CSC473: Advanced Algorithm Design Winter 2019

Week 5: Random Walks and Markov Chains

Aleksandar Nikolov

1 The Simple Random Walk

Suppose we are given a directed graph $G = (\Omega, E)$. You have studied a number of different ways to explore such a graph, for example Breadth and Depth First Search. There is, however, another, arguably even simpler, way to explore a graph: randomly wander around it, at each step picking a random edge to follow out of the current vertex. I.e. you could run the following algorithm:

RANDOM-WALK(G, x_0)
1 $X_0 = x_0$
2 **for** $t = 1$ **to** ∞
3 Set X_t to a uniformly random y such that $(X_{t-1}, y) \in E$

The sequence of visited vertices $X_0, X_1, \ldots$ is called the *simple random walk* in G. As an illustration, see Figure 1 which shows two instances of 100 steps of a simple random walk on a regular grid graph. (The grid graph is undirected, which we take to be equivalent as having an edge in each direction for any two neighboring points.)

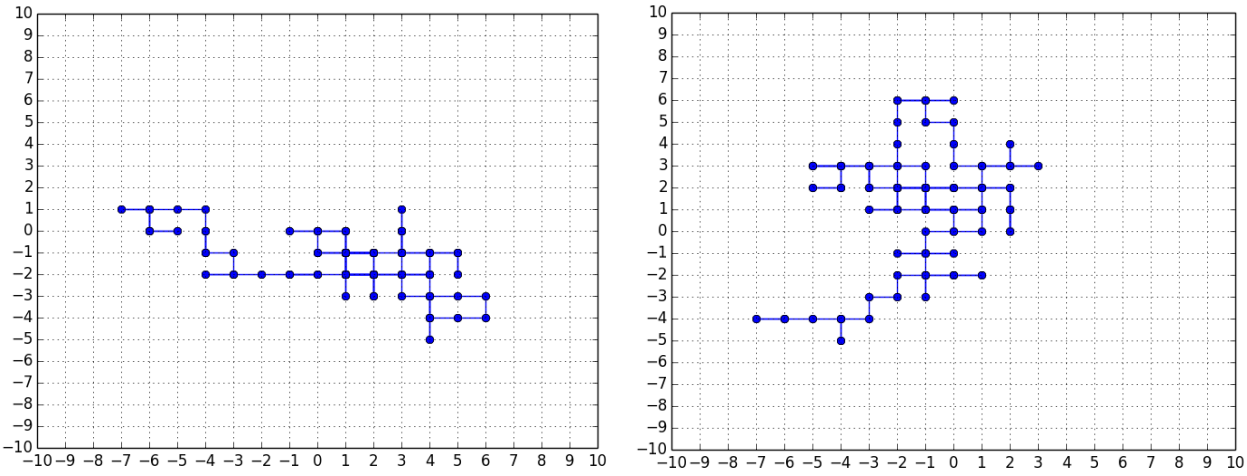

Figure 1: Two random walks on a 10 by 10 grid graph.

In a way this is the most minimalist way to explore a graph: to run the algorithm, you don't even need to remember the nodes that you have already visited, just the current node where you are. Despite this simplicity, random walks have many algorithmic applications: low-space algorithms to decide if a graph is connected; the PageRank algorithm used by Google to rank web pages; algorithms to sample from very complex distributions, often used in machine learning; algorithms to compute matchings efficiently, and more. In most applications, what we are really interested in

is the the distribution of X_t for a very large t. This distribution can provide valuable information about the structure of the graph, or be an interesting distribution in itself from which we are trying to draw a sample.

In these lecture notes we look at a broad generalization of the simple random walk, called Markov Chains. We prove their most fundamental property: that (under some conditions), the distribution of the X_t gets closer and closer to a unique distribution on the nodes of the graph which is independent of the starting distribution. We also look at some applications: Page Rank and sampling using the Markov Chain Monte Carlo method

2 Markov Chains

The random walk $(X_0, X_1, \ldots)$ above is an example of a *discrete stochastic process*. One easy generalization is to add a weight $P_{x,y} > 0$ to any edge (x, y) of the directed graph $G = (\Omega, E)$ and choose the next vertex not uniformly at random from the out-neighbors of the current one, but with probability proportional to the weight of the edge. I.e if we normalize the weights so that for any node x, $\sum_{y:(x,y)\in E} P_{x,y} = 1$, then we can modify the process to

GENERAL-WALK(P, x_0)
1 $X_0 = x$
2 **for** $t = 1$ **to** ∞
3 Choose X_t to equal y with probability $P_{X_{t-1},y}$.

Above, P is the matrix that encodes the weights, called *transition probabilities*: it has one row and one column for every vertex in Ω, and the entry for row x and column y is $P_{x,y}$ if $(x, y) \in E$, and 0 otherwise. The matrix P is called the *transition matrix*. It has the property that all its entries are non-negative, and the sum of entries in every row equals 1. Such a matrix is called *(row)-stochastic*. Clearly, this matrix entirely defines the graph G: the edges of G correspond to non-zero entries in P. Note that we allow the graph G to have self-loops, i.e. edges that go from x back to x, or, equivalently, we allow $P_{x,x} > 0$. This corresponds to allowing the random process to stay at the current vertex x with probability $P_{x,x}$.

As you can see from the pseudocode, the process $(X_0, X_1, \ldots)$ is defined by $X_0 = x_0$ and

$$\mathbb{P}(X_t = y \mid X_{t-1} = x, X_{t-2} = x_{t-2} \ldots, X_0 = x_0) = \mathbb{P}(X_t = y \mid X_{t-1} = x) = P_{x,y}. \tag{1}$$

In other words, the next step in the process depends only on the current state. Such a process is called a *Markov chain*. The set of "vertices" Ω is called the *state space*, and the vertices themselves are called *states*.

More generally, the starting state X_0 does not have to be fixed, and can be random itself. Suppose that the distribution of X_0 is given by a vector p, with entries indexed by Ω, i.e. $\mathbb{P}(X_0 = x) = p_x$. Then, by the law of total probability,

$$\mathbb{P}(X_1 = y) = \sum_{x\in\Omega} \mathbb{P}(X_1 = y \mid X_0 = x) \cdot \mathbb{P}(X_0 = x) = \sum_{x\in\Omega} P_{x,y} p_x = (pP)_y.$$

Above, p is treated as a row vector, i.e. a matrix with one row and $|\Omega|$ columns indexed by Ω, and $(pP)_y$ simply means q_y where $q = pP$ is the vector-matrix product of p and P. Since the probability

distribution of X_2 only depends on X_1, and the distribution of X_1 is given by pP, we have that the distribution of X_2 is given by $(pP)P = p(PP) = pP^2$. Going like this by induction, we have the distribution of X_t is given by pP^t, where P^t is the t-th power of P. If we look at the special case where p puts all its mass on x, i.e. $p_x = 1$ and $p_{x'} = 0$ for all $x' \neq x$, then we have

$$\mathbb{P}(X_t = y \mid X_0 = x) = (P^t)_{x,y}.$$

In general we will use the notation $P^t_{x,y}$ to denote $(P^t)_{x,y}$, i.e. the x, y entry of the matrix P^t, rather than the t-th power of the number $P_{x,y}$. We will also identify a row vector $p \in \mathbb{R}^\Omega$ with non-negative entries summing up to 1 with the probability distribution given by $\mathbb{P}(x) = p_x$.

The equation (1) intuitively means that a Markov chain has no memory: it only remembers the last state it has visited, and forgets it as soon as it moves to the next state. This suggests that, as t gets large, the probability distribution of X_t should depend less and less on the starting state x_0. This is often true, but there are some caveats. Consider for example the two Markov chains represented in Figure 2. In the left Markov chain, if we start in one of the states denoted by lower case letters, then we will always stay among them, and similarly for the upper case letters. In the right one, if we start at state $X_0 = a$, then at every *even* time t we will have $X_t \in \{a, b\}$, and $X_{t+1} \in \{A, B\}$; if we start at $X_0 = A$, then the opposite is true. So, if we want to claim that Markov chains truly forget their past, we must add further conditions on them.

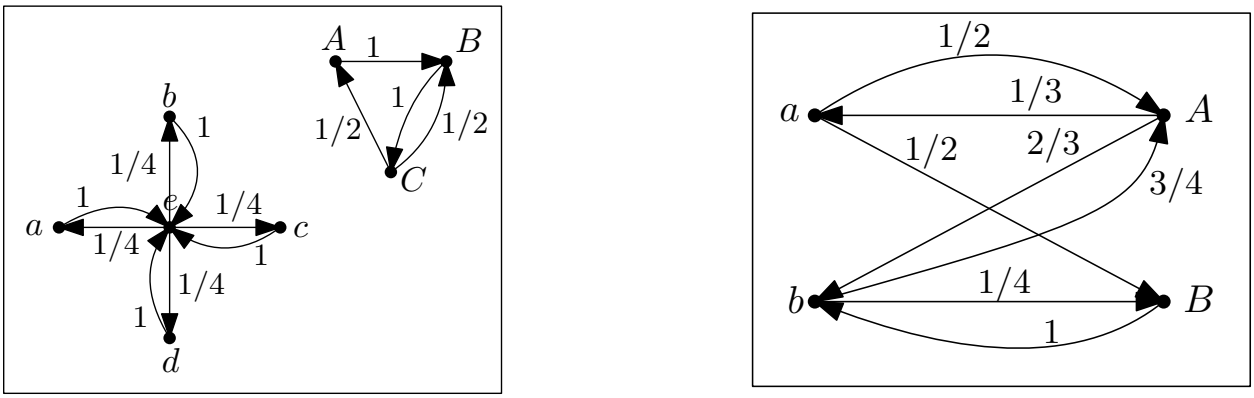

Figure 2: Two Markov chains that do not forget their starting states.

The following two definitions are the main properties that we will need.

Definition 1. *A Markov chain with transition matrix P is called* irreducible *if for all states $x, y \in \Omega$, there exists a t such that $\mathbb{P}(X_t = y \mid X_0 = x) = P^t_{x,y} > 0$. Equivalently, the graph G represented by P is strongly connected.*

Definition 2. *A Markov chain with transition matrix P is called* aperiodic *if for all states $x \in \Omega$ we have that the greatest common divisor of $\{t \geq 1 : \mathbb{P}(X_t = x \mid X_0 = x) = P^t_{x,x} > 0\}$ is 1.*

The left chain in Figure 2 fails to be irreducible: for example $\mathbb{P}(X_t = A \mid X_0 = a)$ is 0 for all t. The right chain in the figure fails to be aperiodic: the times t such that $\mathbb{P}(X_t = a \mid X_0 = a) > 0$ are all even, so their greatest common divisor is at least 2.

The following exercises should get you a little more comfortable with the notion of being aperiodic.

Exercise 1. *Suppose that P represents an irreducible Markov chain with a graph G which has at least one self-loop. Show that the Markov chain is aperiodic.*

Exercise 2. *Suppose that the graph G represented by P is symmetric, i.e. if $(x, y) \in E$, then $(y, x) \in E$ as well. Prove that if G is connected and contains an odd-length cycle, then it is aperiodic.*

NOTE: *The undirected graph $G = (V, E)$ that have no odd-length cycle are exactly the* bipartite graphs*: those whose vertices V can be partitioned into two sets U and W such that all edges go between U and W. So, this exercise says that if we take an undirected non-bipartite graph, and replace each of its edges by two directed edges in each direction, then any Markov chain on the resulting graph is aperiodic.*

Exercise 3. *Suppose that P represents an irreducible Markov chain. Define the* period *of $x \in \Omega$ by $\gcd\{t : \mathbb{P}(X_t = x \mid X_0 = x) > 0\}$. Show that the period of every x is the same.*

Exercise 4. *Show that if P represents an irreducible aperiodic Markov chain, then there exists a positive integer t_0 such that for all $t \geq t_0$, and all $x, y \in \Omega$, $P^t_{x,y} > 0$.*

HINT: *First show that if the condition holds for $x = y$, then it also holds for all x, y (by maybe taking a slightly larger t_0). Then, to show the condition for $x = y$, you can use the following fact from elementary number theory. Let $S = \{s_1, \ldots, s_N\}$ be a set of positive integers whose greatest common divisor is 1. Then there exists a positive integer t_0 such that all integers $t \geq t_0$ can be written as $t = \sum_{i=1}^{N} a_i s_i$ for some non-negative integers a_i.*

If a Markov chain is irreducible and aperiodic, then it is truly forgetful. This is formalized by the *fundamental theorem of Markov chains*, stated next. First, however, we give one last important definition.

Definition 3. *A probability distribution π is* stationary *for a Markov chain with transition matrix P if $\pi P = \pi$. In other words, π is stationary if, when the distribution of X_0 is given by $\mathbb{P}(X_0 = x) = \pi_x$, then the distribution of X_1 is also given by $\mathbb{P}(X_1 = x) = \pi_x$.*

Exercise 5. *Suppose that G is a directed graph such that if $(x, y) \in E$, then $(y, x) \in E$ as well. Let d_x be the out-degree of x in G, and let m be the number of edges in G (where (x, y) and (y, x) are counted as two different edges). Show that π given by $\pi_x = \frac{d_x}{m}$ is stationary for the simple random walk in G.*

Theorem 4 (Fundamental Theorem)**.** *For every irreducible and aperiodic Markov chain with transition matrix P, there exists a unique stationary distribution π. Moreover, for all $x, y \in \Omega$, $P^t_{x,y} \to \pi_y$ as $t \to \infty$. Equivalently, for every starting point $X_0 = x$, $\mathbb{P}(X_t = y \mid X_0 = x) \to \pi_y$ as $t \to \infty$.*

As already hinted, most applications of Markov chains have to do with the stationary distribution π. Sometimes π gives valuable information about the graph G represented by P. Other times, we want to sample from some complicated distribution π on a huge state space Ω. For example, we may want to sample a uniformly random matching in a graph $H = (V_H, E_H)$[1]. However, the set Ω of such matchings is usually exponentially large in the size of H. Instead of writing down all possible matchings of H, which would take exponential time and space, we design a Markov chain on Ω such that the uniform distribution π on Ω is stationary for the chain. We never actually write down the transition matrix P (which is also exponentially large), but we make sure that given a

[1] A matching in a graph is a set of edges that do not share any vertices.

state (i.e. matching) x, we can quickly sample the next state. Then we simply run the Markov chain, starting from some arbitrary starting state, for long enough so that the distribution of X_t gets close to π. This will eventually happen by Theorem 4. Of course, we also want to know how fast we converge to the stationary distribution. This depends a lot on the Markov chain and can be a very difficult question in general. While this example with matchings is mostly theoretical, the same strategy is applied very frequently in statistics and machine learning to sample from complicated distributions, and in physics to simulate complex processes. This is called the Markov Chain Monte Carlo (MCMC) method.

Our chief goal in the rest of the notes is to prove Theorem 4 and to give some idea about its applications.

3 Existence of Stationary Distributions

Before we prove Theorem 4, let us argue that stationary distributions in fact exist. Let us first consider the easy, but important, special case of *time reversible* Markov chains.

Definition 5. *A Markov chain with transition matrix P is reversible if there exists a probability distribution over Ω given by a vector π, such that*

$$\pi_x P_{x,y} = \pi_y P_{y,x}. \tag{2}$$

Note that, in particular, this implies that the graph G represented by P is symmetric.

Exercise 6. *Let $x_0, \ldots, x_t$ be a sequence of states. Prove that in a time-reversible Markov chain,*

$$\pi_{x_0}\mathbb{P}(X_1 = x_1, \ldots, X_t = x_t \mid X_0 = x) = \pi_{x_t}\mathbb{P}(X_1 = x_{t-1}, \ldots, X_t = x_0 \mid X_0 = x_t).$$

This is why such chains are called time-reversible: reversing time gives a Markov chain with the same transition probabilities.

Exercise 7. *Consider a Markov chain on the integers $\Omega = \{1, \ldots n\}$ with transition probabilities*

$$\begin{aligned} P_{i,i+1} &= p & 1 \le i \le n-1, \\ P_{i,i-1} &= 1-p & 2 \le i \le n, \end{aligned}$$

and $P_{1,1} = 1 - p$, $P_{n,n} = p$. Show that this Markov chain is time-reversible, and give a π that satisfies (2).

Lemma 6. *If P defines a time-reversible Markov chain, and π satisfies* (2)*, then π is a stationary distribution for P.*

Proof. By (2),

$$(\pi P)_y = \sum_{x\in\Omega} \pi_x P_{x,y} = \sum_{x\in\Omega} \pi_y P_{y,x} = \pi_y \sum_{x\in\Omega} P_{y,x} = \pi_y,$$

with the final inequality holding true because P is a stochastic matrix. □

Many chains we consider in practice are time-reversible, which is useful, since the condition (2) is usually easy to verify. Note also that Exercise 5 is a special case of Lemma 6, where $\pi_x = \frac{d_x}{m}$.

Lemma 7. *Any irreducible Markov chain has a stationary distribution.*

Proof Sketch. Suppose the transition matrix is P, and let's start the Markov chain at some arbitrary state x. Let T be the first (random) time after $t = 0$ when $X_T = x$ again. Because the chain is irreducible, we can show that $\mathbb{E}[T] < \infty$: we omit the technical details, but you can try to fill them in as an exercise. Define a vector q indexed by Ω by $q_x = 1$ and, for any $y \neq x$,

$$q_y = \mathbb{E}[\ |\{1 \leq t \leq T : X_t = y\}|\ ].$$

I.e. q_y is the expected number of times the Markov chain, when started at x, visits y before returning to x. Then we claim that $qP = q$. If we can prove that, we can just set $\pi_y = q_y/(\sum_{z\in\Omega} q_z)$ for any $y \in \Omega$, and we have found our stationary distribution.

Let us then prove $qP = q$. Let $Z_{y,t}$ be 1 if $X_t = y$ and $t \leq T$ and 0 otherwise. Then, the number of times y is visited before returning to x is $Z_y = \sum_{t=1}^{\infty} Z_{y,t}$. By linearity of expectation,

$$q_y = \mathbb{E}[Z_y] = \sum_{t=1}^{\infty} \mathbb{E}[Z_{y,t}] = \sum_{t=1}^{\infty} \mathbb{P}(X_t = y, t \leq T). \tag{3}$$

Note that $t \leq T$ simply means that none of $X_1, \ldots, X_{t-1}$ is x. Then, by the definition of the Markov chain, we have

$$\sum_{t=1}^{\infty} \mathbb{P}(X_t = y, t \leq T) = \sum_{t=1}^{\infty} \sum_{z\in\Omega} \mathbb{P}(X_{t-1} = z, t \leq T)P_{z,y} = P_{x,y} + \sum_{t=2}^{\infty} \sum_{z\neq x} \mathbb{P}(X_{t-1} = z, t-1 \leq T)P_{z,y}. \tag{4}$$

where in the second equality we used the fact that $X_0 = x$ with probability 1, and the observation that $2 \leq t \leq T$ is the same as $t - 1 \leq T$ and $X_{t-1} \neq z$. Since $q_x = 1$, and by a change of variables, we have

$$\begin{aligned} P_{x,y} + \sum_{t=2}^{\infty} \sum_{z\neq x} \mathbb{P}(X_{t-1} = z, t-1 \leq T)P_{z,y} &= q_x P_{x,y} + \sum_{t=1}^{\infty} \sum_{z\neq x} \mathbb{P}(X_t = z, t \leq T)P_{z,y} \\ &= q_x P_{x,y} + \sum_{z\neq x} P_{z,y} \sum_{t=1}^{\infty} \mathbb{P}(X_t = z, t \leq T) \\ &= \sum_{z\in\Omega} q_z P_{z,y} = (qP)_y. \end{aligned}$$

In the second equality, we just changed the order of summation,[2] and in the third we used (3). Combining with this sequence of equalities with (3) and (4), we get that $q_y = (qP)_y$. □

4 Couplings

It will be convenient to have a measure of the distance between two probability distribution. A very natural such distance is the *total variation distance*. Given two random variables X and Y,

[2] Actually, this step and the use of the linearity of expectation in (3) require further justification, because we are dealing with infinite sums. Rigorously proving that these two steps are ok requires the dominated convergence theorem, and Fubini's theorem, respectively. This is where we need $\mathbb{E}[T] < \infty$.

taking values in the same finite set Ω, their total variation distance is

$$d_{tv}(X,Y) = \max_{S \subseteq \Omega} |\mathbb{P}(X \in S) - \mathbb{P}(Y \in S)|.$$

I.e. this is the biggest difference in the probability assigned to any set by the distribution of X and the distribution of Y. When the total variation distance of X and Y is very small, we can treat the two random variables as being the same for all intents and purposes.

Exercise 8. *Suppose that the probability distribution of X is given by the vector p, and the distribution of Y is given by the vector q. Prove that*

$$d_{tv}(X,Y) = \frac{1}{2} \sum_{x \in \Omega} |p_x - q_x|.$$

HINT: *Consider the set $S = \{x \in \Omega : p_x \geq q_x\}$ and its complement.*

It turns out that the total variation distance is closely related to another central notion in probability theory: that of a *coupling*.

Definition 8. *A coupling of two random variables X and Y taking values in Ω is a random variable $Z = (Z_1, Z_2)$ taking values in $\Omega \times \Omega$, such that Z_1 has the same probability distribution as X, and Z_2 has the same probability distribution as Y. I.e. for any $x, y \in \Omega$:*

$$\mathbb{P}(X = x) = \sum_{z \in \Omega} \mathbb{P}(Z_1 = x, Z_2 = z)$$
$$\mathbb{P}(Y = y) = \sum_{z \in \Omega} \mathbb{P}(Z_1 = z, Z_2 = y).$$

As an example, consider X which is distributed uniformly in $\{1,2,3\}$, and Y which has equal probability to be 1 or 2, but is never 3. One possible coupling Z is given in the following table, where rows indicate values for Z_1, and columns values for Z_2, and each entry is the probability of the corresponding pair of values.

$Z_1 \setminus Z_2$	1	2	3
1	1/6	1/6	0
2	1/6	1/6	0
3	1/6	1/6	0

This coupling corresponds to taking independent copies of X and Y, respectively, as Z_1 and Z_2. I.e. we take $\mathbb{P}(Z_1 = x, Z_2 = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$. However, we could also first pick Z_2 to be uniform in $\{1,2\}$, and then, with probability $2/3$, pick $Z_1 = Z_2$, and with probability $1/3$ pick $Z_1 = 3$. This corresponds to the probabilities in the following table.

$Z_1 \setminus Z_2$	1	2	3
1	1/3	0	0
2	0	1/3	0
3	1/6	1/6	0

Usually we do not explicitly write Z but instead just say there is a coupling (X,Y) of X and Y, or that we have coupled X and Y, and write just X instead of Z_1 and Y instead of Z_2.

The following lemma is the key connection between couplings and total variation distance.

Lemma 9. *Two random variables X and Y taking values in a finite set Ω have $d_{tv}(X,Y) \leq \alpha$ if and only if there exists a coupling (X,Y) such that $\mathbb{P}(X \neq Y) \leq \alpha$.*

Proof. We will prove only the "if" direction, leaving the "only if" as an exercise. Suppose we have coupled X and Y such that $\mathbb{P}(X \neq Y) \leq \alpha$. For any set $S \subset \Omega$, we have

$$\begin{aligned}\mathbb{P}(X \in S) &= \mathbb{P}(X \in S, X = Y) + \mathbb{P}(X \in S, X \neq Y) \\ &= \mathbb{P}(Y \in S, X = Y) + \mathbb{P}(X \in S \mid X \neq Y)\mathbb{P}(X \neq Y) \\ &\leq \mathbb{P}(Y \in S, X = Y) + \alpha \leq \mathbb{P}(Y \in S) + \alpha.\end{aligned}$$

By the exact same reasoning, $\mathbb{P}(Y \in S) \leq \mathbb{P}(X \in S) + \alpha$. This finishes the proof. □

Exercise 9. *Prove the "only if" direction of Lemma 9.*

5 Proof of the Fundamental Theorem

We are now ready to prove the fundamental theorem. We will show the following lemma.

Lemma 10. *Suppose that $X = (X_0, X_1, \ldots)$ and $Y = (Y_0, Y_1, \ldots)$ are two instances of an irreducible aperiodic Markov chain, where X_0 and Y_0 are allowed to be random and to have different distributions. We can couple X and Y so that $\mathbb{P}(X_t \neq Y_t) \to 0$ as $t \to \infty$.*

Let us first see why this proves the theorem. By Lemma 7 there exists a stationary distribution π. Let's take $X_0 = x$, and take Y_0 to be distributed according to π. Then the coupling of X and Y gives also a coupling of X_t and Y_t for any t, and, by Lemmas 9 and 10,

$$d_{tv}(X_t, Y_t) \leq \mathbb{P}(X_t \neq Y_t) \xrightarrow[t\to\infty]{} 0.$$

By the definition of tv-distance, this means that $|\mathbb{P}(X_t = y) - \mathbb{P}(Y_t = y)|$ goes to 0 with t going to infinity. But remember that Y_0 is distributed according to the stationary distribution π, so Y_t is also distributed according to π, and $\mathbb{P}(X_t = y) - \pi_y|$ goes to 0 with t, as we want. Uniqueness follows because, if there were two stationary distributions, π and π', then we could take X_0 to be distributed according to π, and Y_0 to be distributed according to π', and, by the above reasoning, we would get that, for all $y \in \Omega$,

$$|\pi_y - \pi'_y| = |\mathbb{P}(X_t = y) - \mathbb{P}(Y_t = y)| \xrightarrow[t\to\infty]{} 0.$$

But the left hand side is independent of t, so $|\pi_y - \pi'_y| = 0$ for all $y \in \Omega$, i.e. $\pi = \pi'$.

It then remains to prove the Lemma 10.

Proof of Lemma 10. Let us couple X_t and Y_t so that, initially, X_0 and Y_0 are independent, and X_t and Y_t stay independent, and evolve according to the Markov chain, until the first (random) time T when $X_T = Y_T$. Then, for any $t \geq T$, we make X_t and Y_t stay the same, i.e. we enforce $X_t = Y_t$ for $t \geq T$, but we otherwise evolve them according to the Markov chain. You should verify that this is indeed a valid coupling of X and Y.

Let us take another coupling of X_t and Y_t, and call it (X'_t, Y'_t), in which (X'_0, Y'_0) has the same distribution as (X_0, Y_0), but X'_t and Y'_t stay independent forever. Then, (X_t, Y_t) is distributed like (X'_t, Y'_t) until the first moment T when $X_t = Y_t$; after that moment, X_t and Y_t stay "glued", while X'_t and Y'_t continue independently. Observe that

$$\mathbb{P}(X_t \neq Y_t) = \mathbb{P}(\forall s \leq t : X'_s \neq Y'_s).$$

Our goal then is to show that

$$\mathbb{P}(\forall s \leq t : X'_s \neq Y'_s) \xrightarrow[t\to\infty]{} 0.$$

Suppose that X'_0 (and X_0) has a distribution given by the vector p, and Y'_0 (and Y_0) has one given by q. Let P be the transition matrix of the Markov chain. Then the distribution of X'_t is given by pP^t, and that of Y'_t is given by qP^t. By Exercise 4, there exists a positive integer t_0 such that $P^t_{x,y} > 0$ for all x and y and all $t \geq t_0$. Then, it must be the case that all entries of the vectors pP^t and qP^t are strictly positive for all $t \geq t_0$. Combining this with the independence of X'_{t_0} and Y'_{t_0}, we get

$$\begin{aligned}\mathbb{P}(X'_{t_0} = Y'_{t_0}) &= \sum_{x\in\Omega} \mathbb{P}(X'_{t_0} = x, Y'_{t_0} = x) \\ &= \sum_{x\in\Omega} \mathbb{P}(X'_{t_0} = x)\mathbb{P}(Y'_{t_0} = x) = \sum_{x\in\Omega} (pP^{t_0})_x (qP^{t_0})_x > 0.\end{aligned}$$

Taking complements, we get $\mathbb{P}(X'_{t_0} \neq Y'_{t_0}) < 1$. Let $c < 1$ be the maximum of $\mathbb{P}(X'_{t_0} \neq Y'_{t_0})$ over all starting distributions p and q. (As an exercise, convince yourself that this maximum is achieved for p and q that put all their probability mass, respectively, on single states x and y.) Then, if $X'_s \neq Y'_s$ for all $s \leq t$, it must be true that $X'_{t_0} \neq Y'_{t_0}$, and $X'_{2t_0} \neq Y'_{2t_0}$, $X'_{3t_0} \neq Y'_{3t_0}$, etc. We have

$$\mathbb{P}(X'_{2t_0} \neq Y'_{2t_0} \mid X'_{t_0} \neq Y'_{t_0}) \leq c,$$

because we can treat X'_{2t_0} conditioned on $X'_{t_0} \neq Y'_{t_0}$ as the distribution of the Markov chain after t_0 steps, starting from the initial state X'_{t_0} conditioned on $X'_{t_0} \neq Y'_{t_0}$, and similarly for Y'_{2t_0}. (Here we use the fact that c was defined as a maximum over all starting distributions.) Then,

$$\mathbb{P}(X'_{2t_0} \neq Y'_{2t_0}, X'_{t_0} \neq Y'_{t_0}) = \mathbb{P}(X'_{2t_0} \neq Y'_{2t_0} \mid X'_{t_0} \neq Y'_{t_0})\mathbb{P}(X'_{t_0} \neq Y'_{t_0}) \leq c^2.$$

Going on like this, we see that

$$\mathbb{P}(\forall s \leq t : X'_s \neq Y'_s) \leq c^{\lfloor t/t_0 \rfloor},$$

which clearly goes to 0 as $t \to \infty$, because $c < 1$. This finishes the proof. □

Exercise 10. *Prove that $d_{tv}(X_t, Y_t)$ is non-increasing with t, where X_t and Y_t are two instances of the same Markov chain, started in two different, possibly random, initial states X_0 and Y_0. You may want to use Lemma 9. (Note: the coupling in the proof above may not actually achieve the tv-distance between X_t and Y_t.)*

There are other proofs of fundamental theorem, some of them based on what's known as the Perron-Frobenius theorem in linear algebra. The one we just gave is more elementary. At the same time, the technique of using couplings has the benefit that in many cases we can actually get useful *quantitative* bounds on how fast the Markov chain converges to its stationary distribution.

Such bounds are, of course, very important for any application: if the chain converges very slowly, then simulating it would be a very poor method of sampling from π. In general, we care about $d_{tv}(X_t, Y)$, where Y is distributed according to the stationary distribution π. In particular, we care about the *mixing time* function

$$\tau(\alpha) = \min\{t : d_{tv}(X_t, Y) \leq \alpha\}.$$

Some examples where couplings give good quantitative bounds are explored in the following exercises.

Exercise 11. *Let $\Omega = \{0,1\}^n$. Define a Markov chain on Ω by setting $P_{x,y} = \frac{1}{2n}$ if x and y have Hamming distance 1, and $P_{x,x} = \frac{1}{2}$. In other words, in every step of the Markov chain, we stay at the current state with probability $\frac{1}{2}$, and with probability $\frac{1}{2}$ we flip a random bit of it.*

- ***a.*** *Show that the uniform distribution on Ω is stationary for this Markov chain. (Hint: you can use (2).)*
- ***b.*** *Let $X_0 = x$ and let Y_0 be uniform in Ω. Let $X = (X_0, X_1, \ldots, X_t)$ and $Y = (Y_0, Y_1, \ldots, Y_t)$ evolve according to the Markov chain. Give a coupling of X and Y such that, if $t \geq Cn\ln(n/\alpha)$ for some big enough constant C, then $\mathbb{P}(X_t \neq Y_t) \leq \alpha$.*

Exercise 12. *Let Ω be the set of all permutations σ of the set $[n] = \{1, \ldots, n\}$. Consider a Markov chain defined as follows: from a state $X_t = \sigma$, we sample X_{t+1} by picking a uniformly random $i \in [n]$, and moving σ_i to the front of σ.*

- ***a.*** *Show that the uniform distribution π on Ω is stationary for this Markov chain.*
- ***b.*** *Given two copies X and Y of the Markov chain, started at two (possibly random) permutations X_0 and Y_0, consider the following coupling of X and Y. At the t-th step, we pick a uniformly random $i \in [n]$. Suppose that $X_{t-1} = \sigma$ and $Y_{t-1} = \sigma'$, and let j and k be such that $\sigma_j = \sigma'_k = i$. Then move σ_j to the front of σ to produce X_t, and σ'_k to the front of σ' to produce Y_t. Verify that this is a valid coupling of X and Y.*
- ***c.*** *Let $E_{t,i}$ be the event that, with X and Y defined and coupled as above, the location of i in X_t is the same as the location of i in Y_t. Show that if $E_{t-1,i}$ holds, then $E_{t,i}$ also holds. Also, show that*

$$\mathbb{P}(E_{t,i} \mid \text{not } E_{t-1,i}) \geq \frac{1}{n}.$$

- ***d.*** *Use the subproblems above to show that $\tau(\alpha) = O(n\log(n/\alpha))$.*

Exercise 13. *Let $\Omega = \{0, \ldots, n-1\}$, and consider a Markov chain with transition matrix P given by $P_{i,j} = \frac{1}{4}$ if if $j = i+1 \bmod n$ or $j = i-1 \bmod n$, and $P_{i,i} = \frac{1}{2}$. (This is a "lazy" simple random walk on an n-cycle, where lazy means that we have probability $\frac{1}{2}$ not to move from the current vertex.)*

- ***a.*** *Let $f(i,j) = \max\{|i-j|, n-|i-j|\}$. I.e. this is the length of the* longer *of the two paths connecting i and j. Suppose that X and Y are two instances of the above Markov chain, where $X_0 = x$ and $Y_0 = y$, and X and Y are coupled so that they are independent. Show that, if $X_{t-1} \neq Y_{t-1}$, then $\mathbb{E}[f(X_t, Y_t)^2 \mid X_{t-1}, Y_{t-1}] \geq f(X_{t-1}, Y_{t-1}) + C$ for a constant $C > 0$.*

b. *Let X and Y be defined and coupled as above, and let T be the first time when $X_T = Y_T$. Use the previous subproblem to argue that $\mathbb{E}[T] = O(n^2)$.*

c. *Use the subproblems above to bound $\tau(1/4)$.*

6 The Metropolis-Hastings Algorithm

Suppose we have a state space Ω, and a vector of positive weights w, indexed by Ω. These weights define a probability distribution π, given by $\pi_x = \frac{w_x}{\sum_{y\in\Omega} w_x}$. We want to sample a random variable X, taking values in Ω, so that, for every $x \in \Omega$, $\mathbb{P}(X = x) = \pi_x$. As we already mentioned several times, often Ω is a very large set, and we would like the sampling to take time much less than $|\Omega|$. In many cases, even computing the normalizing factor $Z = \sum_{y\in\Omega} w_x$ in time less than $|\Omega|$ is a difficult problem. However, sometimes we can sample from a distribution which is close to π by a running a Markov chain designed so that its stationary distribution is π. In this section we give one very general method to achieve this, due to Metropolis and Hastings.

Let's start with a very natural example. Suppose that $H = (V_H, E_H)$ is an undirected graph with n vertices and m edges. Recall that a *matching* of H is a set of edges $M \subseteq E_H$ so that no two edges in M share an endpoint. We let Ω be the set of matchings in H; this set can easily be exponential in the size of H. We give each matching M of H weight $w_M = 1$. Then the probability distribution π defined, as above, by $\pi_M = \frac{w_M}{\sum_{M'\in\Omega} w_{M'}}$ is just the uniform distribution over Ω, and $\pi_M = \frac{1}{|\Omega|}$. Sampling from π then corresponds to sampling a uniformly random matching of H. *Can we we draw such a sample in time which is polynomial in the size of H?*

The bad news is that, by a famous theorem of Valiant, even computing $|\Omega|$ is at least as hard as any problem in NP. The good news is that it is easy to design a Markov chain whose stationary distribution is π. A step of this chain is given, in pseudocode, next. The algorithm Match-Step takes a matching X_t, which is the state of the Markov chain at time t, and produces a matching X_{t+1}, which is the state at time $t + 1$.

Match-Step(X_t)

1 Sample a uniformly random $e \in E_H$
2 **if** $e \in X_t$
3 $\quad Y = X_t \setminus \{e\}$
4 **else** $Y = X_t \cup \{e\}$
5 **if** Y is *not* a matching
6 $\quad$ Output $X_{t+1} = X_t$
7 **else**

$$\text{Output } X_{t+1} = \begin{cases} Y & \text{with probability } \frac{1}{2} \\ X_t & \text{with probability } \frac{1}{2} \end{cases}$$

Clearly this is a Markov chain over the set Ω of matchings in H. In the graph G corresponding to the Markov chain, every vertex M is a matching in H, and there is an edge from M to M' if $M' = M$ or if we can get M' from M by adding or removing a single edge. An example of G for a 4-vertex graph H is given in Figure 3. We claim that G is strongly connected, so the chain is

irrediciuble. To get from any matching M to M', first remove all edges in $M \setminus M'$ from M, one by one. Then you are left with $M \cap M'$. Now you can add, one by one, all edges in $M' \setminus M$. This describes a path in G. Moreover, since each M has a self-loop in G, by Exercise 1 the Markov chain is also aperiodic. If we can also verify that the stationary distribution of the chain is uniform, then, by the fundamental theorem, we would know that running it long enough would produce a random matching in H whose distribution is close to uniform.

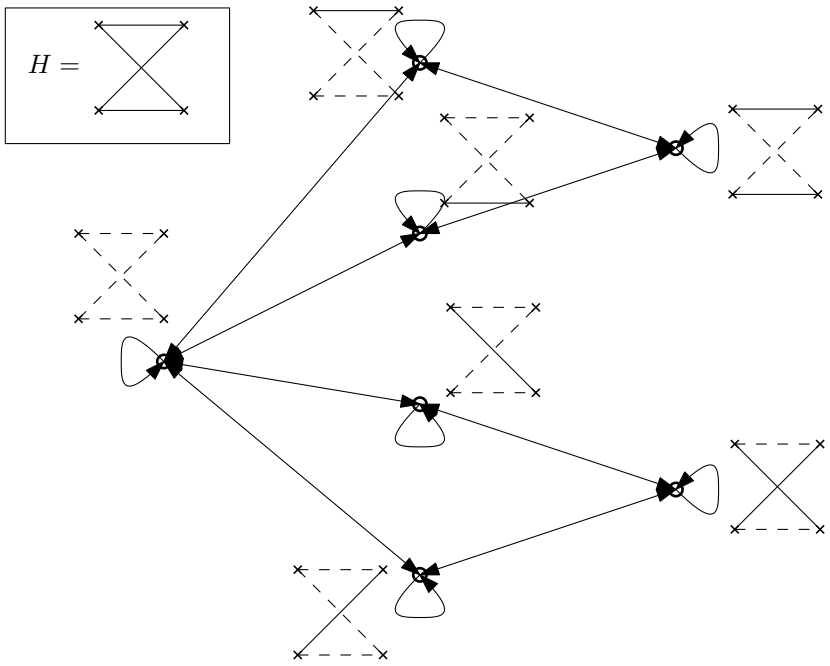

Figure 3: The transition graph G for the Markov chain on matchings of the graph H.

Let d_M be the out-degree of M in G (not counting self-loops). I.e. d_M is the number of matchings we can get from M by adding or removing an edge, or, equivalently, d_M is the number of edges that can be added to or removed from M so that we still have a matching. Then, for $X_t = M$, the probability that Y is a matching is equal to $\frac{d_M}{m}$. We have that, for any matching M, the probability that the Markov chain stays at M is

$$P_{M,M} = \left(1 - \frac{d_M}{m}\right) + \frac{d_M}{m} \cdot \frac{1}{2} = 1 - \frac{d_M}{2m},$$

where the first term is the probability that Y is not a matching, and the second term is the probability that Y is a matching, but we still stay at M. If M and M' are distinct matchings of H such that (M, M') is an edge of G, then the probability of transitioning from M to M' is

$$P_{M,M'} = \frac{d_M}{m} \cdot \frac{1}{2} \cdot \frac{1}{d_M} = \frac{1}{2m},$$

because, if $X_t = M$, the probability that Y is a matching is $\frac{d_M}{m}$, and, conditional on this event, the probability of moving to a new matching is $\frac{1}{2}$, and any neighboring matching M' of M is equally likely.

Now we can verify that the Markov chain is time reversible. Notice first that if (M, M') is an edge of G, then so is (M', M): if we can get M' from M by adding an edge to M, then we can get M

from M' by removing the same edge, and vice versa. Then, for $\pi_M = \frac{1}{|\Omega|}$ and any edge (M, M') of G where M and M' are distinct matchings of H, we have

$$\pi_M P_{M,M'} = \frac{1}{2m|\Omega|} = \pi_{M'} P_{M',M}.$$

The time-reversible condition (2) is of course trivial for self-loops. It follows that the Markov chain is time reversible, and, by Lemma 6, has stationary distribution π, i.e. uniform over Ω.

Running this Markov chain for $\tau(\varepsilon)$ steps then produces a random matching whose probability distribution is within ε from the uniform one in total variation distance. Since a a single step of the Markov chain can be executed in time $O(n+m)$, the total running time of this algorithm is $O((n+m)\tau(\varepsilon))$. If we knew that $\tau(\varepsilon)$ is polynomial in n and m, we would have an efficient algorithm to approximately sample a random matching in H. It turns out that $\tau(\varepsilon) = O(m^2 n \log(n) \log(1/\varepsilon))$, but proving this is beyond the scope of these notes.

The design of the Markov chain above follows from a general method. Suppose now that Ω is some arbitrary state space, and $G = (\Omega, E)$ is a *connected* graph such that if $(x, y) \in E$, then $(y, x) \in \Omega$ too. Let d be an upper bound on the maximum out-degree in G of any state $x \in \Omega$. Let w be a vector of positive weights, one for each state in Ω. Define a a single step of a Markov chain on Ω as follows:

METROPOLIS-STEP(X_t)
1 Let $\Gamma(X_t) = \{y : (X_t, y) \in E\}$ be the out-neighbors of x
2 Pick Y so that, for any $y \in \Gamma(X_t)$, $\mathbb{P}(Y = y) = \frac{1}{d}$ and $\mathbb{P}(Y = \perp) = 1 - \frac{|\Gamma(X_t)|}{d}$
3 **if** $Y == \perp$
4 Output $X_{t+1} = X_t$
5 **else**

$$\text{Output } X_{t+1} = \begin{cases} Y & \text{with probability } \frac{1}{2} \cdot \min\left\{1, \frac{w_Y}{w_{X_t}}\right\} \\ X_t & \text{with probability } 1 - \frac{1}{2} \cdot \min\left\{1, \frac{w_Y}{w_{X_t}}\right\} \end{cases}$$

Theorem 11. *The Markov chain defined by* METROPOLIS-STEP(X_t) *is irreducible, aperiodic, and has stationary distribution π defined by $\pi_x = \frac{w_x}{\sum_{y \in \Omega} w_y}$.*

Proof. The proof is analogous to the special case of matchings we considered above.

The chain is irreducible because we assumed that the graph G is connected and symmetric, so it must be strongly connected. Moreover, it is aperiodic by Exercise 1, because at every step we have probability at least $\frac{1}{2}$ to stay at the current state.

For any $x \neq y$ such that $(x, y) \in E$, the transition probability of the chain is

$$P_{x,y} = \frac{1}{d} \cdot \frac{1}{2} \cdot \min\left\{1, \frac{w_y}{w_x}\right\}.$$

Moreover, if $(x, y) \in E$, then $(y, x) \in E$, and, by symmetry,

$$P_{y,x} = \frac{1}{d} \cdot \frac{1}{2} \cdot \min\left\{1, \frac{w_x}{w_y}\right\}.$$

It follows that

$$\pi_x P_{x,y} = \frac{1}{2d \sum_{z \in \Omega} w_z} \cdot \min\{w_x, w_y\} = \pi_y P_{y,x}.$$

Therefore, the chain is time reversible, and, by Lemma 6 has stationary distribution π. □

The algorithm which simulates this Markov chain in order to sample approximately from π is called the Metropolis-Hastings algorithm.

Exercise 14. *Verify that* Match-Step *is a special case of* MH-Step.

Let $\lambda > 0$, and, for any matching M in H, define $w_M = \lambda^{|M|}$. Using MH-step*, modify* Match-Step *so that the modified Markov chain has stationary distribution π given by $\pi_M = \frac{w_M}{\sum_{M' \in \Omega} w_{M'}}$, where Ω is the set of matchings of H. Every step of the Markov chain should still run in time $O(n+m)$*