{TV} = D_2(A_2) - D_1(A_2)$$

so

$$\|D_1 - D_2\|_{TV} \leq \max_{A\subset\Omega} |D_1(A) - D_2(A)|.$$

For the other direction, for any $A \subset \Omega$, if $D_1(A) - D_2(A) \geq 0$ notice that $D_1(A) - D_2(A) \leq D_1(A_1) - D_2(A_2)$ and otherwise $D_2(A) - D_1(A) \leq D_2(A_2) - D_1(A_2)$, so the maximum of $|D_1(A) - D_2(A)|$ is obtained at these two sets, so

$$\|D_1 - D_2\|_{TV} \geq \max_{A\subset\Omega} |D_1(A) - D_2(A)|$$

and we are done. ■

Definition 6. *For any Markov chain over states Ω with stationary distribution π, and for any initial distribution P and $\epsilon > 0$, the mixing time is*

$$\tau_P(\epsilon) = \min\{t : \|P^s, \pi\|_{TV} \leq \epsilon, s \geq t\}$$

where P^t is the distribution after t steps of the Markov chain.

That is, the mixing time is the first time such that the actual distribution becomes ϵ-close to the stationary, in the total variation sense. We say that the Markov chain is rapidly mixing if $\tau(\epsilon)$ is asymptotically polynomial in $\log|\Omega|$ and $\log(\epsilon^{-1})$.

4.3 Coupling

Coupling is a simple and elegant technique that sees a variety of application in probability. Here we will use it to bound mixing times.

Definition 7. *Given a Markov chain on Ω, a coupling is a Markov chain on $\Omega \times \Omega$ defining a stochastic process (X_t, Y_t) so that*

1. *X_t and Y_t alone are faithful copies of the original Markov chain; that is, there are initial conditions X_0' and Y_0' on the original Markov chain defining stochastic processes X_t' and Y_t' so that*

$$\Pr((X_t, Y_t) \in A \times \Omega) = \Pr(X_t' \in A)$$

and

$$\Pr((X_t, Y_t) \in \Omega \times A) = \Pr(Y_t' \in A)$$

for all $A \subset \Omega$.

2. *If $X_t = Y_t$ then $X_{t+1} = Y_{t+1}$ pointwise.*

Intuitively, a coupling is when two walks in the Marko chain walk together, in some sense.

Example. Consider the random walk on the hypercube $\mathbb{F}_2^n$. At each step, we choose a random coordinate i and a random bit $b \in \{0, 1\}$ and we change the ith bit to b. Given two starting positions x, y we may construct a coupled walk; that is, consider the Markov chain over $\mathbb{F}_2^n \times \mathbb{F}_2^n$ where given two states (a, b), we transition by choosing a random coordinate and a random bit as above, then altering both a and b in the way described above, so that the two walks transition together. If we take the starting position of this new Markov chain to be (x, y) then the resulting stochastic process (X_t, Y_T) is clearly a coupling: as each marginal transitions according to the original Markov chain, alone both X_t and Y_t must be walks along the original Markov chain, and clearly since we transition together if $X_t = Y_t$ then $X_{t+1} = Y_{t+1}$. We will use this coupling, along with the lemma we present next, to bound the time it takes for two random walks on the hypercube to meet.

Lemma 8 (Coupling Lemma)**.** *Let $Z_t = (X_t, Y_t)$ be a coupling where $Y_0 = \pi$ and $X_0 = X$, where X is some arbitrary distribution. Suppose there exists a T so that,*

$$\Pr(X_T \neq Y_T | X_0 = X) \leq \epsilon.$$

Then the mixing time starting at X is bounded by T, that is, $\tau_X(\epsilon)$.

Proof Suppose X started at some arbitrary X_0. For all $A \subseteq \Omega$,

$$\begin{aligned}
\Pr(X_T \in A) &= \Pr(X_T = Y_T \cap Y_T \in A) + \Pr(X_t \neq Y_T \cap X_T \in A) \\
&\geq \Pr(X_T = Y_T \cap Y_T \in A) \\
&= 1 - \Pr(X_t \neq Y_T \cup Y_T \notin A) \\
&\geq 1 - \Pr(Y_T \notin A) - \Pr(X_t \neq Y_T) \geq \pi(A) - \epsilon
\end{aligned}$$

where the fourth inequality follows from the law of total expectation. Similarly, $\Pr(X_t \notin A) = \Pr(X_T \in A^c) \geq \pi(A^c) - \epsilon$ so $\Pr(X_T \in A) \leq \pi(A) + \epsilon$ and so we are done. ■

Examples to follow next lecture.