CSE525: Randomized Algorithms and Probabilistic Analysis June 4, 2013

Lecture 17

Lecturer: Anna Karlin *Scribe: Eric Lei and Tanner Schmidt*

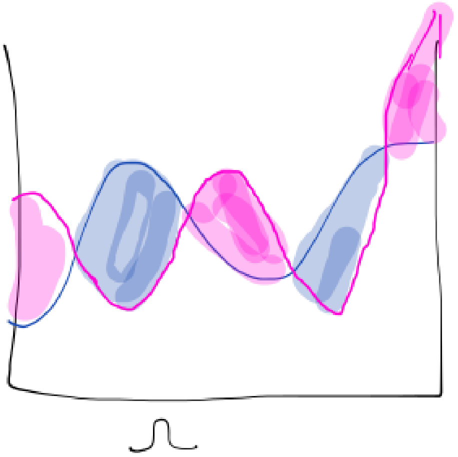
1 Coupling

1.1 Definitions and Notation

Coupling is a useful tool in the analysis of the mixing time of Markov chains. The basic idea is that a Markov chain that is initialized to some arbitrary distribution can be compared via coupling with another Markov chain that is initialized to the stationary distribution. The two chains then progress simultaneously, and the distance between the two chains at any time indicates how close the randomly initialized distribution is to the stationary distribution.

The *total variation distance* between two distributions D_1 and D_2 on the same sample space Ω is defined as

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



$||D_1 - D_2||_{TV}$ = pink area = blue area

One common definition of mixing time using this definition is

$$\tau(\epsilon) = \min\{t \,|\, ||P^t - \pi||_{\text{TV}} \leq \epsilon\}.$$

We can then say a Markov chain is rapidly mixing if $\tau(\epsilon)$ is polynomial in $\log(|\Omega|)$ and $\log(\frac{1}{\epsilon})$. This is also related to the spectral gap of transition matrix P.

Coupling is a simple and elegant approach to bounding mixing times. Given a Markov chain on Ω, a coupling is a Markov chain on $\Omega \times \Omega$ defining a stochastic process (X_t, Y_t) such that:

1. each X_t and Y_t in isolation is a faithful copy of the Markov chain
2. if $X_t = Y_t$, then $X_{t+1} = Y_{t+1}$

Lemma 1 (Coupling Lemma)**.** *Let $Z_t = (X_t, Y_t)$ be a coupling.*
Suppose $\exists T$ such that $\forall x, y : Pr(X_T \neq Y_T | X_0 = x, Y_0 = y) \leq \epsilon$. Then $\tau(\epsilon) \leq T$.

Proof Consider a coupling with Y_0 chosen according to π, with an arbitrary x_0.

$$\begin{aligned}\forall A \subset \Omega : \Pr(X_T \in A) &\geq \Pr(X_T = Y_T \wedge Y_T \in A)\\ &= 1 - \Pr(X_T \neq Y_T \vee Y_T \notin A)\\ &\geq 1 - \Pr(Y_T \notin A) - \Pr(X_T \neq Y_T)\\ &= \Pr(Y_T \in A) - \Pr(X_T \neq Y_T)\\ &= \pi_A - \epsilon\end{aligned}$$

Similarly, $\Pr(X_T \notin A) \geq \pi_{\Omega \setminus A} - \epsilon$ and $\Pr(X_T \in A) \leq \pi_A + \epsilon$. ■

1.2 Examples

1.2.1 Random walk on hypercube

Imagine a random walk on a hypercube in $\mathbb{R}^n$, with nodes at the vertices. There are therefore $N = 2^n$ nodes. At each step, choose a random coordinate i, and a random bit $b \in \{0, 1\}$, and then change the ith bit to b.

Let's define a random variable $X_t \in \mathbb{R}^n$ to be the coordinates of the random walk after t steps in the random walk. Now imagine coupling this with another random variable Y_t, which represents another random walk which at each step uses the same i and b as X_t. If one walk starts at the stationary distribution, and the other starts at some arbitrary position, how long will it take for the walks to meet?

1.2.2 Independent sets

Assume we have a graph $G = (V, E)$, with the sample space Ω defined to be the set of all independent sets of size k in G. The problem is to generate an independent set uniformly at random. We do so by constructing a MCMC sampler, and we must prove that it mixes rapidly. Consider a random walk over Ω, represented by random variable X_t which is the independent set at time t. The walk is defined by the following process:

- choose a vertex $v \in X_t$ uniformly at random and a vertex $w \in V$ uniformly at random
- if $w \notin X_t$ and $X_t - v + w$ is independent, $X_{t+1} = X_t - v + w$
- otherwise, $X_{t+1} = X_t$

In other words, at each step, we pick a random vertex that is already in the set and swap it with a randomly selected vertex in the graph if the result is also an independent set.

Claim 2. *This Markov chain mixes rapidly if $k \leq \frac{n}{3\Delta+3}$, where n is the number of vertices and δ is the maximum degree of the graph.*

Proof For a random walk defined by X_t that is initialized arbitrarily, couple it with another random walk defined by Y_t that starts in the stationary distribution π. Then consider an arbitrary bijection $f(v) : X_t \to Y_t$, i.e. a matching between vertices in X_t and Y_t. This could, for example, be determined by a random permutation of V at the beginning of execution. Then at each time step, pick $w \in V$ uniformly at random and $v \in X_t$ uniformly at random, and update X_t as before. If $v \in Y_t$, set $Y_{t+1} = Y_t - v + w$ if the result is an independent set and $Y_{t+1} = Y_t$ otherwise. If $v \notin Y_t$, then use $f(v)$ and set $Y_{t+1} = Y_t - f(v) + w$ if the result is an indpendent set, and leave it unchanged otherwise.

Define the distance between the walks as $d_t = |X_t - Y_t|$, and note that it can change by at most 1 in each step, but also that it can go up as well as down. The chains will have met when $d_t = 0$. This process is then similar to a random walk on a line, and if we can show that the corresponding walk of d_t on the line has negative drift, we can claim that this won't take too long.

Let's first analyze the probability of increasing the distance when the paths have not met yet. This can only happen when $v \in X_t$ and $v \in Y_t$, which allows for one of the sets to swap the vertex v while the other does not. The probability that $v \in X_t$ and $v \in Y_t$ is the probability that a random vertex $v \in X_t$ is also in Y_t, which is $\frac{k-d_t}{k}$. Even when this is the case, the distance will only increase if w is in the neighborhood of either X_t or Y_t (and therefore will swap with one but not the other) of if w is in X_t but not Y_t or vice versa, which has probability $\frac{2d_t(\Delta+1)}{n}$. Therefore

$$\Pr(d_{t+1} = d_t + 1 \,|\, d_t > 0) \leq \left(\frac{k-d_t}{k}\right)\left(\frac{2d_t(\Delta+1)}{n}\right).$$

For the distance to decrease, we again need two conditions. The first is that $v \notin Y_t$, which happens with probability $\frac{d_t}{k}$. Then, we need for w to be outside of the neighborhood of both X_t and Y_t, and not in X_t or Y_t. Then both sets will swap a vertex that they do not share with w, which they do share, and decrease the distance by one. The second condition has probability greater than or equal to $\frac{2d_t(\Delta+1)}{n}$. Therefore:

$$\Pr(d_{t+1} = d_t - 1 \,|\, d_t > 0) \geq \left(\frac{d_t}{k}\right)\left(\frac{2d_t(\Delta+1)}{n}\right).$$

Putting these two together forms our expectation:

$$\begin{aligned} E(d_{t+1} \,|\, d_t) &= \Pr(d_{t+1} = d_t + 1)(d_t + 1) + \Pr(d_{t+1} = d_t - 1)(d_t - 1) \\ &\leq d_t\left(1 - \frac{n - (3k-3)(\Delta+1)}{kn}\right) \\ &= d_t\alpha \end{aligned}$$

where α is simply defined to be the entire term in the parenthesis. Noting that $\alpha < 1$ for $k \leq \frac{n}{3\Delta+3}$ means that in this case

$$\lim_{t\to\infty} E(d_t) \leq \lim_{t\to\infty} d_0\alpha^t = 0$$

and

$$\lim_{t\to\infty} \Pr(d_t \geq 1) \leq \lim_{t\to\infty} E(d_t) = 0.$$

■

2 Martingales

This section introduces martingales, a class of stochastic process capturing the idea of a fair game. Their study originated from gambling theory and has many useful results.

2.1 Motivating Example: Coin-Flipping

One problem for which martingales are especially relevant is computing the expected number of flips of a fair coin until a given sequence is observed. For instance, how many flips would it take on average to observe HH or HTH? These values can be computed directly, but martingales provide a more elegant and general solution. This section demonstrates the use of martingales in order to provide direction before we give definitions.

Consider the sequence $S = HTH$. Let C_t denote the result of the flip at time t. Imagine that before each flip, a new gambler arrives and bets \$1 that $C_t = H$. If he loses, he his net loss is \$1, and he stops playing. If he wins, he bets his \$2 winnings that $C_{t+1} = T$, which would mean that the next element of S is observed. Again, losing means his net loss is \$1. If he wins, he bets his total winnings on observing H next. In the next step, he either ends with a net loss of \$1, or wins an \$8 bet and ends with a net win of \$7. Remember, a new gambler is introduced at each t, creating a sequence of gamblers at different states in their betting strategies.

Let τ be the time at which S is first observed. Let X_t denote the net profit of all gamblers introduced up to time t. After the more rigorous next section, we will be able to verify that X_t is a martingale. This fact lets us apply the intuition[1] that X_t is "fair"; i.e., $E[X_\tau] = 0$, meaning that a gambler can neither win nor lose on average. When $t = \tau$, exactly one person–the third most recent one—will have won his \$8 bet; another—the most recent one—will have just won his \$2 bet because the last flip was H. Everyone else must have lost a bet at some point because there are no other heads in S. In total, the gains are $8 + 2$, and all τ people paid 1. Therefore,

$$0 = E[X_\tau] = E[8 + 2 - \tau] \implies E[\tau] = 10$$

2.2 Definitions and Notation

First we define a special case of martingales. A stochastic process $\{X_t\}$ is a *martingale* if for all t,

$$E[X_{t+1}|X_0, \dots, X_t] = X_t$$

An interpretation would be that given the history of the process, one can make no guess as to whether it will increase or decrease.

We give a simple example. Consider a gambler with initial wealth X_0 who repeatedly plays a fair game. Then X_t, her wealth at time t, is a martingale because the expected winnings from each game are 0. This says $E[X_{t+1}|X_0, \dots, X_t] - X_t = 0$.

In general, $\{X_t\}$ is a martingale[2] with respect to another stochastic process $\{Y_t\}$, where $X_t = f(Y_0, \dots, Y_t)$ for some function f, if

$$E[X_{t+1}|Y_0, \dots, Y_t] = X_t$$

When X_t is said to be a martingale, and there is no mention of Y_t, then it is assumed that $\{X_t\} = \{Y_t\}$. We now make one change to notation. For a martingale defined with respect to $\{Y_t\}$, instead of writing $Y_0, \dots, Y_t$ in the conditional probabilities, we write F_t.

Another relevant definition is stopping time. A random variable $\tau \in \mathbb{Z}_{\geq 0}$ is a *stopping time* with respect to $\{Y_t\}$ if for all t, we know whether the event $\{\tau = t\}$ occurs if we observe $Y_0, \dots, Y_t$.

2.3 Common Examples

2.3.1 Sums of i.i.d. Random Variables

Let $Y_0 = 0$, and for all $t \geq 1$ let Y_t be distributed i.i.d. with $E[Y_t] = 0$. For all $t \geq 0$, define $X_t = \sum_{k=1}^t Y_k$. To check that $\{X_t\}$ is a martingale with respect to $\{Y_t\}$, we have

$$E[X_{t+1}|F_t] = E[X_t + Y_{t+1}|F_t] = X_t + E[Y_{t+1}|F_t] = X_t + E[Y_{t+1}] = X_t$$

2.3.2 Variance of a Sum

Let $\{Y_t\}$ be defined as in the previous section. Let $\sigma^2 = E[Y_t^2]$. For all $t \geq 0$, define $X_t = (\sum_{k=1}^t Y_k)^2 - n\sigma^2$. To check that $\{X_t\}$ is a martingale with respect to $\{Y_t\}$, we have

$$E[X_{t+1}|F_t] = E[(Y_{t+1} + \sum_{k=1}^t Y_k)^2 - (n+1)\sigma^2|F_t]$$

$$= E[Y_{t+1}^2 + 2Y_{t+1}\sum_{k=1}^t Y_k + (\sum_{k=1}^t Y_k)^2 - (n+1)\sigma^2|F_t]$$

[1] This intuition will be formalized by the optional sampling theorem.

[2] A rigorous definition of martingale would involve σ-algebras and filtrations, which you are encouraged to learn.

$$= X_t + E[Y_{t+1}^2] + 2E[Y_{t+1}] \sum_{k=1}^{t} Y_k - \sigma^2$$

$$= X_t + \sigma^2 + 0 - \sigma^2 = X_t$$

2.3.3 Doob Martingale

This example is useful yet tricky. Let $\{Y_t\}$ be an arbitrary sequence of random variables. Let X be a random variable with $|E[X]| < \infty$. Let $X_t = E_t[X|F_t]$, where the subscript t denotes expectation over all Y_k with $k > t$. This process is called a *Doob martingale* or *Levy martingale.*

To check that $\{X_t\}$ is a martingale with respect to $\{Y_t\}$, recall the *law of iterated expectation*,

$$E_{XY}[X] = E_Y[E_X[X|Y]]$$

which says that on the LHS the outer expectation over B averages out the condition on B. A consequence is

$$E_B[E_A[f(A,B,C)|B,C]|C] = E_{AB}[f(A,B,C)|C]$$

Therefore,

$$E_t[X_{t+1}|F_t] = E_t\big[E_{t+1}[X|F_{t+1}]|F_t\big]$$

$$= \underset{Y_{t+1},Y_{t+2},\dots}{E} \left\{ \underset{Y_{t+2},\dots}{E} \big[X|Y_{t+1}, F_t\big]|F_t \right\}$$

$$= \underset{Y_{t+1}}{E} \left\{ \underset{Y_{t+2},\dots}{E} \big[X|Y_{t+1}, F_t\big]|F_t \right\}$$

$$= E_t[X|F_t] = X_t$$

In the above, we first use the definition of X_{t+1}. Then we explicitly write out which variables the expectations are over. Next we drop redundant variables from the outer expectation. Lastly we invoke the law of iterated expectation.

An application would be to random graphs $G(n,p)$ in which there are n vertices, and each possible edge is present with probability p. Let the present edges be denoted by $e_1, \dots, e_m$. Let X be some graph property, such as chromatic number. Let Y_t be the indicator of the presence of e_t, so that F_t means we have observed whether $e_1, \dots, e_t$ are present. Then X_t is the expected chromatic number after observing whether the first t edges are present. Now X_t is called an *edge exposure martingale.* We also see that $X_0 = E[X]$ and $X_m = X$. These facts can be combined to lead to interesting results.

2.4 Martingale Facts

Fact 3. *For all t, $E[X_t] = E[X_0]$, where $E[\cdot]$ is expectation with respect to all Y_k.*

Proof Let $E_t[\cdot]$ denote expectation with respect to Y_k for all $k > t$. We induct on t:

$$E_t[X_{t+1}|F_t] = X_t$$

$$\Longrightarrow E[X_{t+1}] = E[E_t[X_{t+1}|F_t]] = E[X_t] = E[X_0]$$

where we use the law of iterated expectation. ■

Fact 4. *Optional Sampling Theorem (a.k.a. Optional Stopping Theorem). Let $\{X_t\}$ be a martingale w.r.t. $\{Y_t\}$. Let τ be a stopping time. Then $E[X_\tau] = E[X_0]$, where the expectation is over all Y_k, if at least one of the following holds almost surely (with probability 1):*

- $|X_t| < \infty \ \forall t$

- $|\tau| < \infty$
- $E[\tau] < \infty$, $E[|X_{t+1} - X_t||F_t] < \infty$

Remark The proof of the OST is beyond the scope of this class. However, the result is highly intuitive and is basically a stronger version of martingales being "fair" no matter what strategy is played.

2.5 Applications of Optional Sampling Theorem: Symmetric Random Walk

Consider the symmetric random walk $\{X_t\}$ on the line. Given $a, b > 0$, let $\tau_a = \min\{t : X_t = -a\}$ be the first time at which $-a$ is hit. Let $\tau_b = \min\{t : X_t = b\}$ be the first time at which b is hit. We will compute $p_a = \Pr(\tau_a < \tau_b)$.

First, we formally define X_t. Let Y_t equal 1 or -1 with equal probability. Now the random walk $X_t = \sum_{k=1}^{t} Y_k$ is a martingale. Let $\tau = \min(\tau_a, \tau_b)$.

In this class, we must take on faith that $\tau < \infty$ almost surely.[3] By the OST,

$$0 = E[X_0] = E[X_\tau] = p_a(-a) + (1 - p_a)b \implies p_a = \frac{b}{a+b}$$

Next, we will compute $E[\tau]$. Let $Z_t = X_t^2 - n$. Then Z_t is a martingale because

$$E[Z_{t+1} - Z_t|F_t] = E[X_{n+1}^2 - 1 + Z_n - Z_n|F_t] = E[X_{n+1}^2 - 1] = 0$$

By the OST,

$$0 = E[Z_0] = E[Z_\tau] = (p_a a^2 + (1-p)b^2) - E[\tau] \implies E[\tau] = ab$$

[3] The reason is that X_t is a *recurrent* and *irreducible* Markov chain.