Probability on Graphs: Techniques and Applications to Data Science

3 - Spectral Techniques

Sébastien Roch
UW–Madison
Mathematics

July 26, 2018

Random walk on a network

Definition

Let $G = (V, E)$ be a graph. Let $c : E \to \mathbb{R}_+$ be a positive edge weight function on G. We call $\mathcal{N} = (G, c)$ a *network*. Random walk on $\mathcal{N}$ is the Markov chain on V, started at an arbitrary vertex, which at each time picks a neighbor of the current state proportionally to the weight of the corresponding edge.

Transition matrix

Let (X_t) be a a Markov chain on V and let

$$P^t(x,y) := \mathbb{P}[X_t = y \,|\, X_0 = x].$$

The one-step probabilities $P(x,y) := P^1(x,y)$ are the elements of its *transition matrix* $P = (P(x,y))_{x,y}$. We have

$$\mathbb{P}_\mu[X_0 = x_0, \ldots, X_t = x_t] = \mu(x_0)P(x_0,x_1)\cdots P(x_{t-1},x_t),$$

and $P^t(x,y) = (P^t)_{x,y}$.

Stationary distribution I

Definition (Stationary distribution)

Let (X_t) be a Markov chain with transition matrix P. A *stationary measure* π is a measure such that

$$\sum_{x \in V} \pi(x) P(x, y) = \pi(y), \qquad \forall y \in V,$$

or in matrix form $\pi = \pi P$. We say that π is a *stationary distribution* if in addition π is a probability measure.

When P is *irreducible*, i.e. $\forall x, y, \exists t$ s.t. $P^t(x, y) > 0$, then the stationary distribution is unique and positive. This is the case for a random walk on a connected network.

Stationary distribution II

Definition (Reversible chain)

A transition matrix P is *reversible* w.r.t. a measure η if $\eta(x)P(x,y) = \eta(y)P(y,x)$ for all $x, y \in V$. By summing over y, one sees such a measure is necessarily stationary.

Let (X_t) be random walk on a network $\mathcal{N} = (G, c)$. Then (X_t) is reversible w.r.t. $\eta(v) := c(v)$, where

$$c(v) := \sum_{x \sim v} c(v,x).$$

If all edge weights are 1, then $\eta(v) := \delta(v)$.

Convergence I

A transition matrix P is *aperiodic* if, for all x, $P^t(x,x) > 0$ for all sufficiently large t. The *lazy walk* on $\mathcal{N}$ is the Markov chain that, at each time, stays put with probability $1/2$ or else takes a step according to the random walk on $\mathcal{N}$. This modified walk is aperiodic.

Theorem (Convergence to stationarity)

Suppose P is irreducible, aperiodic and has stationary distribution π. Then, for all x, y, $P^t(x,y) \to \pi(y)$ as $t \to +\infty$.

Convergence II

For probability measures μ, ν on V, let their *total variation distance* be $\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A \subseteq V} |\mu(A) - \nu(A)|$.

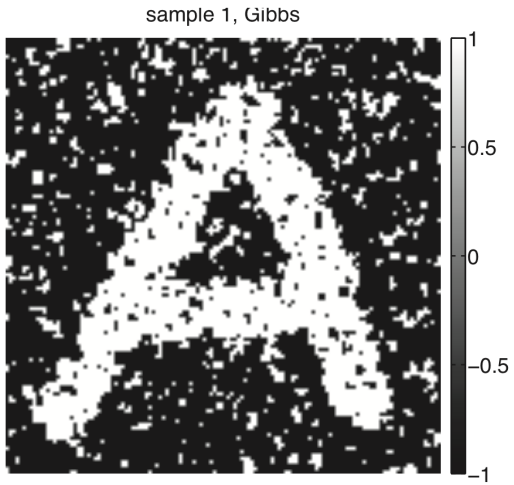
Definition (Mixing time)

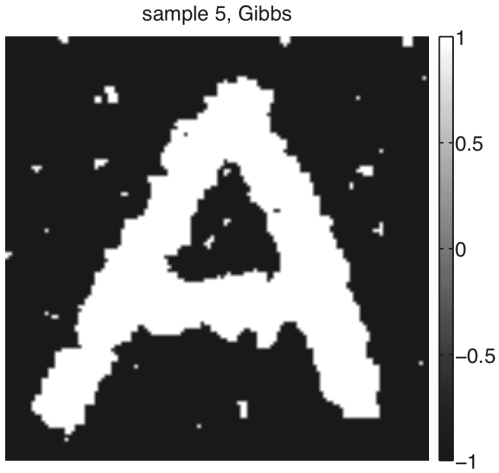
The *mixing time* is $t_{\mathrm{mix}}(\varepsilon) := \min\{t \geq 0 : d(t) \leq \varepsilon\}$, where $d(t) := \max_{x \in V} \|P^t(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}}$.

Other useful random walk quantities

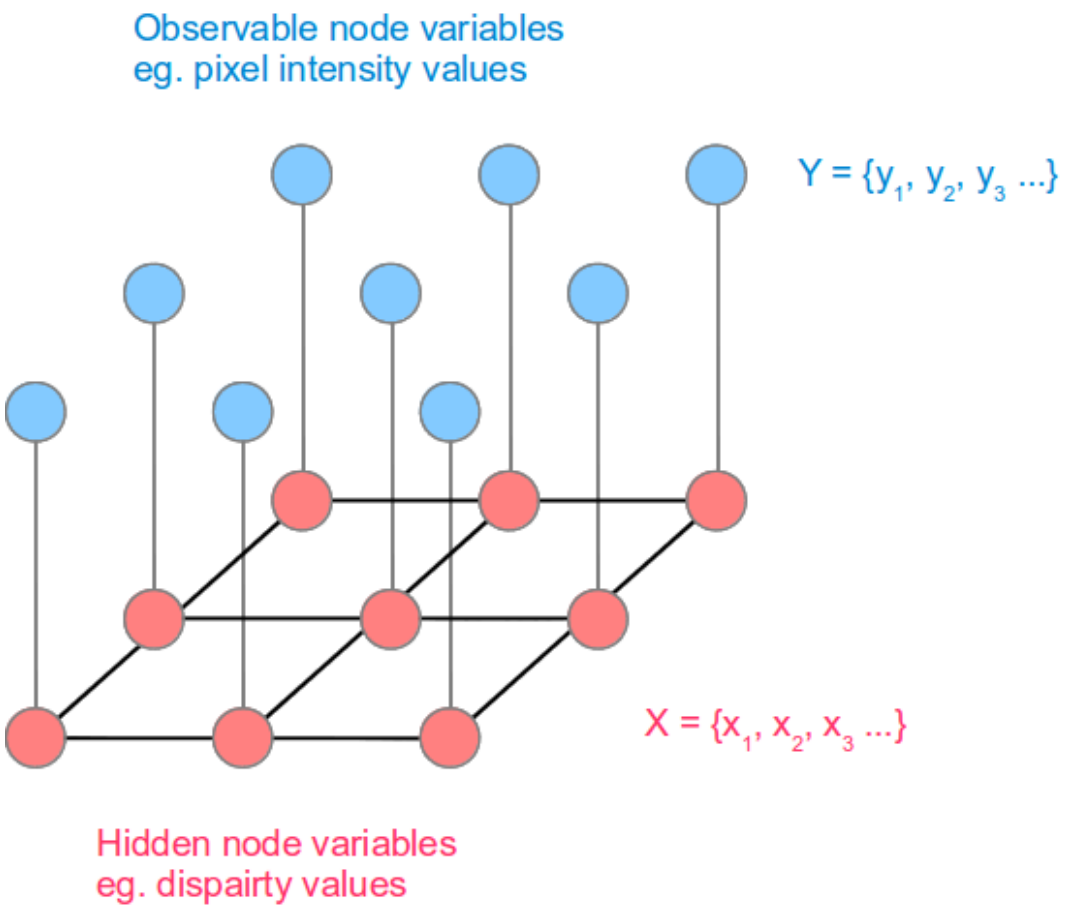
- Hitting times
- Cover times
- Heat kernels

Application: Bayesian image analysis I





Application: Bayesian image analysis II



Application: Undirected graphical models I

Definition

Let S be a finite set and let $G = (V, E)$ be a finite graph. Denote by $\mathcal{K}$ the set of all cliques of G. A positive probability measure μ on $\mathcal{X} := S^V$ is called a *Gibbs random field* if there exist *clique potentials* $\phi_K : S^K \to \mathbb{R}$, $K \in \mathcal{K}$, such that

$$\mu(x) = \frac{1}{\mathcal{Z}} \exp\left(\sum_{K \in \mathcal{K}} \phi_K(x_K)\right),$$

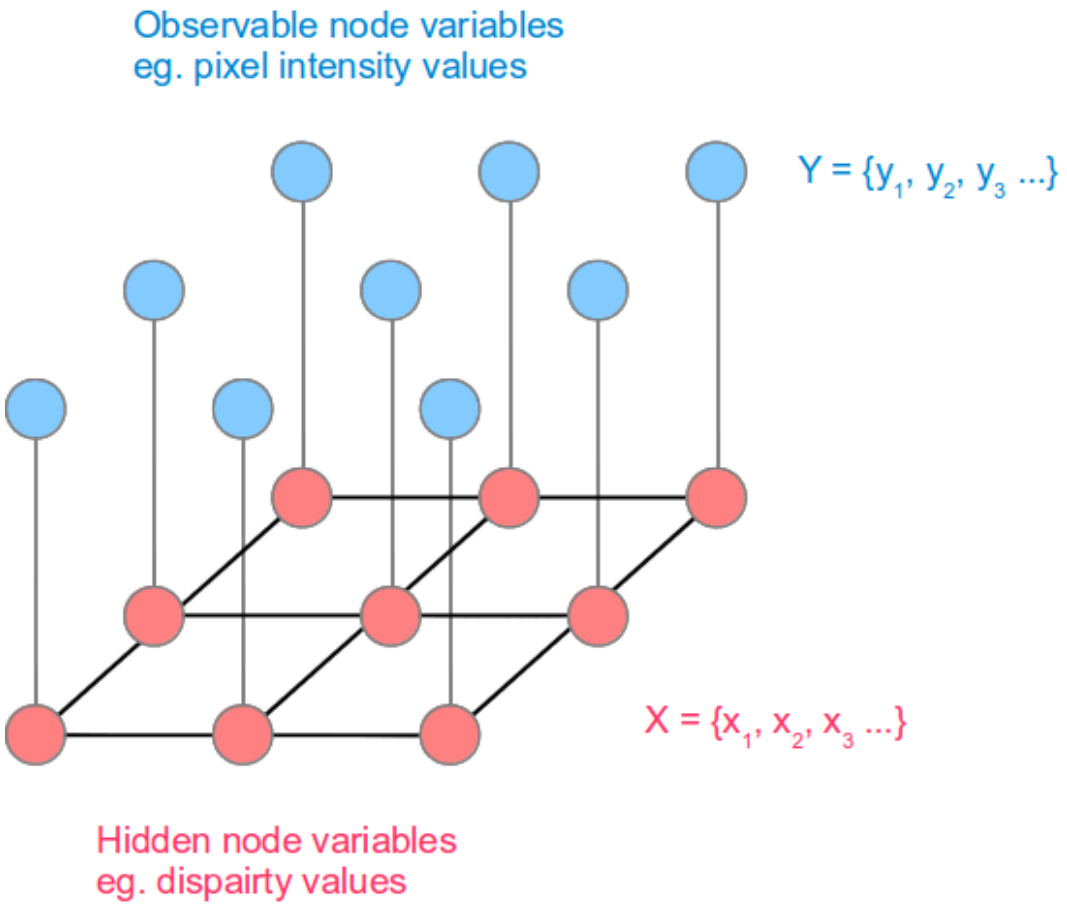
where x_K is x restricted to the vertices of K and $\mathcal{Z}$ is a normalizing constant.

Application: Undirected graphical models II

Example

For $\beta > 0$, the *ferromagnetic Ising model* with inverse temperature β is the Gibbs random field with $S := \{-1, +1\}$, $\phi_{\{i,j\}}(\sigma_{\{i,j\}}) = \beta\sigma_i\sigma_j$ and $\phi_K \equiv 0$ if $|K| \neq 2$. The function $\mathcal{H}(\sigma) := -\sum_{\{i,j\}\in E} \sigma_i\sigma_j$ is known as the *Hamiltonian*. The normalizing constant $\mathcal{Z} := \mathcal{Z}(\beta)$ is called the *partition function*. The states $(\sigma_i)_{i\in V}$ are referred to as *spins*.

Application: Back to Bayesian image analysis I



Application: Back to Bayesian image analysis II

We assume the prior (i.e. distribution of hidden variables) is an Ising model $\mu_\beta(\sigma)$ on the $L \times L$ grid $G = (V, E)$. The observed variables τ are independent flips of the corresponding hidden variables with flip probability $q \in (0, 1/2)$, i.e.,

$$\begin{aligned}
\mathbb{P}[\tau \,|\, \sigma] &= \prod_{i \in V} (1-q)^{\mathbf{1}_{\tau_i = \sigma_i}} q^{\mathbf{1}_{\tau_i \neq \sigma_i}} \\
&= \exp\left(\sum_{i \in V} \left\{ \log(1-q) \frac{1 + \sigma_i \tau_i}{2} + \log(q) \frac{1 - \sigma_i \tau_i}{2} \right\} \right) \\
&= \exp\left(\sum_{i \in V} \sigma_i \frac{\tau_i}{2} \log \frac{1-q}{q} + \mathcal{Y}(q) \right).
\end{aligned}$$

Application: Back to Bayesian image analysis III

By Bayes' rule, the posterior is then given by

$$\begin{aligned}\mathbb{P}[\sigma \,|\, \tau] &= \frac{\mathbb{P}[\tau \,|\, \sigma]\mu_\beta(\sigma)}{\sum_\sigma \mathbb{P}[\tau \,|\, \sigma]\mu_\beta(\sigma)} \\ &= \frac{1}{\mathcal{Z}(\beta, q)} \exp\left(\beta \sum_{i\sim j} \sigma_i\sigma_j + \sum_i h_i\sigma_i\right),\end{aligned}$$

where $h_i = \frac{\tau_i}{2} \log \frac{1-q}{q}$.

Application: Gibbs sampling I

Definition

Let μ_β be the Ising model with inverse temperature $\beta > 0$ on a graph $G = (V, E)$. The *(single-site) Glauber dynamics* is the Markov chain on $\mathcal{X} := \{-1, +1\}^V$ which at each time:

- selects a site $i \in V$ uniformly at random, and
- updates the spin at i according to μ_β conditioned on agreeing with the current state at all sites in $V \backslash \{i\}$.

Application: Gibbs sampling II

Specifically, for $\gamma \in \{-1, +1\}$, $i \in \Lambda$, and $\sigma \in \mathcal{X}$, let $\sigma^{i,\gamma}$ be the configuration σ with the spin at i being set to γ. Let $n = |V|$ and $S_i(\sigma) := \sum_{j \sim i} \sigma_j$. Then

$$\begin{aligned} Q_\beta(\sigma, \sigma^{i,\gamma}) &:= \frac{1}{n} \frac{\frac{1}{\mathcal{Z}(\beta)} \exp\left(\beta \sum_{j \sim k} \sigma_j^{i,\gamma} \sigma_k^{i,\gamma}\right)}{\sum_{i'=-,+} \frac{1}{\mathcal{Z}(\beta)} \exp\left(\beta \sum_{j \sim k} \sigma_j^{i',\gamma} \sigma_k^{i',\gamma}\right)} \\ &= \frac{1}{n} \cdot \frac{e^{\gamma \beta S_i(\sigma)}}{e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}}. \end{aligned}$$

The Glauber dynamics is reversible w.r.t. μ_β. How quickly does the chain approach μ_β?

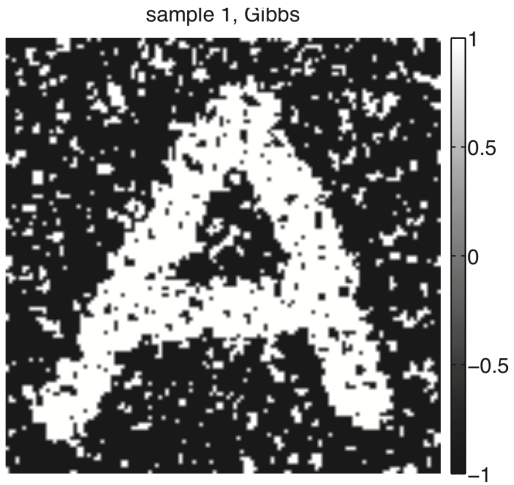
Application: Gibbs sampling III

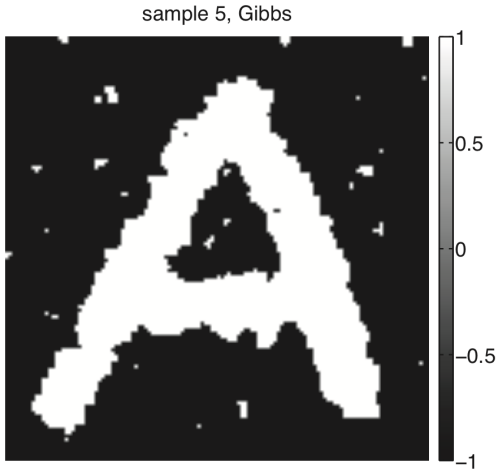
Proof of reversibility: This chain is clearly irreducible. For all $\sigma \in \mathcal{X}$ and $i \in V$, let $S_{\neq i}(\sigma) := \mathcal{H}(\sigma^{i,+}) + S_i(\sigma) = \mathcal{H}(\sigma^{i,-}) - S_i(\sigma)$. We have

$$\begin{aligned}
\mu_\beta(\sigma^{i,-})\, Q_\beta(\sigma^{i,-}, \sigma^{i,+}) &= \frac{e^{-\beta S_{\neq i}(\sigma)} e^{-\beta S_i(\sigma)}}{\mathcal{Z}(\beta)} \cdot \frac{e^{\beta S_i(\sigma)}}{n[e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}]} \\
&= \frac{e^{-\beta S_{\neq i}(\sigma)}}{n\mathcal{Z}(\beta)[e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}]} \\
&= \frac{e^{-\beta S_{\neq i}(\sigma)} e^{\beta S_i(\sigma)}}{\mathcal{Z}(\beta)} \cdot \frac{e^{-\beta S_i(\sigma)}}{n[e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}]} \\
&= \mu_\beta(\sigma^{i,+})\, Q_\beta(\sigma^{i,+}, \sigma^{i,-}).
\end{aligned}$$

■

Application: Back to Bayesian image analysis





Eigenbasis I

Let P be the transition matrix of an irreducible, reversible Markov chain with stationary distribution $\pi > 0$. Define

$$\langle f, g \rangle_\pi := \sum_{x \in V} \pi(x) f(x) g(x), \quad \|f\|_\pi^2 := \langle f, f \rangle_\pi,$$
$$(Pf)(x) := \sum_y P(x, y) f(y).$$

We let $\ell^2(V, \pi)$ be the Hilbert space of real-valued functions on V equipped with the inner product $\langle \cdot, \cdot \rangle_\pi$ (equivalent to the vector space $(\mathbb{R}^n, \langle \cdot, \cdot \rangle_\pi)$).

Theorem

There is an orthonormal basis of $\ell^2(V, \pi)$ formed of eigenfunctions $\{f_j\}_{j=1}^n$ of P with real eigenvalues $\{\lambda_j\}_{j=1}^n$. We can take $f_1 \equiv 1$ and $\lambda_1 = 1$.

Eigenbasis II

Proof: Let D_π be the diagonal matrix with π on the diagonal. By reversibility,

$$M(x,y) := \sqrt{\frac{\pi(x)}{\pi(y)}}P(x,y) = \sqrt{\frac{\pi(y)}{\pi(x)}}P(y,x) =: M(y,x).$$

So $M = (M(x,y))_{x,y} = D_\pi^{1/2} P D_\pi^{-1/2}$ is symmetric and has orthonormal eigenvectors $\{\phi_j\}_{j=1}^n$ and real eigenvalues $\{\lambda_j\}_{j=1}^n$. Define $f_j := D_\pi^{-1/2}\phi_j$. Then

$$Pf_j = PD_\pi^{-1/2}\phi_j = D_\pi^{-1/2}D_\pi^{1/2}PD_\pi^{-1/2}\phi_j = D_\pi^{-1/2}M\phi_j = \lambda_j D_\pi^{-1/2}\phi_j = \lambda_j f_j,$$

and

$$\begin{aligned}\langle f_i, f_j\rangle_\pi &= \langle D_\pi^{-1/2}\phi_i, D_\pi^{-1/2}\phi_j\rangle_\pi \\ &= \sum_x \pi(x)[\pi(x)^{-1/2}\phi_i(x)][\pi(x)^{-1/2}\phi_j(x)] = \langle \phi_i, \phi_j\rangle.\end{aligned}$$

Because P is stochastic, the all-one vector is a right eigenvector of P with eigenvalue 1.

Spectral decomposition I

Theorem

Let $\{f_j\}_{j=1}^n$ be the eigenfunctions of a reversible and irreducible transition matrix P with corresponding eigenvalues $\{\lambda_j\}_{j=1}^n$, as defined previously. Assume $\lambda_1 \geq \cdots \geq \lambda_n$. We have the decomposition

$$\frac{P^t(x,y)}{\pi(y)} = 1 + \sum_{j=2}^{n} f_j(x) f_j(y) \lambda_j^t.$$

Spectral decomposition II

Proof: Let F be the matrix whose columns are the eigenvectors $\{f_j\}_{j=1}^n$ and let D_λ be the diagonal matrix with $\{\lambda_j\}_{j=1}^n$ on the diagonal. Using the notation of the eigenbasis theorem,

$$D_\pi^{1/2} P^t D_\pi^{-1/2} = M^t = (D_\pi^{1/2} F) D_\lambda^t (D_\pi^{1/2} F)',$$

which after rearranging becomes

$$P^t D_\pi^{-1} = F D_\lambda^t F'.$$

■

Eigenvalues

Lemma

Any eigenvalue λ of P satisfies $|\lambda| \leq 1$.

Proof: $Pf = \lambda f \implies |\lambda| \|f\|_\infty = \|Pf\|_\infty = \max_x |\sum_y P(x,y) f(y)| \leq \|f\|_\infty$ ■

We order the eigenvalues $1 \geq \lambda_1 \geq \cdots \geq \lambda_n \geq -1$.

Spectral gap

Definition (Spectral gap)

The *(absolute) spectral gap* is $\gamma_* := 1 - |\lambda_2| \vee |\lambda_n|$. The *relaxation time* is defined as $t_{\text{rel}} := \gamma_*^{-1}$.

Note that the eigenvalues of the lazy version $\frac{1}{2}P + \frac{1}{2}I$ of P are $\{\frac{1}{2}(\lambda_j + 1)\}_{j=1}^n$ which are all nonnegative.

Theorem

Let P be reversible, irreducible, and aperiodic with stationary distribution π. Let $\pi_{\min} = \min_x \pi(x)$. For all $\varepsilon > 0$,

$$(t_{\text{rel}} - 1) \log\left(\frac{1}{2\varepsilon}\right) \leq t_{\text{mix}}(\varepsilon) \leq \log\left(\frac{1}{\varepsilon \pi_{\min}}\right) t_{\text{rel}}.$$

Example: Random walk on the cycle I

Consider random walk on an n-cycle. That is, $V := \{0, 1, \ldots, n-1\}$ and $P(x, y) = 1/2$ if and only if $|x - y| = 1 \mod n$.

Lemma (Eigenbasis on the cycle)

For $j = 0, \ldots, n-1$, the function

$$f_j(x) := \cos\left(\frac{2\pi j x}{n}\right), \qquad x = 0, 1, \ldots, n-1,$$

is an eigenfunction of P with eigenvalue

$$\lambda_j := \cos\left(\frac{2\pi j}{n}\right).$$

Example: Random walk on the cycle II

Proof: Note that, for all i, x,

$$\begin{aligned}
\sum_y P(x,y) f_j(y) &= \frac{1}{2}\left[\cos\left(\frac{2\pi j(x-1)}{n}\right) + \cos\left(\frac{2\pi j(x+1)}{n}\right)\right] \\
&= \frac{1}{2}\left[\frac{e^{i\frac{2\pi j(x-1)}{n}} + e^{-i\frac{2\pi j(x-1)}{n}}}{2} + \frac{e^{i\frac{2\pi j(x+1)}{n}} + e^{-i\frac{2\pi j(x+1)}{n}}}{2}\right] \\
&= \left[\frac{e^{i\frac{2\pi jx}{n}} + e^{-i\frac{2\pi jx}{n}}}{2}\right]\left[\frac{e^{i\frac{2\pi j}{n}} + e^{-i\frac{2\pi j}{n}}}{2}\right] \\
&= \left[\cos\left(\frac{2\pi jx}{n}\right)\right]\left[\cos\left(\frac{2\pi j}{n}\right)\right] \\
&= \cos\left(\frac{2\pi j}{n}\right) f_j(x).
\end{aligned}$$

■

Example: Random walk on the cycle III

Theorem (Relaxation time on the cycle)

The relaxation time for lazy random walk on the n-cycle is

$$t_{\mathrm{rel}} = \frac{2}{1 - \cos\left(\frac{2\pi}{n}\right)} = \Theta(n^2).$$

Proof: The eigenvalues are $\frac{1}{2}\left[\cos\left(\frac{2\pi j}{n}\right) + 1\right]$. The spectral gap is therefore $\frac{1}{2}(1 - \cos\left(\frac{2\pi}{n}\right))$. By a Taylor expansion,

$$1 - \cos\left(\frac{2\pi}{n}\right) = \frac{4\pi^2}{n^2} + O(n^{-4}).$$

■

Since $\pi_{\min} = 1/n$, we get $t_{\mathrm{mix}}(\varepsilon) = O(n^2 \log n)$ and $t_{\mathrm{mix}}(\varepsilon) = \Omega(n^2)$.

Back to eigenvalues I

Theorem (Rayleigh's quotient)

Let P be irreducible and reversible with respect to π. The second largest eigenvalue is characterized by

$$\lambda_2 = \sup \left\{ \frac{\langle f, Pf \rangle_\pi}{\langle f, f \rangle_\pi} \ : \ f \in \ell^2(V, \pi), \ \sum_x \pi(x) f(x) = 0 \right\}.$$

Proof: Recalling that $f_1 \equiv 1$, the condition $\sum_x \pi(x) f(x) = 0$ is equivalent to $\langle f_1, f \rangle_\pi = 0$.

Back to eigenvalues II

For such an f, the eigendecomposition is

$$f = \sum_{j=1}^{n} \langle f, f_j \rangle_\pi f_j = \sum_{j=2}^{n} \langle f, f_j \rangle_\pi f_j,$$

and

$$Pf = \sum_{j=2}^{n} \langle f, f_j \rangle_\pi \lambda_j f_j,$$

so that

$$\frac{\langle f, Pf \rangle_\pi}{\langle f, f \rangle_\pi} = \frac{\sum_{i=2}^{n} \sum_{j=2}^{n} \langle f, f_i \rangle_\pi \langle f, f_j \rangle_\pi \lambda_j \langle f_i, f_j \rangle_\pi}{\sum_{j=2}^{n} \langle f, f_j \rangle_\pi^2} = \frac{\sum_{j=2}^{n} \langle f, f_j \rangle_\pi^2 \lambda_j}{\sum_{j=2}^{n} \langle f, f_j \rangle_\pi^2} \leq \lambda_2.$$

Taking $f = f_2$ achieves the supremum. ■

Dirichlet energy I

Note that

$$
\begin{aligned}
&2\langle f, (I-P)f\rangle_\pi \\
&\quad = \sum_x \pi(x) f(x)^2 + \sum_y \pi(y) f(y)^2 - 2\sum_x \pi(x) f(x) f(y) P(x,y) \\
&\quad = \sum_{x,y} f(x)^2 \pi(x) P(x,y) + \sum_{x,y} f(y)^2 \pi(y) P(y,x) - 2\sum_x \pi(x) f(x) f(y) P(x,y) \\
&\quad = \sum_{x,y} f(x)^2 \pi(x) P(x,y) + \sum_{x,y} f(y)^2 \pi(x) P(x,y) - 2\sum_x \pi(x) f(x) f(y) P(x,y) \\
&\quad = 2\mathcal{E}(f)
\end{aligned}
$$

where the *Dirichlet energy* is defined as (using $c(x,y) = \pi(x)P(x,y)$)

$$\mathcal{E}(f) := \frac{1}{2}\sum_{x,y} c(x,y)[f(x) - f(y)]^2.$$

Dirichlet energy II

We note further that if $\sum_x \pi(x)f(x) = 0$ then

$$\langle f, f\rangle_\pi = \langle f - \langle \mathbf{1}, f\rangle_\pi, f - \langle \mathbf{1}, f\rangle_\pi\rangle_\pi = \mathrm{Var}_\pi[f],$$

where the last expression denotes the variance under π. So the variational characterization of λ_2 translates into

$$\gamma \le \frac{\mathcal{E}(f)}{\mathrm{Var}_\pi[f]} = \frac{\frac{1}{2}\sum_{x,y} c(x,y)[f(x)-f(y)]^2}{\mathrm{Var}_\pi[f]},$$

where $\gamma = 1 - \lambda_2$, for all f such that $\sum_x \pi(x)f(x) = 0$ (in fact for any f by considering $f - \langle \mathbf{1}, f\rangle_\pi$ and noticing that both numerator and denominator are unaffected by adding a constant).

Bottleneck ratio I

Let $\mathcal{N} = (G, c)$ be a finite or infinite network with $G = (V, E)$. For a subset $S \subseteq V$, we let the *edge boundary* of S be

$$\partial_{\mathrm{E}} S := \{e = (x, y) \in E \,:\, x \in S, y \in S^c\}.$$

Let $g : E \to \mathbb{R}_+$ be an edge weight function. For $F \subseteq E$ we define

$$|F|_g := \sum_{e \in F} g(e).$$

For $S \subseteq V$, we let

$$\Phi_{\mathrm{E}}(S; g, h) := \frac{|\partial_{\mathrm{E}} S|_g}{|S|_h}.$$

Bottleneck ratio II

For disjoint subsets $S_0, S_1 \subseteq V$, we let $c(S_0, S_1) := \sum_{x_0 \in S_0} \sum_{x_1 \in S_1} c(x_0, x_1)$.

Definition (Bottleneck ratio)

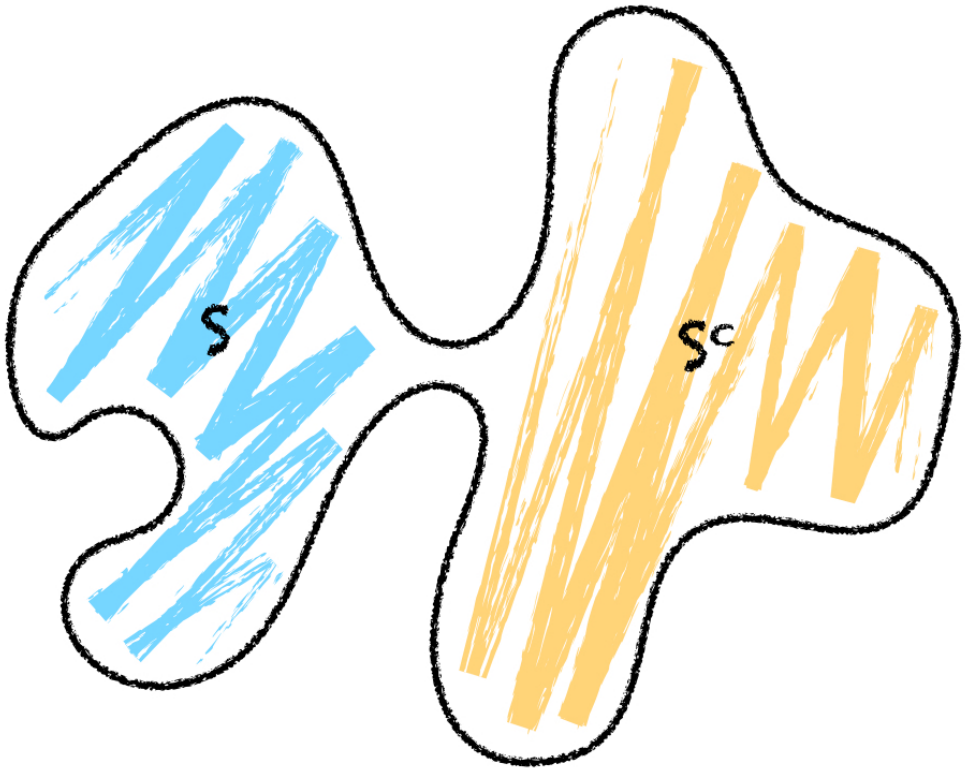
For a subset of states $S \subseteq V$, the *bottleneck ratio* of S is

$$\Phi_{\mathrm{E}}(S; c, \pi) = \frac{|\partial_{\mathrm{E}} S|_c}{|S|_\pi} = \frac{c(S, S^c)}{\pi(S)}.$$

The *bottleneck ratio* of $\mathcal{N}$ is

$$\Phi_* := \min \left\{ \Phi_{\mathrm{E}}(S; c, \pi) \, : \, S \subseteq V, \, 0 < \pi(S) \leq \frac{1}{2} \right\}.$$

A bottleneck



Example: Clique

Example

Let $G = K_n$ be the clique on n vertices and assume $c(x, y) = 1$ for all $x \neq y$. For simplicity, take n even. Then for a subset S of size $|S| = k$,

$$\Phi_{\mathrm{E}}(S; c, \pi) = \frac{|\partial_{\mathrm{E}} S|_c}{|S|_\pi} = \frac{k(n-k)}{k/n} = \frac{n-k}{n}.$$

Thus, the minimum is achieved for $k = n/2$ and

$$\Phi_* = \frac{n - n/2}{n} = \frac{1}{2}.$$

Cheeger's inequality

Theorem (Spectral gap and the bottleneck ratio)

Let P be a finite, irreducible, reversible Markov transition matrix and let $\gamma = 1 - \lambda_2$ be the spectral gap of P. Then

$$\frac{\Phi_*^2}{2} \leq \gamma \leq 2\Phi_*.$$

In terms of the relaxation time $t_{\text{rel}} = \gamma^{-1}$, these inequalities have an intuitive meaning: the presence or absence of a strong bottleneck in the state space leads to slow or fast mixing respectively.

Cheeger's inequality: Proof I

Proof: We only prove the upper bound. To get an upper bound on For $S \subseteq V$ with $\pi(S) \in (0, 1/2]$, we let

$$f_S(x) := \begin{cases} -\sqrt{\frac{\pi(S^c)}{\pi(S)}}, & x \in S, \\ \sqrt{\frac{\pi(S)}{\pi(S^c)}}, & x \in S^c. \end{cases}$$

Then

$$\sum_x \pi(x) f_S(x) = \pi(S) \left[-\sqrt{\frac{\pi(S^c)}{\pi(S)}} \right] + \pi(S^c) \left[\sqrt{\frac{\pi(S)}{\pi(S^c)}} \right] = 0,$$

and

$$\sum_x \pi(x) f_S(x)^2 = \pi(S) \left[-\sqrt{\frac{\pi(S^c)}{\pi(S)}} \right]^2 + \pi(S^c) \left[\sqrt{\frac{\pi(S)}{\pi(S^c)}} \right]^2 = 1.$$

Cheeger's inequality:Proof II

Proof (continued): From the variational characterization,

$$\gamma \leq \frac{\mathcal{E}(f_S)}{\mathrm{Var}_\pi[f_S]} = \mathcal{E}(f_S)$$

$$= \frac{1}{2}\sum_{x,y} c(x,y)[f_S(x) - f_S(y)]^2 = \sum_{x\in S, y\in S^c} c(x,y)\left[\sqrt{\frac{\pi(S^c)}{\pi(S)}} + \sqrt{\frac{\pi(S)}{\pi(S^c)}}\right]^2$$

$$= \frac{c(S,S^c)}{\pi(S)\pi(S^c)} \leq 2\frac{c(S,S^c)}{\pi(S)}.$$

■

Example: Cycle I

Let (Z_t) be lazy random walk on the n-cycle. Assume n is even. Consider a subset of vertices S. Note by symmetry $\pi(S) = \frac{|S|}{n}$. Moreover, for all $i \sim j$, $c(i,j) = \pi(i)P(i,j) = \frac{1}{n} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4n}$. Among all sets of size $|S|$, consecutive vertices minimize the size of the boundary. So

$$\Phi_* \leq \frac{2\frac{1}{4n}}{\frac{\ell}{n}} = \frac{1}{2\ell},$$

for all $\ell \leq n/2$. This expression is minimized for $\ell = n/2$ so

$$\Phi_* = \frac{1}{n}.$$

Example: Cycle II

By Cheeger's inequality,

$$\frac{1}{2n^2} = \frac{\Phi_*^2}{2} \le \gamma \le 2\Phi_* = \frac{2}{n}$$

and

$$\frac{n}{2} \le t_{\mathrm{rel}} = \gamma^{-1} \le 2n^2.$$

Thus

$$t_{\mathrm{mix}}(\varepsilon) \ge (t_{\mathrm{rel}} - 1) \log\left(\frac{1}{2\varepsilon}\right) = \Omega(n),$$

and

$$t_{\mathrm{mix}}(\varepsilon) \le \log\left(\frac{1}{\varepsilon \pi_{\min}}\right) t_{\mathrm{rel}} = O(n^2 \log n).$$

Background I

Let $G = (V, E)$ be a connected graph and $\mathcal{X} := \{-1, +1\}^V$. Recall that the (ferromagnetic) Ising model on V with *inverse temperature* β is the probability distribution over *spin configurations* $\sigma \in \mathcal{X}$ given by

$$\mu_\beta(\sigma) := \frac{1}{\mathcal{Z}(\beta)} e^{-\beta \mathcal{H}(\sigma)},$$

where

$$\mathcal{H}(\sigma) := -\sum_{i \sim j} \sigma_i \sigma_j,$$

is the *Hamiltonian* and $\mathcal{Z}(\beta) := \sum_{\sigma \in \mathcal{X}} e^{-\beta \mathcal{H}(\sigma)}$.

Background II

The single-site Glauber dynamics of the Ising model is the Markov chain on $\mathcal{X}$ which, at each time, selects a site $i \in V$ uniformly at random and updates the spin σ_i according to $\mu_\beta(\sigma)$ conditioned on agreeing with σ at all sites in $V \backslash \{i\}$. Specifically, for $\gamma \in \{-1, +1\}$, $i \in V$, and $\sigma \in \mathcal{X}$, let $\sigma^{i,\gamma}$ be the configuration σ with the state at i being set to γ. The transition matrix is

$$Q_\beta(\sigma, \sigma^{i,\gamma}) := \frac{1}{n} \cdot \frac{e^{\gamma\beta S_i(\sigma)}}{e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}} = \frac{1}{n}\left\{\frac{1}{2} + \frac{1}{2}\tanh(\gamma\beta S_i(\sigma))\right\},$$

where

$$S_i(\sigma) := \sum_{j \sim i} \sigma_j.$$

All other transitions have probability 0.

Curie-Weiss model I

Let $G = K_n$ be the complete graph on n vertices. In this case, the Ising model is often referred to as the *Curie-Weiss model*. It is customary to scale β with n. We define $\alpha := \beta(n-1)$.

Theorem (Curie-Weiss model: slow mixing at low temperature)

For $\alpha > 1$, $t_{\mathrm{mix}}(\varepsilon) = \Omega(\exp(r(\alpha)n))$ for some function $r(\alpha) > 0$ not depending on n.

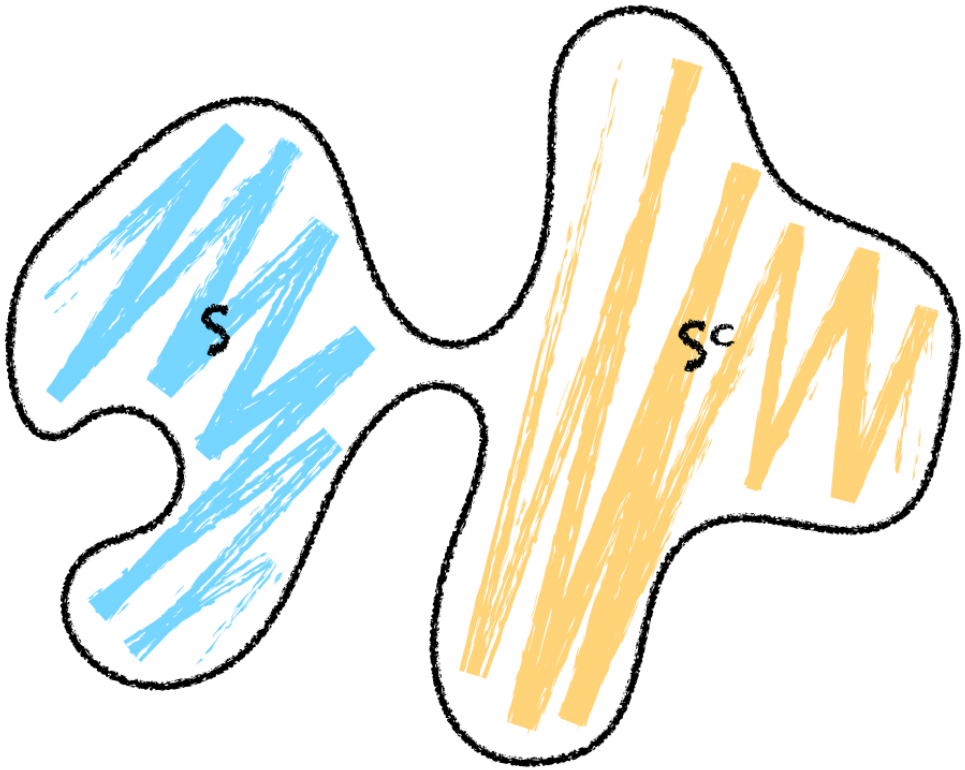
Curie-Weiss model II

Proof: We only prove exponential mixing when α is large enough. The idea of the proof is to bound the bottleneck ratio. To simplify the proof, assume n is odd. We denote the bottleneck ratio of the chain by $\Phi_*^{\mathcal{X}}$ to avoid confusion with the base graph G. Intuitively, because the spins tend to align strongly at low temperature, it takes a considerable amount of time to travel from a configuration with a majority of -1s to a configuration with a majority of $+1$s. A natural place to look for a bottleneck is the set

$$S := \left\{ \sigma \in \mathcal{X} \, : \, \sum_i \sigma_i < 0 \right\},$$

where the quantity $m(\sigma) := \sum_i \sigma_i$ is the *magnetization*.

A bottleneck



Curie-Weiss model III

Proof (continued): Note that the magnetization is positive if and only if a majority of spins are $+1$. Observe further that $\mu_\beta(S) = 1/2$ by symmetry. The bottleneck ratio is hence bounded by

$$\Phi_*^{\mathcal{X}} \leq \frac{\sum_{\sigma \in S, \sigma' \notin S} \mu_\beta(\sigma) Q_\beta(\sigma, \sigma')}{\mu_\beta(S)} = 2 \sum_{\sigma \in S, \sigma' \notin S} \mu_\beta(\sigma) Q_\beta(\sigma, \sigma').$$

Because the Glauber dynamics changes a single spin at a time, in order for $\sigma \in S$ to be adjacent to a configuration $\sigma' \notin S$, it must be that

$$\sigma \in S_{-1} := \{\sigma \in \mathcal{X} \,:\, m(\sigma) = -1\},$$

and that $\sigma' = \sigma^{i,+}$ for some site i such that

$$i \in M_\sigma := \{i \in V \,:\, \sigma_i = -1\}.$$

Curie-Weiss model IV

Proof (continued): Because the number of such sites is $(n+1)/2$ on S_{-1}, that is, $|M_\sigma| = (n+1)/2$ for all $\sigma \in S_{-1}$, and the Glauber dynamics picks a site uniformly at random, it follows that for $\sigma \in S_{-1}$

$$\sum_{\sigma' \notin S} Q_\beta(\sigma, \sigma') \leq \frac{(n+1)/2}{n} = \frac{1}{2}\left(1 + \frac{1}{n}\right).$$

Thus plugging this back

$$\begin{aligned}
\Phi_*^{\mathcal{X}} &\leq 2 \sum_{\sigma \in S, \sigma' \notin S} \mu_\beta(\sigma) Q_\beta(\sigma, \sigma') \\
&\leq \left(1 + \frac{1}{n}\right) \mu_\beta(S_{-1}) = (1 + o(1)) \sum_{\sigma \in S_{-1}} \frac{e^{-\beta \mathcal{H}(\sigma)}}{\mathcal{Z}(\beta)} \\
&= (1 + o(1)) \sum_{\sigma \in S_{-1}} \frac{\exp\left(\frac{\alpha}{n-1}\left[\binom{|M_\sigma|}{2} + \binom{|M_\sigma^c|}{2} - |M_\sigma||M_\sigma^c|\right]\right)}{\mathcal{Z}(\beta)}.
\end{aligned}$$

Curie-Weiss model V

Proof (continued): We bound $\mathcal{Z}(\beta) = \sum_{\sigma\in\mathcal{X}} e^{-\beta\mathcal{H}(\sigma)}$ with the all-(-1) term

$$\begin{aligned}
\Phi_*^{\mathcal{X}} &\le (1+o(1)) \sum_{\sigma\in S_{-1}} \frac{\exp\left(\frac{\alpha}{n-1}\left[\binom{|M_\sigma|}{2} + \binom{|M_\sigma^c|}{2} - |M_\sigma||M_\sigma^c|\right]\right)}{\exp\left(\frac{\alpha}{n-1}\left[\binom{|M_\sigma|}{2} + \binom{|M_\sigma^c|}{2} + |M_\sigma||M_\sigma^c|\right]\right)} \\
&= (1+o(1)) \sum_{\sigma\in S_{-1}} \exp\left(-\frac{2\alpha}{n-1}|M_\sigma||M_\sigma^c|\right) \\
&= (1+o(1)) \binom{n}{n/2} \exp\left(-\frac{2\alpha}{n-1}\left[\frac{n+1}{2}\right]\left[\frac{n-1}{2}\right]\right) \\
&= (1+o(1)) \sqrt{\frac{2}{\pi n}}\, 2^n (1+o(1)) \exp\left(-\frac{\alpha(n+1)}{2}\right) \\
&= C_\alpha \sqrt{\frac{2}{\pi n}} \exp\left(-n\left[\frac{\alpha}{2} - \ln 2\right]\right),
\end{aligned}$$

for some constant $C_\alpha > 0$ depending on α.

Curie-Weiss model VI

Proof (continued): Hence, by Cheeger's inequality, for $\alpha > 2\ln 2$ there is $r(\alpha) > 0$

$$t_{\mathrm{mix}}(\varepsilon) \geq (t_{\mathrm{rel}} - 1) \log\left(\frac{1}{2\varepsilon}\right) \geq \exp(r(\alpha)n) \log\left(\frac{1}{2\varepsilon}\right).$$

■

Go deeper

More details and examples on spectral techniques at:

`http://www.math.wisc.edu/~roch/mdp/`

For more on mixing times in general, see e.g. (available online):

- *Markov Chains and Mixing Times* by Levin, Peres and Wilmer
- *Reversible Markov Chains and Random Walks on Graphs* by Aldous and Fill

Probability on Graphs: Techniques and Applications to Data Science

4 - Coupling

Sébastien Roch
UW–Madison
Mathematics

July 26, 2018

Coupling

Definition

Let μ and ν be probability measures on the same measurable space $(S, \mathcal{S})$. A *coupling* of μ and ν is a probability measure γ on the product space $(S \times S, \mathcal{S} \times \mathcal{S})$ such that the *marginals* of γ coincide with μ and ν, i.e.,

$$\gamma(A \times S) = \mu(A) \quad \text{and} \quad \gamma(S \times A) = \nu(A), \qquad \forall A \in \mathcal{S}.$$

Examples

Example (Bernoulli variables)

Let X and Y be Bernoulli random variables with parameters $0 \leq q < r \leq 1$ respectively. That is, $\mathbb{P}[X = 0] = 1 - q$ and $\mathbb{P}[X = 1] = q$, and similarly for Y. Here $S = \{0, 1\}$ and $\mathcal{S} = 2^S$.

- *(Independent coupling)* One coupling of X and Y is (X', Y') where $X' \overset{\mathrm{d}}{=} X$ and $Y' \overset{\mathrm{d}}{=} Y$ are *independent*. Its law is

$$\Big(\mathbb{P}[(X', Y') = (i, j)]\Big)_{i,j \in \{0,1\}} = \begin{pmatrix} (1-q)(1-r) & (1-q)r \\ q(1-r) & qr \end{pmatrix}.$$

- *(Monotone coupling)* Another possibility is to pick U uniformly at random in $[0, 1]$, and set $X'' = \mathbf{1}_{\{U \leq q\}}$ and $Y'' = \mathbf{1}_{\{U \leq r\}}$. The law of coupling (X'', Y'') is

$$\Big(\mathbb{P}[(X'', Y'') = (i, j)]\Big)_{i,j \in \{0,1\}} = \begin{pmatrix} 1-r & r-q \\ 0 & q \end{pmatrix}.$$

Coupling inequality I

Let μ and ν be probability measures on $(S, \mathcal{S})$. Recall the definition of total variation distance:

$$\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A \in \mathcal{S}} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in S} |\mu(x) - \nu(x)|.$$

Lemma

Let μ and ν be probability measures on $(S, \mathcal{S})$. For any coupling (X, Y) of μ and ν,

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \mathbb{P}[X \neq Y].$$

Coupling inequality II

Proof:

$$\begin{aligned}\mu(A) - \nu(A) &= \mathbb{P}[X \in A] - \mathbb{P}[Y \in A] \\ &= \mathbb{P}[X \in A,\ X = Y] + \mathbb{P}[X \in A,\ X \neq Y] \\ &\quad - \mathbb{P}[Y \in A,\ X = Y] - \mathbb{P}[Y \in A,\ X \neq Y] \\ &= \mathbb{P}[X \in A,\ X \neq Y] - \mathbb{P}[Y \in A,\ X \neq Y] \\ &\leq \mathbb{P}[X \neq Y],\end{aligned}$$

and, similarly, $\nu(A) - \mu(A) \leq \mathbb{P}[X \neq Y]$. Hence

$$|\mu(A) - \nu(A)| \leq \mathbb{P}[X \neq Y].$$

■

Maximal coupling

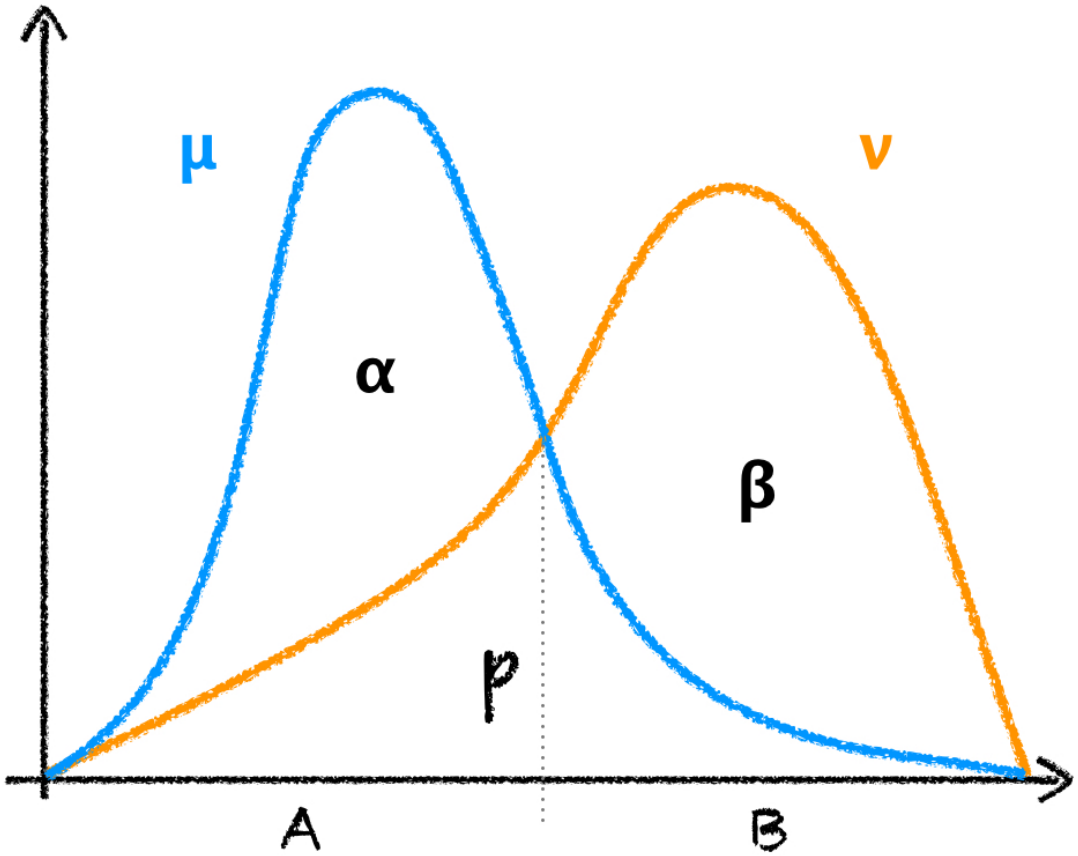
In fact, the inequality is tight.

Lemma

Assume S is finite and let $\mathcal{S} = 2^S$. Let μ and ν be probability measures on $(S, \mathcal{S})$. Then,

$$\|\mu - \nu\|_{\mathrm{TV}} = \inf\{\mathbb{P}[X \neq Y] : \textit{coupling } (X, Y) \textit{ of } \mu \textit{ and } \nu\}.$$

Maximal coupling by picture



Example: Bernoullis

Example (Bernoulli variables, continued)

Let X and Y be Bernoulli random variables with parameters $0 \leq q < r \leq 1$ respectively.Let μ and ν be the laws of X and Y respectively. To construct the maximal coupling as above, we note that

$$p := \sum_x \mu(x) \wedge \nu(x) = (1-r) + q, \qquad 1 - p = \alpha = \beta := r - q,$$

$$A := \{0\}, \qquad B := \{1\},$$

$$(\gamma_{\min}(x))_{x=0,1} = \left(\frac{1-r}{(1-r)+q}, \frac{q}{(1-r)+q} \right), \qquad \gamma_A(0) := 1, \qquad \gamma_B(1) := 1.$$

The law of the maximal coupling (X''', Y''') is

$$\Big(\mathbb{P}[(X''', Y''') = (i,j)] \Big)_{i,j \in \{0,1\}} = \begin{pmatrix} 1-r & r-q \\ 0 & q \end{pmatrix},$$

which coincides with the monotone coupling.

Bounding the mixing time via coupling I

Let P be an irreducible, aperiodic transition matrix on V with stationary distribution π. Recall that, for a fixed $0 < \varepsilon < 1/2$, the mixing time of P is

$$t_{\text{mix}}(\varepsilon) := \min\{t : d(t) \leq \varepsilon\},$$

where

$$d(t) := \max_{x \in V} \|P^t(x, \cdot) - \pi\|_{\text{TV}}.$$

It will be easier to work with

$$\bar{d}(t) := \max_{x,y \in V} \|P^t(x, \cdot) - P^t(y, \cdot)\|_{\text{TV}},$$

which satisfies $d(t) \leq \bar{d}(t) \leq 2d(t)$.

Bounding the mixing time via coupling II

Definition (Markovian coupling)

A *Markovian coupling* of P is a Markov chain $(X_t, Y_t)_t$ on $V \times V$ with transition matrix Q satisfying:

- For all $x, y, x', y' \in V$,

$$\sum_{z'} Q((x,y),(x',z')) = P(x,x'),$$

$$\sum_{z'} Q((x,y),(z',y')) = P(y,y').$$

We say that a Markovian coupling is *coalescing* if further:

- For all $z \in V$, $x' \neq y' \implies Q((z,z),(x',y')) = 0$.

Bounding the mixing time via coupling III

Let (X_t, Y_t) be a coalescing Markovian coupling of P. By the coalescing condition, if $X_s = Y_s$ then $X_t = Y_t$ for all $t \geq s$. That is, once (X_t) and (Y_t) meet, they remain equal. Let τ_{coal} be the *coalescence time* (also called coupling time), i.e.,

$$\tau_{\text{coal}} := \inf\{t \geq 0 \,:\, X_t = Y_t\}.$$

The key to the coupling approach to mixing times is the following immediate consequence of the coupling inequality. For any starting point (x, y),

$$\|P^t(x,\cdot) - P^t(y,\cdot)\|_{\text{TV}} \leq \mathbb{P}_{(x,y)}[X_t \neq Y_t] = \mathbb{P}_{(x,y)}[\tau_{\text{coal}} > t].$$

Bounding the mixing time via coupling IV

Theorem (Bounding the mixing time: coupling method)

Let (X_t, Y_t) be a coalescing Markovian coupling of an irreducible transition matrix P on a finite state space V with stationary distribution π. Then

$$d(t) \leq \max_{x,y \in V} \mathbb{P}_{(x,y)}[\tau_{\text{coal}} > t].$$

In particular

$$t_{\text{mix}}(\varepsilon) \leq \inf\left\{t \geq 0 \,:\, \mathbb{P}_{(x,y)}[\tau_{\text{coal}} > t] \leq \varepsilon,\ \forall x, y\right\}.$$

Example: Hypercube I

Let (Z_t) be lazy random walk on the n-dimensional hypercube $\mathbb{Z}_2^n := \{0,1\}^n$ where $i \sim j$ if $\|i - j\|_1 = 1$. We denote the coordinates of Z_t by $(Z_t^{(1)}, \ldots, Z_t^{(n)})$. The coupling (X_t, Y_t) started at (x, y) is the following:

- At each time t, pick a coordinate i uniformly at random in $[n]$, pick a bit value b in $\{0,1\}$ uniformly at random independently of the coordinate choice.
- Set *both* i coordinates to b, i.e., $X_t^{(i)} = Y_t^{(i)} = b$.

Clearly the chains coalesce when all coordinates have been updated at least once.

Example: Hypercube II

Lemma (Coupon collecting)

Let τ_{coll} be the time it takes to update each coordinate at least once. Then, for any $c > 0$,

$$\mathbb{P}\left[\tau_{\text{coll}} > \lceil n \log n + cn \rceil\right] \leq e^{-c}.$$

Proof: Let B_i be the event that the i-th coordinate has not been updated by time $\lceil n \log n + cn \rceil$. Then

$$\begin{aligned} \mathbb{P}[\tau_{\text{coll}} > \lceil n \log n + cn \rceil] &\leq \sum_i \mathbb{P}[B_i] = \sum_i \left(1 - \frac{1}{n}\right)^{\lceil n \log n + cn \rceil} \\ &\leq n \exp\left(-\frac{n \log n + cn}{n}\right) = e^{-c}. \end{aligned}$$

Example: Hypercube III

Applying the coupling bound, we get

$$\begin{aligned} d(\lceil n\log n + cn\rceil) &\leq \max_{x,y\in V} \mathbb{P}_{(x,y)}[\tau_{\mathrm{coal}} > \lceil n\log n + cn\rceil] \\ &\leq \mathbb{P}[\tau_{\mathrm{coll}} > \lceil n\log n + cn\rceil] \\ &\leq e^{-c}. \end{aligned}$$

Hence for $c := c_\varepsilon > 0$ large enough:

$$\mathrm{t}_{\mathrm{mix}}(\varepsilon) \leq \lceil n\log n + c_\varepsilon n\rceil.$$

Path coupling method I

Path coupling is a method for constructing Markovian couplings from "simpler" couplings. The building blocks are one-step couplings starting from pairs of initial states that are close in some dissimilarity graph. Let (X_t) be an irreducible Markov chain on a finite state space V with transition matrix P and stationary distribution π. Assume that we are given a *dissimilarity graph* $H_0 = (V_0, E_0)$ on $V_0 := V$ with edge weights $w_0 : E_0 \to \mathbb{R}_+$. This graph need not have the same edges as the transition graph of (X_t). We extend w_0 to the *path metric*

$$w_0(x, y) := \inf \left\{ \sum_{i=0}^{m-1} w_0(x_i, x_{i+1}) : x = x_0, \ldots, x_m = y \text{ is a path in } H_0 \right\},$$

where the infimum is over all paths connecting x and y in H_0. We call a path achieving the infimum a *minimum-weight path*. Let

$$\Delta_0 := \max_{x,y} w_0(x, y),$$

be the *weighted diameter* of H_0.

Path coupling method II

Theorem (Path coupling method)

Assume that $w_0(u, v) \geq 1$, *for all* $\{u, v\} \in E_0$. *Assume further that there exists* $\kappa \in (0, 1)$ *such that:*

- *For all* x, y *with* $\{x, y\} \in E_0$, *there is a coupling* (X^*, Y^*) *of* $P(x, \cdot)$ *and* $P(y, \cdot)$ *satisfying the* contraction property

$$\mathbb{E}[w_0(X^*, Y^*)] \leq \kappa \, w_0(x, y).$$

Then

$$d(t) \leq \Delta_0 \, \kappa^t,$$

or

$$t_{\mathrm{mix}}(\varepsilon) \leq \left\lceil \frac{\log \Delta_0 + \log \varepsilon^{-1}}{\log \kappa^{-1}} \right\rceil.$$

Path coupling method III

Proof: The crux of the proof is to extend the contraction property to arbitrary pairs of vertices.

Lemma (Global coupling)

For all $x, y \in V$, there is a coupling (X^*, Y^*) of $P(x, \cdot)$ and $P(y, \cdot)$ such that the contraction property holds.

Iterating the coupling in this claim immediately implies the existence of a coalescing Markovian coupling (X_t, Y_t) of P such that

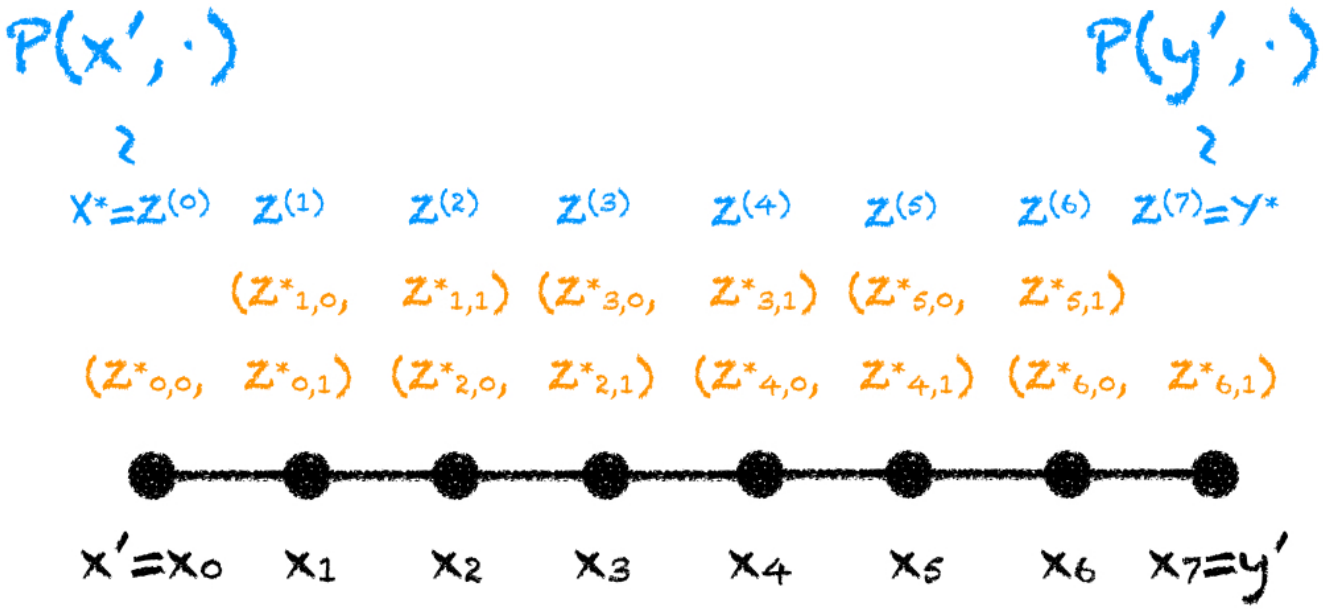
$$\begin{aligned}\mathbb{E}_{(x,y)}[w_0(X_t, Y_t)] &= \mathbb{E}_{(x,y)}\left[\mathbb{E}[w_0(X_t, Y_t) \,|\, X_{t-1}, Y_{t-1}]\right] \\ &\le \mathbb{E}_{(x,y)}\left[\kappa\, w_0(X_{t-1}, Y_{t-1})\right] \le \cdots \le \kappa^t\, \mathbb{E}_{(x,y)}[w_0(X_0, Y_0)] \\ &= \kappa^t\, w_0(x, y) \le \kappa^t\, \Delta_0.\end{aligned}$$

By assumption, $\mathbf{1}_{\{x \neq y\}} \le w_0(x, y)$ so that by the coupling inequality

$$d(t) \le \bar{d}(t) \le \max_{x,y} \mathbb{P}_{(x,y)}[X_t \neq Y_t] \le \max_{x,y} \mathbb{E}_{(x,y)}[w_0(X_t, Y_t)] \le \kappa^t\, \Delta_0,$$

which implies the theorem.

Global coupling lemma: proof by picture



Setup I

Let $G = (V, E)$ be a finite, connected graph with maximum degree $\bar{\delta}$. Define $\mathcal{X} := \{-1, +1\}^V$. Recall that the (ferromagnetic) Ising model on V with *inverse temperature* β is the probability distribution over *spin configurations* $\sigma \in \mathcal{X}$ given by

$$\mu_\beta(\sigma) := \frac{1}{\mathcal{Z}(\beta)} e^{-\beta \mathcal{H}(\sigma)},$$

where

$$\mathcal{H}(\sigma) := -\sum_{i \sim j} \sigma_i \sigma_j,$$

is the *Hamiltonian* and $\mathcal{Z}(\beta) := \sum_{\sigma \in \mathcal{X}} e^{-\beta \mathcal{H}(\sigma)}$.

Setup II

The single-site Glauber dynamics of the Ising model is the Markov chain on $\mathcal{X}$ which, at each time, selects a site $i \in V$ uniformly at random and updates the spin σ_i according to $\mu_\beta(\sigma)$ conditioned on agreeing with σ at all sites in $V \backslash \{i\}$. Specifically, for $\gamma \in \{-1, +1\}$, $i \in V$, and $\sigma \in \mathcal{X}$, let $\sigma^{i,\gamma}$ be the configuration σ with the state at i being set to γ. Then the transition matrix is

$$Q_\beta(\sigma, \sigma^{i,\gamma}) := \frac{1}{n} \cdot \frac{e^{\gamma\beta S_i(\sigma)}}{e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}} = \frac{1}{n}\left\{\frac{1}{2} + \frac{1}{2}\tanh(\gamma\beta S_i(\sigma))\right\},$$

where

$$S_i(\sigma) := \sum_{j \sim i} \sigma_j.$$

All other transitions have probability 0. Recall that this chain is irreducible and reversible with respect to μ_β.

Fast mixing at high temperature I

We show that the Glauber dynamics of the Ising model is fast mixing when the inverse temperature β is small enough as a function of the maximum degree.

Theorem (Glauber dynamics: fast mixing at high temperature)

If $\beta < \bar{\delta}^{-1}$ *then* $\implies$ $\mathrm{t}_{\mathrm{mix}}(\varepsilon) = O(n \log n)$.

Proof: We use path coupling. Let $H_0 = (V_0, E_0)$ where $V_0 := \mathcal{X}$ and $\{\sigma, \omega\} \in E_0$ if $\frac{1}{2}\|\sigma - \omega\|_1 = 1$ with unit w_0-weights on all edges. (To avoid confusion, we reserve the notation $\sim$ for adjacency in G.) Let $\{\sigma, \omega\} \in E_0$ differ at coordinate i.

Fast mixing at high temperature II

Proof (continued): We construct a coupling (X^*, Y^*) of $Q_\beta(\sigma, \cdot)$ and $Q_\beta(\omega, \cdot)$. We first pick the same coordinate i_* to update. If i_* is such that all its neighbors in G have the same state in σ and ω, i.e., if $\sigma_j = \omega_j$ for all $j \sim i_*$, we update X^* from σ according to the Glauber rule and set $Y^* := X^*$. Note that this includes the case $i_* = i$. Otherwise, i.e. if $i_* \sim i$, we proceed as follows. From the state σ, the probability of updating site i_* to state $\gamma \in \{-1, +1\}$ is given by $\frac{1}{2} + \frac{1}{2}\tanh(\gamma\beta S_i(\sigma))$, and similarly for ω. Unlike the previous case, we cannot guarantee that the update is identical in both chains. To minimize the chance of increasing the distance between the two chains, we perform a maximal coupling of Bernoullis: we pick a uniform-$[-1, 1]$ variable U and set

$$X^*_{i_*} := \begin{cases} +1, & \text{if } U \le \tanh(\beta S_{i_*}(\sigma)) \\ -1, & \text{o.w.} \end{cases}$$

and

$$Y^*_{i_*} := \begin{cases} +1, & \text{if } U \le \tanh(\beta S_{i_*}(\omega)) \\ -1, & \text{o.w.} \end{cases}$$

Fast mixing at high temperature III

Proof (continued): We set $X_j^* := \sigma_j$ and $Y_j^* := \omega_j$ for all $j \neq i^*$. The expected distance between X^* and Y^* is then

$$\mathbb{E}[w_0(X^*, Y^*)] = 1 - \underbrace{\frac{1}{n}}_{\text{(a)}} + \underbrace{\frac{1}{n} \sum_{j \sim i} \frac{1}{2} |\tanh(\beta S_j(\sigma)) - \tanh(\beta S_j(\omega))|}_{\text{(b)}},$$

where (a) corresponds to $i_* = i$ in which case $w_0(X^*, Y^*) = 0$ and (b) corresponds to $i_* \sim i$ in which case $w_0(X^*, Y^*) = 2$ with probability $\frac{1}{2} |\tanh(\beta S_{i_*}(\sigma)) - \tanh(\beta S_{i_*}(\omega))|$ by our coupling, and $w_0(X^*, Y^*) = 1$ otherwise. To bound (b), we note that for $j \sim i$

$$|\tanh(\beta S_j(\sigma)) - \tanh(\beta S_j(\omega))| = \tanh(\beta(s+2)) - \tanh(\beta s),$$

where $s := S_j(\sigma) \wedge S_j(\omega)$. The derivative of tanh is maximized at 0 where it is equal to 1. So the r.h.s. above is $\leq 2\beta$.

Fast mixing at high temperature IV

Proof (continued): Plugging this back above we get

$$\mathbb{E}[w_0(X^*, Y^*)] \leq 1 - \frac{1 - \bar{\delta}\beta}{n} \leq \exp\left(-\frac{1 - \bar{\delta}\beta}{n}\right) = \kappa\, w_0(\sigma, \omega),$$

where

$$\kappa := \exp\left(-\frac{1 - \bar{\delta}\beta}{n}\right) < 1,$$

by assumption. The diameter of H_0 is $\Delta_0 = n$. By the path coupling theorem,

$$t_{\mathrm{mix}}(\varepsilon) \leq \left\lceil \frac{\log \Delta_0 + \log \varepsilon^{-1}}{\log \kappa^{-1}} \right\rceil = \left\lceil \frac{n(\log n + \log \varepsilon^{-1})}{1 - \bar{\delta}\beta} \right\rceil,$$

which implies the claim. ∎

Go deeper

More details and examples on coupling at:

`http://www.math.wisc.edu/~roch/mdp/`

For more on mixing times in general, see e.g. (available online):

- *Markov Chains and Mixing Times* by Levin, Peres and Wilmer
- *Reversible Markov Chains and Random Walks on Graphs* by Aldous and Fill