Random walks on graphs

1 Random walks in continuous time

In this section, we shall study continuous time random walks on graphs. The parallels with discrete-time random walks are close, so we won't repeat the analysis for them. However, we will go on to discuss the Page Rank metric for web pages, and point out its connection to discrete-time random walks.

Let $G = (V, E)$ be a connected, undirected graph. We shall consider the following model of a random walk on G. The walk starts out at time 0 in some initial distribution on the vertex set V; the initial condition may be deterministic, which corresponds to the distribution assigning probability 1 to a single vertex. The random walk is constructed as follows. There are independent unit rate Poisson processes associated with each edge of the graph. If the random walk is at some vertex $v \in V$ at time s, its next jump occurs at the first time $t > s$ at which there is an increment of the Poisson process on some edge (v, w) incident on V. If this increment occurs on edge (v, u), then the random walk moves to vertex u at time t. (The process is well defined because the probability of the next increment happening at exactly the same time on two different edges is zero.) Equivalently, if the random walk is at vertex v, it samples independent Exp(1) random variables for each edge (v, w) incident on v, and moves along the edge with the minimum value of these random variables; the time spent at v is equal to this minimum value. From this description, and the fact that the minimum of independent exponential random variables is exponential with the sum of their rates, it is clear that the time spent at a vertex v on each visit has an Exp(d_v) distribution, where d_v denotes the degree of the vertex v. Moreover, the next vertex to be visited after v is chosen uniformly at random from the neighbours of v.

Question Suppose that, in the description of the random walk, the Poisson processes associated with different edges had different rates, $\lambda(v, w)$ for edge (v, w), but were still mutually independent. Then what would be the distribution of the time spent at vertex v during a visit, and what would be the probability of next going to a vertex w?

Let X_t denote the location of the random walk at time t. It should be clear from the description above, and the memorylessness property of Poisson processes, that X_t is a Markov process with state space V, the vertex set. Moreover, every vertex is accessible from every other vertex since the graph G is connected, and so the Markov chain is irreducible. Since it is a finite state chain, it follows that it has a unique invariant distribution, which we shall denote π. Let us now write down the rate matrix for this chain and then work out its invariant distribution. The probability of moving from a vertex v to a vertex w during some small time interval of size δ is the probability the clock on the edge (v, w) goes off during this time interval, and that no other clock on an edge incident on v does so. (If there is no edge between v and w, the probability of moving from v to w during this interval is negligible, $o(\delta)$). As the clocks are indepent Exp(1) random variables, we can work out this probability to be $\delta(1-\delta)^{d_v-1} + o(\delta)$, which reduces to $\delta + o(\delta)$. Hence, the rate associated with this move is 1 if (v, w) is an edge, and 0 otherwise. As the sum applies to all vertices, we conclude that the rate matrix Q has elements

$$q_{vw} = \begin{cases} 1, & \text{if } (v,w) \in E, \\ 0, & \text{if } (v,w) \notin E \text{ and } v \neq w, \\ -d_v, & \text{if } v = w. \end{cases}$$

It is easy to see from this that the rate matrix Q is precisely $-L_G$, the negative of the Laplacian of the graph G.

Let $\mathbf{p}(t)$ denote the distribution of the random variable $X(t)$, i.e., let $\mathbf{p}(t)$ be a vector of length $n = |V|$ with elements $p_x(t) = \mathbb{P}(X_t = x)$ for $x \in V$. Likewise, let $\mathbf{p}(0)$ denote the initial condition, the distribution of X_0, the position of the random walk at time 0. Since X_t is a Markov process with rate matrix $Q = -L_G$, we have

$$\mathbf{p}(t) = \mathbf{p}(0)e^{Qt} = \mathbf{p}(0)e^{-L_G t}.$$

Moreover, $\pi = \pi e^{Qt} = \pi e^{-L_g t}$ for all t, since π is invariant. To see this, note that π satisfies the global balance equations $\pi Q = \mathbf{0}$, and use the Taylor

expansion of e^{Qt}. Hence, we have

$$\mathbf{p}(t) - \pi = (\mathbf{p}(0) - \pi)e^{-L_g t}.$$

We know from the ergodic theorem for Markov chains, and the uniqueness of the invariant distribution π (which, in turn, follows from the fact that G is connected and, consequently, that the Markov chain X_t is irreducible), that $\mathbf{p}(t)$ converges to π as t tends to infinity. We want to know how quickly this convergence happens. First, we shall study the convergence treating $\mathbf{p}(t)$ and π as vectors in Euclidean space and using the usual Euclidean distance between the vectors. Subsequently, we shall introduce a commonly used distance measure between probability distributions known as total variation distance, and restate our results in terms of this metric.

We begin by obtaining bounds on $\|\mathbf{p}(t)-\pi\|^2$, the squared Euclidean distance between $\mathbf{p}(t)$ and π. Recall that the dot product of a vector with itself is the square of its norm. Noting that $\mathbf{p}(t)$ and π are row vectors, we can write

$$\begin{aligned}
\|\mathbf{p}(t) - \pi\|^2 &= (\mathbf{p}(t) - \pi) \cdot (\mathbf{p}(t) - \pi) = (\mathbf{p}(t) - \pi)(\mathbf{p}(t) - \pi)^T \\
&= (\mathbf{p}(0) - \pi)e^{-L_G t}\Big((\mathbf{p}(0) - \pi)e^{-L_G t}\Big)^T \\
&= (\mathbf{p}(0) - \pi)e^{-2L_G t}(\mathbf{p}(0) - \pi)^T. \qquad (1)
\end{aligned}$$

We have used the fact that L_G is symmetric to deduce that the same is true of $e^{-L_G t}$, and hence to obtain the last equality.

Let us denote the eigenvalues of the Laplacian matrix L_G by $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n$. Recall that 0 is always an eigenvalue of the Laplacian, with corresponding eigenvector $\mathbf{1}$, the all-1 vector. We saw in an earlier chapter that the Laplacian of any graph G is positive semi-definite, and that the multiplicity of the eigenvalue 0 is the same as the number of connected components in the graph G. In particular, if G is connected, then all other eigenvalues are strictly positive, so the inequalities claimed above hold. Next, it is easy to see that if $\mathbf{v}$ is an eigenvector of L_G corresponding to the eigenvalue λ, it is also an eigenvector of $e^{-2L_G t}$, with eigenvalue $e^{-2\lambda t}$. Hence, the eigenvalues of $e^{-2L_G t}$ are $1 = e^{-2\lambda_1 t} > e^{-2\lambda_2 t} \geq \ldots \geq e^{-2\lambda_n t}$. The all-1 vector $\mathbf{1}$ is an eigenvector corresponding to the eigenvalue 1. Recall from an earlier chapter that the second largest eigenvalue of the symmetric matrix $e^{-2L_G t}$ is given by

$$e^{-2\lambda_2 t} = \max_{\mathbf{x} \in \mathbb{R}^{\mathbf{n}}: \mathbf{x} \perp \mathbf{1}} \frac{\mathbf{x}^T e^{-2L_G t}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}. \qquad (2)$$

Now $\mathbf{p}(0)$ and π are both probability distributions and sum to 1, so $\mathbf{p}(0)\cdot\mathbf{1} = \pi\cdot\mathbf{1} = 1$. Hence, $(\mathbf{p}(0)-\pi)\cdot\mathbf{1} = 0$, i.e., $\mathbf{p}(0)-\pi$ is orthogonal to $\mathbf{1}$. It follows from (2) that

$$(\mathbf{p}(0)-\pi)^T e^{-2L_G t}(\mathbf{p}(0)-\pi) \leq e^{-2\lambda_2 t}(\mathbf{p}(0)-\pi)^T(\mathbf{p}(0)-\pi).$$

Substituting this in (1), we conclude that

$$\|\mathbf{p}(t)-\pi\|^2 \leq e^{-2\lambda_2 t}\|\mathbf{p}(0)-\pi\|^2. \tag{3}$$

Since λ_2 is strictly positive, the Euclidean distance between the vectors $\mathbf{p}(t)$ and π decays exponentially as t tends to infinity. In that sense, $\mathbf{p}(t)$ converges to π.

We can also establish this convergence in a more commonly used measure of distance between probability distributions, which we now introduce.

Defintion: Let $\mathbf{p}$ and $\mathbf{q}$ denote two probability distributions on a finite sample space S. The total variation distance between them, denote $d_{TV}(\mathbf{p},\mathbf{q})$ is defined as

$$d_{TV}(\mathbf{p},\mathbf{q}) = \max_{A\subseteq S} p(A) - q(A),$$

where, as usual, $p(A) = \sum_{x\in A} p(x)$ denotes the probability of the set A and, with some abuse of notation, we write $p(x)$ for $p(\{x\})$.

Note that the total variation distance between any two probability distributions is a number between 0 and 1. If X and Y are random variables with distributions $\mathbf{p}$ and $\mathbf{q}$, we may write $d_{TV}(X,Y)$ instead to denote the same quantity. The definition can be extended to sample spaces S that are countably or uncountably infinite.

It is easy to see that the maximum is attained by the set $A = \{x \in S : p(x) \geq q(x)\}$, that $p(A^c) - q(A^c) = q(A) - p(A)$, and hence that

$$d_{TV}(\mathbf{p},\mathbf{q}) = \frac{1}{2}\sum_{x\in S} |p(x) - q(x)|.$$

The sum on the RHS is called the ℓ_1 norm of the vector $\mathbf{p}-\mathbf{q}$. If the sample space S is uncountable, then $\mathbf{p}$ and $\mathbf{q}$ can be taken to be probability density functions, and the sum replaced by an integral.

You can verify that the total variation distance is symmetric ($d_{TV}(\mathbf{p},\mathbf{q}) = d_{TV}(\mathbf{q},\mathbf{p})$), that $d_{TV}(\mathbf{p},\mathbf{q}) \geq 0$, with equality if and only if $\mathbf{p} = \mathbf{q}$, and

that it satisfies the triangle inequality $d_{TV}(\mathbf{p}, \mathbf{q}) + d_{TV}(\mathbf{q}, \mathbf{r}) \geq d_{TV}(\mathbf{p}, \mathbf{r})$ for any three probability distributions $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{r}$. These three properties imply that the total variation distance is a metric on the space of probability distributions (for a given sample space - it is not defined for two probability distributions on two different sample spaces).

Suppose the sample space S has cardinality n. We can bound the total variation distance between any two probability distributions $\mathbf{p}$ and $\mathbf{q}$ on S in terms of the Euclidean distance between those vectors, as follows:

$$\begin{aligned} d_{TV}(\mathbf{p}, \mathbf{q}) &= \frac{1}{2} \sum_{x \in S} |p(x) - q(x)| \\ &= \frac{1}{2} (\mathbf{p} - \mathbf{q}) \cdot \text{ sign } (\mathbf{p} - \mathbf{q}) \\ &\leq \frac{1}{2} \|\mathbf{p} - \mathbf{q}\| \| \text{ sign } (\mathbf{p} - \mathbf{q})\| \leq \frac{\sqrt{n}}{2} \|\mathbf{p} - \mathbf{q}\|, \end{aligned}$$

where $\cdot$ denotes the dot product between vectors, sign $(\mathbf{x})$ denotes the $\{-1, 0, 1\}$-valued vector whose elements are the sign of the corresponding elements of $\mathbf{x}$, and we have used the Cauchy-Schwarz inequality to obtain the first inequality above. Combining this bound with (3), we get

$$d_{TV}(\mathbf{p}(t), \pi) \leq \frac{\sqrt{n}}{2} e^{-\lambda_2 t} \|\mathbf{p}(0) - \pi\|.$$

It can be shown that $\|\mathbf{p} - \mathbf{q}\| \leq \sqrt{2}$ for any two probability distributions $\mathbf{p}$ and $\mathbf{q}$. Substituting this above, we conclude that

$$d_{TV}(\mathbf{p}(t), \pi) \leq \sqrt{\frac{n}{2}} e^{-\lambda_2 t}, \tag{4}$$

for arbitrary initial conditions.

Now, given an arbitrary $\epsilon > 0$, we can ask how long we need to wait before the random walk is within ϵ of the invariant distribution π in total variation distance. More precisely, we can ask how large we need to choose T such that, for any $t > T$, we have $d_{TV}(\mathbf{p}(t), \pi) < \epsilon$. We see from (4) that choosing

$$T = \frac{1}{2\lambda_2} \log \frac{n}{2\epsilon}$$

sufficies to guarantee the desired bound on the total variation distance.

The time T is called the mixing time. The main things to take away from the above result are the scalings. The mixing time grows in the size of

the graph as $\log n$. It grows in the stringency of the desired closeness to the invariant distributation as $\log(1/\epsilon)$. And finally, its dependence on the shape of the graph is captured by the second eigenvalue λ_2 of the Laplacian. More commonly, this is expressed as saying that it depends on the spectral gap, $\lambda_2 - \lambda_1 = \lambda_2 - 0$. The reason for using the terminology of a spectral gap is that similar results hold for general Markov processes, and not just for random walks. In general, the eigenvalues of the rate matrix may be complex, but there is always a real eigenvalue of 0 (with $\mathbf{1}$ as the right eigenvector). In the general case, the mixing time depends on how far the remaining eigenvalues are from the imaginary axis, i.e., the gap between the imaginary axis and the other eigenvalues. Similarly, for general discrete-time Markov chains, there is always an eigenvalue at 1, and the mixing time depends on the gap between the remaining eigenvalues, which have to lie within the unit circle in the complex plane (by the Perron-Frobenius theorem, which we shall encounter later), and the boundary of this circle.

2 Bounds on the spectral gap

If we are given a graph on n nodes, then it is quite straightforward to compute the eigenvalues of its Laplacian matrix (to a desired accuracy). There are well-known numerical algorithms for this purpose, whose computational complexity is polynomial in n, growing no faster than n^3. In this section, we provide bounds on the spectral gap λ_2 in terms of geometric properties of the graph, specifically its conductance. These bounds are not very useful computationally, as there is no known polynomial-time algorithm for computing the conductance (though the bounds could be turned around to bound the conductance in terms of the spectral gap). The reason for providing these bounds is that, while the conductance can't be computed easily for a general graph, it can be for graphs with special structure, like the star or cycle graph we have seen earlier in the course. It is also possible to get bounds on the conductance for various models of random graphs, which are widely used as models of complex networks. We shall state our bounds after defining the conductance.

Definition: We define the conductance of a graph $G = (V, E)$ on n nodes as

$$\Phi(G) = \min_{S \subset V, S \neq \emptyset} \frac{|E(S, S^c)|}{\frac{1}{n}|S||S^c|}, \tag{5}$$

where, for any proper subset S of the vertex set, $E(S, S^c)$ denotes the set of edges crossing from S to S^c, namely edges (i, j) such that $i \in S$ and $j \in S^c$ or vice versa.

Example. If G is the triangle with nodes denoted $1, 2, 3$ and $S = \{1\}$, then $E(S, S^c) = \{(1, 2), (1, 3)\}$. For this graph, it is easy to check that $\Phi(G) = 3$, since for any S consisting of either 1 or 2 nodes, $|E(S, S^c)| = 2$.

Remark. The definition above is similar to the definition we gave earlier for the conductance of a non-negative matrix P. Indeed, it is the same definition applied to the adjacency matrix of the graph. An alternative definition of conductance in widespread use has the same numerator, but $\min\{|S|, |S^c|\}$ in the denominator. These two definitions differ by at most a factor of 2 from each other, and yield very similar bounds. We shall stick with the definition above.

Theorem 1 *The second smallest eigenvalue of the Laplacian is related to the conductance as follows:*

$$\lambda_2 \leq \Phi(G) \leq \sqrt{8 d_{\max} \lambda_2}. \tag{6}$$

Here $d_{\max}$ denotes the maximum node degree of all nodes in G.

The second inequality in (6) is called Cheeger's inequality. We shall give a proof of the first inequality below, but will take the second one for granted as its proof is more involved. We begin with the following lemmas.

Lemma 1 *The conductance of the graph G can be expressed as the solution of an integer programming problem, as follows:*

$$\Phi(G) = \min_{\mathbf{x} \in \{0,1\}^n} \frac{\sum_{(i,j) \in E} (x_i - x_j)^2}{\frac{1}{2n} \sum_{i,j \in V} (x_i - x_j)^2}, \tag{7}$$

where the minimum is restricted to vectors which aren't identically 0 or 1, so that the ratio is well defined.

Proof. Fix a subset S of V. Let $\mathbf{x}$ be the vector with components $x_i = 1$ if $i \in S$ and $x_i = 0$ if $i \in S^c$. Therefore, for $(i, j) \in E$, if $i \in S$ and $j \in S^c$ or vice versa, then $(x_i - x_j)^2 = 1$, whereas this quantity is zero if i and j are both in S or both in S^c. Hence, the numerator of the right hand side of (7)

is just $|E(S, S^c)|$ as an edge is counted only if it crosses the cut between S and S^c. Likewise, $(x_i - x_j)^2$ is equal to 1 if $i \in S$ and $j \in S^c$ or vice versa, and equal to 0 if i and j are both in S or both in S^c. As the number of ways of choosing a pair $i \in S$ and $j \in S^c$ is $|S| \cdot |S^c|$, the sum in the denominator is $2|S| \cdot |S^c|/n$ (with the factor of 2 because each pair of nodes is counted twice, once with $i \in S$ and $j \in S^c$, and once with $j \in S$ and $i \in S^c$).

Conversely, given any 0-1 vector of length n, we can identify it with a subset S of V. We have thus established a one-to-one correspondence between subsets of V and 0-1 vectors, and shown that the ratio in the statement of the lemma is equal to the statement in the definition of conducatance. This establishes the claim of the lemma. □

Lemma 2 *The second smallest eigenvalue of the Laplacian of G is given by the solution of the following optimisation problem:*

$$\lambda_2 = \min_{\mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0}} \frac{\sum_{(i,j) \in E} (x_i - x_j)^2}{\frac{1}{2n} \sum_{i,j \in V} (x_i - x_j)^2}. \tag{8}$$

Proof. Recall that $\lambda_1 = 0$ is an eigenvalue of the Laplacian, with eigenvector $\mathbf{1}$, the all-1 vector. Hence, by the variational characterisation of eigenvalues,

$$\lambda_2 = \min_{\mathbf{x} \in \mathbb{R}^n, \mathbf{x} \perp \mathbf{1}} \frac{\mathbf{x}^T L_G \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \tag{9}$$

We saw in an earlier section that $\mathbf{x}^T L_G \mathbf{x} = \sum_{(i,j) \in E} (x_i - x_j)^2$. Moreover,

$$\frac{1}{2n} \sum_{i,j \in V} (x_i - x_j)^2 = \frac{1}{2} \Big(\sum_{i \in V} x_i^2 + \sum_{j \in V} x_j^2 \Big) - \frac{1}{2n} \sum_{i \in V} x_i \sum_{j \in V} x_j.$$

Now, if $\mathbf{x} \perp \mathbf{1}$, then $\sum_{i \in V} x_i = 0$, and so

$$\frac{1}{2n} \sum_{i,j \in V} (x_i - x_j)^2 = \sum_{i \in V} x_i^2 = \mathbf{x}^T \mathbf{x}.$$

Thus, for $\mathbf{x} \perp \mathbf{1}$, the quantity being minimised on the right hand side of (8) is just $\mathbf{x}^T L_G \mathbf{x} / \mathbf{x}^T \mathbf{x}$.

On the other hand, suppose $\mathbf{x}$ is not orthogonal to $\mathbf{1}$. Note that neither the numerator or the denominator on the RHS of (8) is changed if add or subtract a constant vector from $\mathbf{x}$. If we define

$$\mathbf{y} = \mathbf{x} - \frac{(\mathbf{x}^T \mathbf{1})}{n} \mathbf{1},$$

then $\mathbf{y}$ is orthogonal to $\mathbf{1}$ and the RHS of (8) has the same value for $\mathbf{y}$ as for $\mathbf{x}$. Thus, minimising this quantity over all of $\mathbb{R}^n \setminus \mathbf{0}$ is the same as minimising it over $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \perp \mathbf{1}$. In the latter case, the quantity being minimised is the Rayleigh quotient, $x^T L_G x / x^T x$. Hence, the minimum is the second smallest eigenvalue λ_2. This completes the proof of the lemma. □

The proof of the first inequality in Theorem 1is straightforward from the above two lemmas. They show that $\Phi(G)$ and λ_2 both involve the minimisation of the same expression. However, $\Phi(G)$ involves its minimisation only over 0-1 vectors while λ_2 involves its minimisation over all vectors. As λ_2 involves taking the minimum over a larger set, it cannot be bigger than $\Phi(G)$.

3 Coupling

We now study a different technique for bounding the mixing time of Markov chains. The idea of coupling is that, given two random variables X and Y (or two probability distributions) which may be defined on different probability spaces, it is possible to construct two other random variables $\tilde{X}$ and $\tilde{Y}$ on the same probability space such that $\tilde{X}$ has the same distribution as X, and $\tilde{Y}$ has the same distribution as Y. The same can also be done with random processes. The point is that it might be possible to jointly construct $\tilde{X}$ and $\tilde{Y}$ in such a way as to easily prove statements about X and Y.

Example Suppose X has a $\text{Bin}(n, p)$ distribution and Y a $\text{Bin}(m, p)$ distribution, where $n > m$. Let $0 \leq k \leq m$. Show that $\mathbb{P}(X > k) \geq \mathbb{P}(Y > k)$.

It is not that easy to explicitly compare the expressions for probabilities of the two events. However, it is very easy to show this by coupling. We construct the random variables $\tilde{X}$ and $\tilde{Y}$ with the same probability distributions as X and Y respectively, as follows. Let $\xi_1, \xi_2, \xi_3, \ldots$ be a sequence of iid $\text{Bern}(p)$ random variables. Set $\tilde{X} = \sum_{i=1}^{n} \xi_i$ and $\tilde{Y} = \sum_{i=1}^{m} \xi_i$. Since $n > m$ and the ξ_i are non-negative, it is clear by construction that $\mathbb{P}(\tilde{X} \geq \tilde{Y}) = 1$. Hence, for any fixed k, $\mathbb{P}(\tilde{X} > k) \geq \mathbb{P}(\tilde{Y} > k)$. As $\tilde{X}$ has the same distribution as X, and $\tilde{Y}$ has the same distribution as Y, $\mathbb{P}(X > k) \geq \mathbb{P}(Y > k)$.

Let μ and ν denote two probability distributions on a countable set S. The following lemma bounds the total variation distance between the distributions in terms of a coupling between them. In fact, the result holds for an

arbitrary measurable space $(S, \mathcal{S})$, but a countable set suffices for conveying the intuition, without requiring knowledge of measure theory.

Lemma 3 *Let X and Y be S-valued random variables defined on the same probability space, such that X has distribution μ and Y has distribution ν, denoted $X \sim \mu$, and $Y \sim \nu$. In other words, (X, Y) is a coupling of the distributions μ and ν. Then,*

$$d_{TV}(\mu, \nu) \leq \mathbb{P}(X \neq Y). \tag{10}$$

Proof. We have

$$\begin{aligned} d_{TV}(\mu, \nu) &= \max_{A \subseteq S} |\mu(A) - \nu(A)| = \max_{A \subseteq S} \mu(A) - \nu(A) \\ &= \max_{A \subseteq S} \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \\ &\leq \max_{A \subseteq S} \mathbb{P}(X \in A) - \mathbb{P}(X \in A, Y \in A) \\ &= \max_{A \subseteq S} \mathbb{P}(X \in A, Y \in A^c) \\ &\leq \max_{A \subseteq S} \mathbb{P}(X \in A, X \neq Y) \leq \mathbb{P}(X \neq Y). \end{aligned}$$

□

Remark In fact, there is a coupling, namely a pair of random variables X and Y defined on the same probability space, for which equality holds in (10). Such a pair is called a maximal coupling, as they maximise the probability that $X = Y$.

We now turn from random variables to random processes, and specifically to Markov processes. Let $(X_t, t \geq 0)$ be an irreducible Markov process on a finite state space S, starting in some initial state $X_0 = x$. Let π denote the (unique) invariant distribution of this Markov process. Let $(Y_t, t \geq 0)$ be a Markov process with the same probability law (i.e., same generator matrix) as X_t, but started in the invariant distribution π, so that the marginal distribution of Y_t at any fixed t is π. Suppose that X_t and Y_t are constructed on the same probability space, and define the random time

$$T_{cpl} = \inf\{t \geq 0 : X_t = Y_t\},$$

which we call the coupling time. Extend the coupling such that $X_t = Y_t$ for all $t > T_{cpl}$, i.e., ensure that the two processes have exactly the same jumps after they meet.

Lemma 4 *At any time $t > 0$, the total variation distance between the probability distribution of X_t, which we denote $\mathbf{p}(t)$, and the invariant distribution π, is bounded as follows:*

$$d_{TV}\big(\mathbf{p}(t), \pi\big) \leq \mathbb{P}(T_{cpl} > t). \tag{11}$$

Remark Note that the claimed inequality holds for *any* way of constructing the processes X_t and Y_t on the same probability space. The goal is to construct them cleverly so that T_{cpl} is as small as possible, which then gives the best possible bound on the mixing time. An obvious construction is to make the processes X_t and Y_t independent, but this will give very poor bounds. Instead, we need to make them dependent in the right way, as we did in the example with binomial random variables.

Proof. Suppose the X_t and Y_t process are coupled as described above, and let T_{cpl} denote the first (random) time that the two processes are in the same state; they remain in the same state for all $t > T_{cpl}$. Since the Y_t process was started in the invariant distribution, it remains in the invariant distribution for all time, i.e., $\mathbb{P}(Y_t = x) = \pi_x$ for all $x \in S$ and any fixed time $t \geq 0$. As $\mathbf{p}(t)$ denotes the distribution of the X process at time t, it follows from Lemma 4 that

$$d_{TV}\big(\mathbf{p}(t), \pi\big) \leq \mathbb{P}(X_t \neq Y_t) = \mathbb{P}(T_{cpl} > t).$$

This completes the proof of the lemma. □

Example A student has a stack of n different books on her desk. Every day, she picks one book from the stack, uniformly at random, reads from it, and replaces it at the top of the stack. Suppose that initially books are ordered alphabetically by author. What is the steady state distribution of the order of books, and after how many days is the distribution within distance ϵ of the steady state, in total variation?

Number the books $1, 2, \ldots, n$ in alphabetical order by author. Denote by X_t the permutation of books from top to bottom at the start of day t, and by $\mathbf{p}(t)$ the distribution of X_t. Then X_1 is the identity permutation $\{1, 2, \ldots, n\}$, and X_t is a discrete-time Markov chain on the state space S_n, the symmetry group of permutations of n objects. The question asks us what the invariant distribution of this Markov chain is, and how many time steps it takes to get within ϵ of the invariant distribution in total variation distance.

It should be clear by symmetry that the invariant distribution is uniform over all permutations. This can also be checked from the global balance equations. The Markov chain is not reversible, so local balance cannot be used.

We now construct a coupling and use it to obtain a bound on the ϵ-mixing time, defined as

$$T_{mix}(\epsilon) = \sup\{t > 0 : d_{TV}(\mathbf{p}(t), \pi) \geq \epsilon\}.$$

Let Y_t be another Markov chain with the same transition probabilities, but started in the invariant distribution π, namely the uniform distribution on all permutations. We couple the two Markov chains as follows: if book i is selected by the X process on day t, then the same book is selected by the Y process on day t as well (note: the same book, not the book in the same location). Note that this obeys the transition probabilities for the Y process as the choice of book is uniform at each time step, and independent of the past.

The coupling described above demonstrates that, on day $t + 1$, the same book is on top in both processes, for each $t \geq 1$, i.e., from day 2 onwards. But how long does it take until the same book is second from top, third from top and so on, all the way to the bottom of the pile? It is clear that, as soon as book i is chosen in the X process, it is also chosen in the Y process and moves to the top of the pile in both. At all subsequent times, it occupies the same position in both piles, moving down by 1 whenever some other book is selected, and moving to the top whenever it is selected. Thus, when each book has been chosen at least once in the X process, the books are in the same order in both processes. (It could happen that they reached the same order even earlier, because some of the books were in an initial condition that brought them to the same position even before they were selected for reading.) Hence, an upper bound on T_{cpl}, the first time that all books are in the same order in both processes, is the first time that every book has been picked at least once in the X process. Determining the time until this happens is called the coupon collector problem, which we study next. It yields that $T_{cpl} \approx n \log n$.

The coupon collector problem Each box of cereal comes with a coupon (e.g., a toy figure), chosen uniformly at random from a set of n coupons, and independent of the contents of other boxes (for all practical purposes). How many boxes must one open in order to collect a full set of n coupons?

This number is clearly random, but what can we say about this random variable? Let us denote by X_k the number of boxes that must be opened in order to get k distinct coupons. Clearly, $X_1 = 1$. Moreover, if we already have k coupons, then each box yields a coupon distinct from these with probability $(n-k)/n$, independent of how many boxes have already been opened or the identities of the coupons held. Hence, the number of boxes that have to be opened before the next success (new coupon) is geometrically distribution with parameter $(n-k)/n$. In other words,

$$X_{k+1} - X_k \sim Geom\Big(\frac{n-k}{n}\Big), \text{ i.e., } \mathbb{P}(X_{k+1} - X_k = j) = \Big(\frac{k}{n}\Big)^{j-1} \frac{n-k}{n}, \; j \geq 1.$$

Now, recalling that the mean of a Geom(p) random variable is $1/p$ (verify this, using generating functions or otherwise), we obtain that

$$\begin{aligned} \mathbb{E}[X_n] &= \sum_{k=1}^{n-1} \mathbb{E}[X_{k+1} - X_k] + \mathbb{E}[X_1] \\ &= \sum_{k=0}^{n-1} \frac{n}{n-k} \;=\; n \sum_{k=1}^{n} \frac{1}{k} \;\approx\; n \log n. \end{aligned}$$

Thus, we need approximately $n \log n$ boxes in order to collect n coupons.

4 Page Rank

In the last part of this chapter, we shall study Google's Page Rank metric, which is a measure of the importance of a web page based purely on its position in the web graph. Google uses this metric as one of the factors in deciding how to rank results in web searches.

The web graph is a directed graph defined as follows. Each web page is a node in the graph. If a page x has a hyperlink to a page y, this is represented by the directed edge (x, y). (The graph may have loops, namely pages with links to themselves, and multiple edges, namely multiple hyperlinks from one page to another, but we will ignore them. It is not hard to generalise the analysis to multigraphs, the term for graphs which may have loops and multiple edges.)

We start with some intuitive ideas about what a measure of the rank, or importance, of a node (web page) ought to capture. We will assume that

we want the rank of each page to be a positive real number, and that we want the rank of a page to be bigger if it is more important. This raises the question of what we mean by important, and we will build up some intuitive criteria. It is reasonable to expect that if a page is important, then many other pages will have hyperlinks pointing to it. Thus, the in-degree of a node could be one measure of its importance, and we could define the rank of a node simply as its in-degree. But this measure is arguably "too local". It might be that a page should be assigned more importance if there are many "important" pages linking to it. This suggests defining the rank r_x of a node x as

$$r_x = \sum_{y:(y,x)\in E} r_y.$$

But there are disdvantages to this definition as well. There are some important pages in the web graph which are "directory pages" for a domain, say the University of Bristol. These pages have little content of their own but have pointers to pages containing the information you may be seeking. It is less informative that a web page is listed in some directory than that it is linked to by another page on the same topic. For this reason, we might want to give less importance to links from directory pages than to links from pages with content. However, it is not possibly to automatically detect the nature of a web page without looking at its content, which would be infeasible to do for the whole web graph! So we seek a proxy indicator that a page might be a directory page. One obvious property of a directory page is that it has a large number of hyperlinks pointing out from it. We exploit this proxy information by distributing the "weight" of a page equally among its outward hyperlinks. Thus, if a page with a rank (weight) of, say 12.6, has hyperlinks to 3 pages, it will contributed 4.2 towards the rank of each of those pages. This modification leads us to define the rank r_x of a node x as

$$r_x = \sum_{y:(y,x)\in E} \frac{r_y}{d_y}, \tag{12}$$

where d_y denotes the out-degree of node y. This is the definition of Page Rank that we shall use (and is very close to the definition used by Google).

The definition is somewhat circular, in that the rank of each page is defined in terms of the rank of other pages. This raises the question of whether the page rank is defined at all (do the equations in (12) have a solution?), and then whether it is well defined (do they have a unique solution?). We now attempt to address these questions.

Let A denote the adjacency matrix of the web graph, with elements

$$a_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E \\ 0, & \text{otherwise.} \end{cases}$$

Note that A need not be symmetric, since the graph is directed. Let D be the diagonal matrix with d_i, the degree of node i, as the i^{th} diagonal entry, and let $W = D^{-1}A$. Then, we can rewrite the linear equations in (12) in matrix notation as

$$\mathbf{r} = \mathbf{r}W, \tag{13}$$

where $\mathbf{r}$ denotes the vector of page ranks. The question of whether the above equation has a unique non-negative solution for $\mathbf{r}$ is thus equivalent to asking whether the matrix W has 1 as a non-repeated eigenvalue, and whether the corresponding eigenvector is non-negative. The answers to these questions are provided by the following theorem. A matrix is said to be non-negative if all its elements are non-negative.

Theorem 2 (Perron-Frobenius theorem) *Suppose that $A \in \mathbb{R}^{n\times n}$ is a non-negative matrix and that A^k is strictly positive for some $k > 0$. Then, the following hold:*

(i) A has a positive eigenvalue λ such that $|\lambda_i| < \lambda$ for all other eigenvalues λ_i, which could be real or complex.

(ii) The eigenvector corresponding to λ is non-negative, and it is the only non-negative eigenvector.

If A is non-negative and non-zero, but there is no k such that A^k is strictly positive, then (i) holds with $|\lambda_i| \leq \lambda$ for all other eigenvalues λ_i. Moreover, the only non-negative eigenvectors are those corresponding to the positive eigenvalue λ, which could now have multiplicity bigger than one.

Remarks. The positive eigenvalue λ which has the largest absolute value among all eigenvalues of A is referred to as the Perron root or Perron eigenvalue of A. Part (ii) refers to right eigenvectors, but also applies to left eigenvectors, as can be seen by applying the theorem to A^T.

We will not prove this theorem. A proof can be found in any book on non-negative matrices, e.g., E. Seneta, *Non-negative matrices and Markov*

chains, Springer, 2006, or R. B. Bapat and T. E. S. Raghavan, *Non-negative matrices and applications*, Cambridge University Press, 1997.

Returning to page ranks, it is easy to see that the matrix $W = D^{-1}A$ is non-negative. It can also be verified that the row sums of W are all 1, so W is a stochastic matrix. In fact, it is the transition probability matrix of the Markov chain describing the following discrete-time random walk on the web graph. If the walk is currently at a node x, then in the next time step, it moves to a node y chosen uniformly at random among nodes z such that $(x, z) \in E$, i.e., nodes to which there are edges directed out of x. (We can't use the terminology of neighbours unambiguously as the graph is directed; we have to distinguish between neighbours to whom there is an edge directed outwards and from whom there is an edge directed inwards.)

For a stochastic matrix W, it is easy to check that the all-1 vector $\mathbf{1}$ is a right eigenvector with eigenvalue 1. Since it is a non-negative eigenvector, it must correspond to the Perron eigenvalue λ. In other words, $\lambda = 1$, and $|\lambda_i| \leq 1$ for all other eigenvalues. The page rank vector $\mathbf{r}$ is a left eigenvector of W with eigenvalue 1. As we have established that 1 is an eigenvalue of W, there is at least one such eigenvector. This addresses the question of whether there are page ranks, but doesn't yet tell us if they are uniquely defined.

The requirement in the Perron-Frobenius theorem that W^k be positive for some k is met if the Markov chain with transition probability matrix W is irreducible and aperiodic. It can be shown that this is equivalent to the graph G being strongly connected and non-bipartite, terms which we explain below.

A directed graph G is said to be strongly connected if, for any two nodes u and v, there is a directed path from u to v. A graph G, directed or undirected, is said to be bipartite if the vertex set V can be partitioned into disjoint sets V_1 and V_2 (i.e., $V_1 \cap V_2 = \emptyset$, $V_1 \cup V_2 = V$) such that there are no edges between two vertices in V_1 or two vertices in V_2. In other words, $E \subseteq V_1 \times V_2 \cup V_2 \times V_1$, i.e., all edges either go from a node in V_1 to a node in V_2 or the other way round.

If the graph is not strongly connected, then the discrete time random walk on it is not irreducible.The multiplicity of the Perron root $\lambda = 1$ is equal to the number of recurrent communicating classes in the Markov chain. It could be bigger than 1, in which case page ranks are not well defined.

If the graph is strongly connected but bipartite, it turns out that the random walk is irreducible and $\lambda = 1$ has multiplicity 1, so page ranks are well-defined. What happens in this case is that the random walk is a periodic Markov chain, with period 2, and there is an eigenvalue of -1, which has the same absolute value as the Perron root. This has bad consequences for the convergence of the algorithm to compute page ranks (which we don't study), but doesn't affect the uniqueness of page ranks.

In practice, it is highly unlikely that the web graph would be bipartite (finding a single triangle or odd cycle in it would rule that out), but it is quite possible that it is not strongly connected. In order to get around this, the actual Page Rank algorithm introduces a dummy node, and dummy edges from it to all other nodes. It also works with a weighted rather than unweighted adjacency matrix, with the dummy edges having very small weights. But we won't go into those details here.