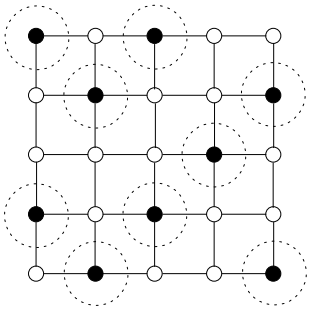

Figure 10: Hard-core colouring of a lattice.

2 Markov Chain Monte Carlo Sampling

We now introduce *Markov chain Monte Carlo (MCMC) sampling*, which is an extremely important method for dealing with "hard-to-access" distributions.

Assume that one needs to generate samples according to a probability distribution π, but π is too complicated to describe explicitly. A clever solution is then to construct a Markov chain that converges to stationary distribution π, let it run for a while and then sample states of the chain. (However, one obvious problem that this approach raises is determining how long is "for a while"? This leads to interesting considerations of the convergence rates and "rapid mixing" of Markov chains.)

Example 2.1 *The hard-core model.*

A *hard-core colouring* of a graph $G = (V, E)$ is a mapping

$$\xi : V \to \{0,1\} \qquad \text{("empty" vs. "occupied" sites)}$$

such that

$$(i,j) \in E \;\Rightarrow\; \xi(i) = 0 \vee \xi(j) = 0 \qquad \text{(no two occupied sites are adjacent)}$$

E.g. on a lattice graph, the hard-core colouring condition models an exclusion principle, whereby a "particle" at one site excludes the presence of "particles" at neighbouring sites, cf. Figure 10. In computer science terms, a hard-core colouring of a graph G corresponds to an independent set of nodes from G.

Denote by μ_G the uniform distribution over all the Z_G valid hard-core colourings of G. We would like to sample colourings according to μ_G, e.g. in order to compute the expected number of ones in a valid colouring:

$$E(n(X)) = \sum_{\xi \in \{0,1\}^V} n(\xi)\mu_G(\xi) = \frac{1}{Z_G} \sum_{\xi \in \{0,1\}^V} n(\xi) I_{[\xi \text{ is valid}]},$$

where $n(\xi)$ denotes the number of ones in colouring ξ.

However, the combinatorial structure of distribution μ_G is quite complicated; it is far from clear how one could pick a random valid hard-core colouring of graph G. (Even computing their total number Z_G is likely to be a so called #P-complete problem, and thus not solvable in polynomial time unless P = NP.)

Given a graph $G = (V,E)$, $V = \{1,\dots,n\}$, let us consider the following Markov chain $(X_0, X_1, \dots)$ on the space of valid hard-core colourings of G:

- Initially choose X_0 to be any valid hard-core colouring of G.
- Then, given colouring X_t, generate colouring X_{t+1} as follows:
  1. Choose some node $i \in V$ uniformly at random.
  2. If all the neighbours of i have colour 0 in X_t, then let $X_{t+1}(i) = 1$ with probability $1/2$ and $X_{t+1}(i) = 0$ with probability $1/2$.
  3. At all other nodes j, let $X_{t+1}(j) = X_t(j)$.

It can be seen that the chain thus determined is irreducible (since all colourings communicate via the all-zeros colouring) and aperiodic (since for any colouring ξ, $P_{\xi\xi} > 0$).

To see that the chain has μ_G as its unique stationary distribution, it suffices to check the detailed balance conditions with respect to μ_G. Let ξ, ξ' be two different colourings. If they differ at more than one node, then $P_{\xi\xi'} = P_{\xi'\xi} = 0$, so it suffices to check the case where $\xi(i) \neq \xi'(i)$ at a single node i. But then

$$\mu_G(\xi)P_{\xi\xi'} = \frac{1}{Z_G} \cdot \frac{1}{n} \cdot \frac{1}{2} = \mu_G(\xi')P_{\xi'\xi}.$$

The above hard-core sampling algorithm is a special case of a *Gibbs sampler* for a target distribution π on a state space of the form $S = C^V$.

The general principle is to choose in step 2 of the state update rule the new value for $X_{t+1}(i)$ according to the *conditional π-distribution*:

$$\Pr_{MC}(X_{t+1}(i) = c) = \Pr_\pi(\xi(i) = c \mid \xi(j) = X_t(j), j \neq i).$$

(In addition, the chain needs to be initialised in a state X_0 that has nonzero π-probability.) It can be seen that the chain so obtained is aperiodic and has π as a stationary distribution. Whether the chain is also irreducible depends on which states ξ have nonzero π-probability.

Example 2.2 *Sampling graph k-colourings.* Let $G = (V,E)$ be a graph. The following is a Gibbs sampler for the uniform distribution in the space $S = \{1,\dots,k\}^V$ of k-colourings of G:

- Initially choose X_0 to be any valid k-colouring of G. (Of course, finding a valid k-colouring is an NP-complete problem for $k \geq 3$, but let us not worry about that).
- Then, given colouring X_t, generate colouring X_{t+1} as follows:
  1. Choose some node $i \in V$ uniformly at random.
  2. Let C' be the set of colours assigned by X_t to the neighbours of i:
     $$C' = \{X_t(j) \mid (i,j) \in E\}.$$
     (Note that $|C'| < k$.) Choose a colour for $X_{t+1}(i)$ uniformly at random from the set $\{1,\dots,k\} \setminus C'$.
  3. At all other nodes j, let $X_{t+1}(j) = X_t(j)$.

Note that it is a nontrivial question whether this chain is irreducible or not.

Another general family of MCMC samplers are the *Metropolis chains*.

Let the state space S have some neighbourhood structure, so that it may be viewed as a (large) connected graph (S,N). Denote by $N(i)$ the set of neigbours of state i, and let $d_i = |N(i)|$. We assume that the neighbourhood structure is symmetric, so that $i \in N(j)$if and only if $j \in N(i)$.

Then the (basic) *Metropolis sampler* for distribution π on S operates as follows:

- Initially choose X_0 to be some state $i \in S$.
- Then, given state $X_t = i$, state X_{t+1} is obtained as follows:
  1. Choose some $j \in N(i)$ uniformly at random.
  2. With probability $\min\left\{\frac{\pi_j d_i}{\pi_i d_j}, 1\right\}$, accept $X_{t+1} = j$. Otherwise let $X_{t+1} = i$.

Thus, fully written out the transition probabilities are:

$$p_{ij} = \begin{cases} \frac{1}{d_i}\min\left\{\frac{\pi_j d_i}{\pi_i d_j}, 1\right\}, & \text{if } j \in N(i) \\ 0, & \text{if } j \notin N(i), j \neq i \\ 1 - \sum_{j \in N(i)} p_{ij}, & \text{if } j = i \end{cases}$$

To show that this chain has π as its stationary distribution, it suffices to check the detailed balance conditions:

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \forall\, i, j \in S.$$

The conditions are trivial if $i = j$ or $j \notin N(i)$, so let us consider the case $j \in N(i)$. There are two subcases:

(i) Case $\frac{\pi_j d_i}{\pi_i d_j} \geq 1$: Then:

$$\begin{cases} \pi_i p_{ij} = \pi_i \cdot \dfrac{1}{d_i} \cdot 1 \\ \pi_j p_{ji} = \pi_j \cdot \dfrac{1}{d_j} \cdot \dfrac{\pi_i d_j}{\pi_j d_i} = \dfrac{\pi_i}{d_i} \end{cases}$$

(ii) Case $\frac{\pi_j d_i}{\pi_i d_j} < 1$: Then:

$$\begin{cases} \pi_i p_{ij} = \pi_i \cdot \dfrac{1}{d_i} \cdot \dfrac{\pi_j d_i}{\pi_i d_j} = \dfrac{\pi_j}{d_j} \\ \pi_j p_{ji} = \pi_j \cdot \dfrac{1}{d_j} \cdot 1 \end{cases}$$

(Note that in both cases $\pi_i p_{ij} = \pi_j p_{ji} = \min\{\frac{\pi_i}{d_i}, \frac{\pi_j}{d_j}\}$.) Hence π is a stationary distribution of the chain.

Furthermore, the chain is guaranteed to be aperiodic if there is at least one $i \in S$ such that $\frac{\pi_j d_i}{\pi_i d_j} < 1$ ($\Rightarrow\ p_{ii} > 0$) i.e. it is not the case that for all $i, j \in S$:

$$\frac{\pi_i}{d_i} = \frac{\pi_j}{d_j} = \text{ const.}$$

In the latter case the chain reduces to a simple random walk on the graph (S, N) with stationary distribution

$$\pi = \left[\frac{d_1}{d} \quad \frac{d_2}{d} \quad \cdots \quad \frac{d_n}{d} \right]$$

as seen earlier. Such a random walk is aperiodic, if and only if the graph (S, N) contains at least one odd cycle, i.e. if (S, N) is not bipartite.

3 Estimating the Convergence Rate of a Markov Chain

3.1 Second Eigenvalue, Conductance, Canonical Paths

Consider a regular Markov Chain on state set $S = \{1, \ldots, n\}$, with transition probability matrix $P = (p_{ij})$ and stationary distribution π.

We would like to measure the rate of convergence of the chain to π, e.g. in terms of the *total variation distance*:

$$\Delta_V^{(i)}(t) = d_V(\pi^{(i,t)}, \pi),$$

where $\pi_j^{(i,t)} = p_{ij}^{(t)}$, and

$$d_V(\rho, \pi) = \max_{A \subseteq S} |\rho(A) - \pi(A)| = \frac{1}{2} \sum_{j \in S} |\rho_j - \pi_j|.$$

However, we get somewhat tighter results by using instead of d_V the *relative pointwise distance*

$$d_{rp}^U(\rho, \pi) = \max_{j \in U} \frac{|\rho_j - \pi_j|}{\pi_j}.$$

Hence, we define our convergence rate function as:

$$\Delta^U(t) = \max_{i \in U} d_{rp}^U(\pi^{(i,t)}, \pi) = \max_{i,j \in U} \frac{|p_{ij}^{(t)} - \pi_j|}{\pi_j}.$$

When we consider convergence over the whole state space, i.e. $U = S$, we denote simply:

$$\Delta(t) = \Delta^S(t).$$

Proposition 3.1 *For any two distributions ρ, π, where $\pi_j > 0$ for all j:*

$$d_V(\rho, \pi) \leq \frac{1}{2} d_{rp}(\rho, \pi) \leq \frac{1}{\min_j \pi_j} d_V(\rho, \pi).$$

Consequently, $\Delta_V^{(i)}(t) \leq \frac{1}{2}\Delta(t)$ for all i, t. □

Define the *mixing time* of a given regular chain as

$$\tau(\varepsilon) = \min\{t \mid \Delta(t') \leq \varepsilon \quad \forall\, t' \geq t\}.$$

In algorithmic applications, the details of the chain are often determined by some input x, in which case we write $\Delta_x(t)$, $\tau_x(\varepsilon)$ correspondingly.

A chain (more precisely, a family of chains determined by inputs x) is *rapidly mixing* if

$$\tau_x(\varepsilon) = \text{poly}\left(|x|, \ln\frac{1}{\varepsilon}\right).$$

Our goal is now to establish some techniques for analysing the convergence rates of Markov chains and proving them to be rapidly mixing.

Lemma 3.2 *A regular Markov chain with transition matrix P and stationary distribution π is reversible, if and only if the matrix $D^{1/2}PD^{-1/2}$ is symmetric, where $D^{1/2} = diag(\sqrt{\pi_1}, \sqrt{\pi_2}, \ldots, \sqrt{\pi_n})$.*

Proof. $D^{1/2}PD^{-1/2} = \left(D^{1/2}PD^{-1/2}\right)^T \Leftrightarrow DP = P^TD.$

Inspecting this condition coordinatewise shows that it is exactly the same as the reversibility condition $\pi_i p_{ij} = p_{ji}\pi_j \quad \forall\, i, j.$ □

Now it is easy to see that the matrix $A = D^{1/2}PD^{-1/2}$ has the same eigenvalues as P: if λ is an eigenvalue of P with left eigenvector u, then for the vector $v = uD^{-1/2}$ we obtain

$$vA = uD^{-1/2}\left(D^{1/2}PD^{-1/2}\right) = uPD^{-1/2} = \lambda uD^{-1/2} = \lambda v.$$

Since matrix A is symmetric for reversible P, this shows that reversible P have real eigenvalues. By the Perron-Frobenius theorem they can thus be ordered as

$$\lambda_1 = 1 > \lambda_2 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n > -1.$$

Denote $\lambda_{max} = (|\lambda_i| : 2 \leq i \leq n)$.

Theorem 3.3 *Let P be the transition matrix of a regular, reversible Markov chain, and other notations as above. Then for any $U \subseteq S$,*

$$\Delta^U(t) \leq \frac{\lambda_{max}^t}{\min\limits_{i\in U}\pi_i}.$$

Proof. Let $e^1, \ldots, e^n$ be an orthonormal basis for $\mathbb{R}^n$ consisting of left eigenvectors of A, where vector e^i is associated to eigenvalue λ_i. Especially, $e^1 = \pi D^{-1/2} = [\sqrt{\pi_1}, \sqrt{\pi_2}, \ldots, \sqrt{\pi_n}]$.

Then A has a spectral representation

$$A = \sum_{i=1}^{n} \lambda_i (e^i)^T e^i = \sum_{i=1}^{n} \lambda_i E_i,$$

where $E_i = (e^i)^T e^i$. Clearly $E_i^2 = E_i$, and $E_i E_j = 0$ if $i \neq j$.
Thus, for any $t \geq 0$, $A^t = \sum_{i=1}^{n} \lambda_i^t E_i$, and hence

$$\begin{aligned} P^t &= D^{-1/2} A^t D^{1/2} = \sum_{i=1}^{n} \lambda_i^t \left(D^{-1/2} (e^i)^T \right) \left(e^i D^{1/2} \right) \\ &= 1\pi + \sum_{i=2}^{n} \lambda_i^t \left(D^{-1/2} (e^i)^T \right) \left(e^i D^{1/2} \right). \end{aligned}$$

In component form, this means:

$$p_{jk}^{(t)} = \pi_k + \sqrt{\frac{\pi_k}{\pi_j}} \sum_{i=2}^{n} \lambda_i^t e_j^i e_k^i.$$

Computing the relative pointwise distance convergence rate, we thus get for any $U \subseteq S$:

$$\begin{aligned} \Delta^U(t) &= \max_{j,k \in U} \frac{\left| \sum_{i=2}^{n} \lambda_i^t e_j^i e_k^i \right|}{\sqrt{\pi_j \pi_k}} \qquad (5) \\ &\leq \lambda_{max}^t \frac{\max_{j,k \in U} \left| \sum_{i=2}^{n} e_j^i e_k^i \right|}{\min_{j \in U} \pi_j} \\ &\leq \frac{\lambda_{max}^t}{\min_{j \in U} \pi_j} \qquad \text{(by orthonormality of right eigenvectors).} \ \square \end{aligned}$$

Theorem 3.4 *With notation and assumptions as above,*

$$\Delta(t) \geq \lambda_{max}^t$$

for all even t. Moreover, if all eigenvalues of P are nonnegative, then the bound holds for all t.

Proof. Continue from equation (5) above:

$$\Delta(t) = \Delta^S(t) \geq \max_{j \in S} \frac{\left| \sum_{i=2}^{n} \lambda_i^t (e_j^i)^2 \right|}{\pi_j} \geq \lambda_{max}^t \max_{j \in S} \frac{(e_j^{i_0})^2}{\pi_j},$$

where e^{i_0} is an eigenvector corresponding to eigenvalue with absolute value λ_{max}. Necessarily $(e_j^{i_0})^2 \geq \pi_j$ for some j for otherwise

$$||e^{i_0}|| = \sum_{j=1}^{n} (e_j^{i_0})^2 < \sum_{j=1}^{n} \pi_j = 1,$$

contradicting the normality of e^{i_0}. □

Note that always $\lambda_{max} = \max(\lambda_2, |\lambda_n|)$.

Negative eigenvalues are often a nuisance, but they can always be removed, without affecting the convergence properties of the chain much, by adding appropriate self-loops to the states. E.g.:

Proposition 3.5 *With notation and assumptions as above, consider the chain determined by transition matrix $P' = \frac{1}{2}(I+P)$. This chain is then also regular and reversible, has same stationary distribution π, and its eigenvalues satisfy $\lambda_n' > 0$ and $\lambda_{max}' = \lambda_2' = \frac{1}{2}(1+\lambda_2)$.* □

With Theorem 3.3 and Proposition 3.5 in mind, it is clear that the key to analysing convergence rates of reversible Markov chains is to find good techniques for bounding the second eigenvalue λ_2 away from 1.

An interesting and intuitive approach to this task is via the notion of "conductance" of a chain.

Given a finite, regular, reversible Markov chain $\mathcal{M}$ on the state space $S = \{1, \ldots, n\}$, transition probability matrix $P = (p_{ij})$ and stationary distribution $\pi = (\pi_i)$, we associate to $\mathcal{M}$ a weighted graph $G = (S, E, W)$, where $E = \{(i, j) \mid p_{ij} > 0\}$, and the weights on the edges correspond to the *ergodic flows* between states:

$$w_{ij} = \pi_i p_{ij} = \pi_j p_{ji}.$$

Given a state set $A \subseteq S$, the *capacity* of A is defined as

$$C_A = \pi(A) = \sum_{i \in A} \pi_i,$$

and the *ergodic flow* out of A as

$$F_A = \sum_{\substack{i\in A\\ j\notin A}} \pi_i p_{ij} = \sum_{\substack{i\in A\\ j\notin A}} w_{ij} = w(A,\bar{A}).$$

(Note that $0 < F_A \leq C_A < 1$.)

Then the *conductance* of the cut $(A,\bar{A})$, or the *(weighted) expansion* of A is

$$\Phi_A = \frac{F_A}{C_A} = \frac{w(A,\bar{A})}{\pi(A)},$$

and finally the *conductance* of $\mathcal{M}$, or G, is obtained as

$$\Phi_M = \Phi(G) = \min_{0<\pi(A)\leq 1/2} \Phi_A.$$

Since clearly $F_A = F_{\bar{A}}$ for any $\varnothing \neq A \subsetneq S$, this may equally well be defined as:

$$\Phi = \min_{\varnothing\neq A\subsetneq S} \max(\Phi_A, \Phi_{\bar{A}}).$$

Theorem 3.6 *For a regular reversible Markov chain with underlying graph G, the second eigenvalue of the transition matrix satisfies:*

(i)

$$\lambda_2 \leq 1 - \frac{\Phi(G)^2}{2};$$

(ii)

$$\lambda_2 \geq 1 - 2\Phi(G).$$

Proof. Later. □

Corollary 3.7 *With notation and assumptions as above, the convergence rates for the chain under consideration satisfy, for any $\varnothing \neq A \subsetneq S$ and $t \geq 0$:*

(i)

$$\Delta^A(t) \leq \frac{\left(1-\Phi^2/2\right)^t}{\min\limits_{i\in A}\pi_i};$$

(ii)

$$\Delta(t) \geq (1-2\Phi)^t .$$

Corollary 3.8 *Consider a family of regular reversible chains where all eigenvalues are nonnegative, parameterised by some input string x, and with underlying graphs G_x. Then the chains are rapidly mixing, if and only if*

$$\Phi(G_x) \geq \frac{1}{p(|x|)},$$

for some polynomial p and all x.

Proof. According to Corollary 3.7 (i):

$$\begin{array}{lrcll}
 & \Delta(t) & \leq & \varepsilon & \\
\text{if} & \frac{\left(1-\Phi^2/2\right)^t}{\min_{i\in A}\pi_i} & \leq & \varepsilon & \\
\text{if} & t\cdot \underbrace{\ln\left(1-\frac{\Phi^2}{2}\right)}_{\leq -\Phi^2/2} & \leq & \ln\varepsilon + \ln\pi_{min} & \\
\text{if} & -t\Phi^2/2 & \leq & \ln\varepsilon + \ln\pi_{min} & \\
\text{if} & t & \geq & \frac{2}{\Phi^2}\left(\ln\frac{1}{\varepsilon} + \ln\frac{1}{\pi_{min}}\right). &
\end{array}$$

Conversely, by Theorem 3.4 and Corollary 3.7 (ii):

$$\begin{array}{lrcll}
 & \Delta(t) & > & \varepsilon & \\
\text{if} & \lambda_2^t & > & \varepsilon & \\
\text{if} & t\ln\lambda_2 & > & \ln\varepsilon & \\
\text{if} & t\ln\frac{1}{\lambda_2} & < & \ln\frac{1}{\varepsilon} & \\
\text{if} & t\cdot\frac{1-\lambda_2}{\lambda_2} & < & \ln\frac{1}{\varepsilon} & \ln\frac{1}{\lambda} = \ln\left(1+\frac{1-\lambda}{\lambda}\right) \leq \frac{1-\lambda}{\lambda},\ 0<\lambda\leq\lambda_2 \\
\text{if} & t & < & \frac{\lambda_2}{1-\lambda_2}\cdot\ln\frac{1}{\varepsilon} & \\
\text{if} & t & < & \frac{1-2\Phi}{2\Phi}\ln\frac{1}{\varepsilon} & \frac{\lambda}{1-\lambda}\nearrow,\ 1-2\Phi\leq\lambda_2.
\end{array}$$

Consequently,

$$\frac{1-2\Phi(G_x)}{2\Phi(G_x)}\ln\frac{1}{\varepsilon} \leq \tau_x(\varepsilon) \leq \frac{2}{\Phi(G_x)^2}\left(\ln\frac{1}{\varepsilon}+\ln\frac{1}{\pi^x_{min}}\right).\square$$

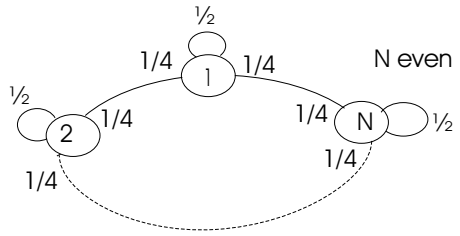

Figure 11: A simple cyclic random walk.

Example 3.1 *Simple cyclic random walk.* Consider the regular, reversible Markov chain described by the graph in Figure 11.

Clearly the stationary distribution is $\pi = [\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}]$.

The conductance $\Phi_A = F_A/C_A$ of a cut $(A, \bar{A})$ is minimised by choosing A to consist of any $n/2$ consecutive nodes on the cycle, e.g. $A = \{1, 2, \ldots, n/2\}$. Then

$$\Phi = \Phi_A = \frac{F_A}{C_A} = \frac{\sum_{\substack{i \in A \\ j \notin A}} \pi_i p_{ij}}{\sum_{i \in A} \pi_i} = \frac{2 \cdot \frac{1}{n} \cdot \frac{1}{4}}{\frac{n}{2} \cdot \frac{1}{n}} = \frac{1/2n}{1/2} = \frac{1}{n}.$$

Thus, by Theorem 3.6, the second eigenvalue satisfies:

$$1 - \frac{2}{n} \leq \lambda_2 \leq 1 - \frac{1}{2n^2},$$

by Corollary 3.7, the convergence rate satisfies

$$\left(1 - \frac{2}{n}\right)^t \leq \Delta(t) \leq n \cdot \left(1 - \frac{1}{2n^2}\right)^t,$$

and by Corollary 3.8, the mixing time satisfies:

$$\begin{aligned} & \frac{1 - 2/n}{2/n} \ln \frac{1}{\varepsilon} \leq \tau(\varepsilon) \leq 2n^2 \left(\ln \frac{1}{\varepsilon} + \ln n\right) \\ \Leftrightarrow \quad & \left(\frac{n}{2} - 1\right) \cdot \ln \frac{1}{\varepsilon} \leq \tau(\varepsilon) \leq 2n^2 \left(\ln n + \ln \frac{1}{\varepsilon}\right). \end{aligned}$$

Let us now return to the proof of Theorem 3.6, establishing the connection between the second-largest eigenvalue and the conductance of a Markov chain. Recall the statement of the Theorem:

Theorem 3.6 *Let $\mathcal{M}$ be a finite, regular, reversible Markov chain and λ_2 the second-largest eigenvalue of its transition probability matrix. Then:*