# Probability on Graphs: Techniques and Applications to Data Science

## *0 - Preliminaries*

Sébastien Roch

*UW–Madison*

*Mathematics*

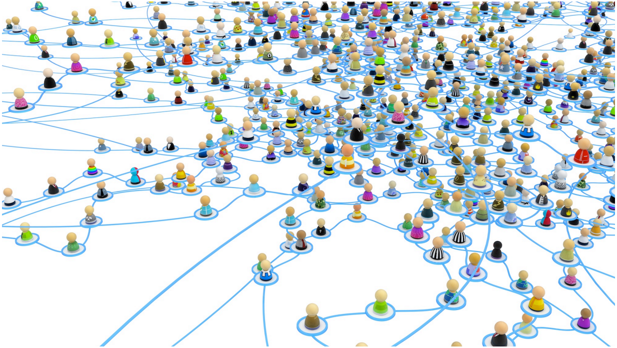
July 25, 2018

# Go deeper

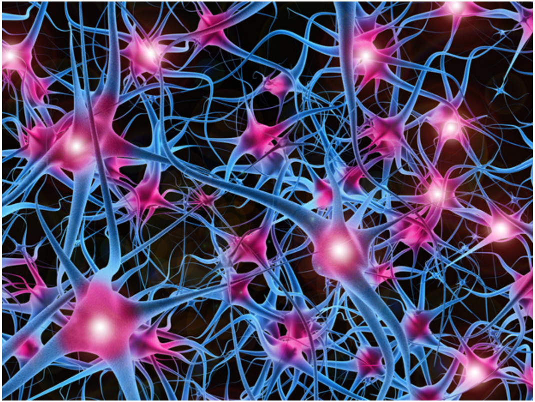
A lot more details and examples in the lecture notes at:

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

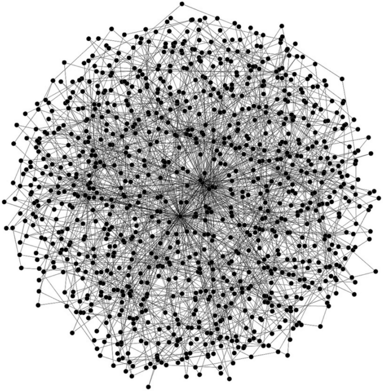
# Networks are ubiquitous: Social networks



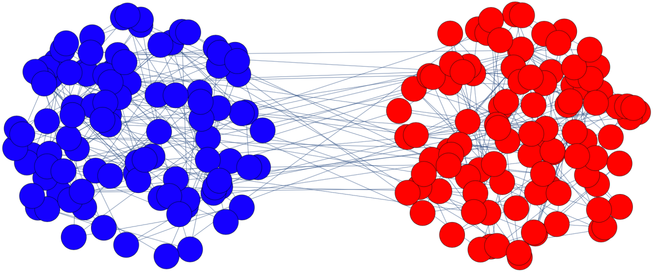
# Networks are ubiquitous: Biological networks



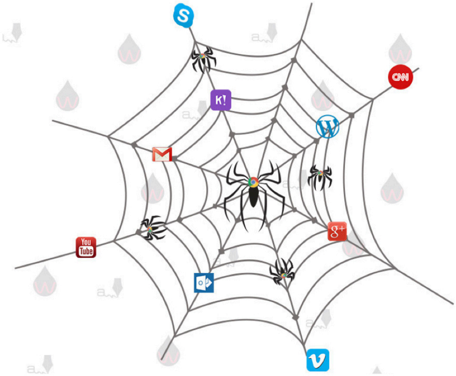
# Data science: Network modeling



# Data science: Network processes



# Data science: Network sampling



1. Graph terminology

2. Basic examples of stochastic processes on graphs

# Graphs

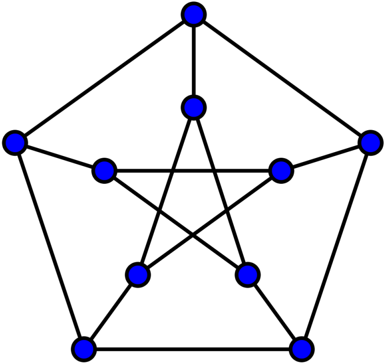
## Definition

An *(undirected) graph* is a pair $G = (V, E)$ where $V$ is the set of *vertices* and

$$E \subseteq \{\{u, v\} \,:\, u, v \in V\},$$

is the set of *edges*.

# An example: the Petersen graph



# Basic definitions

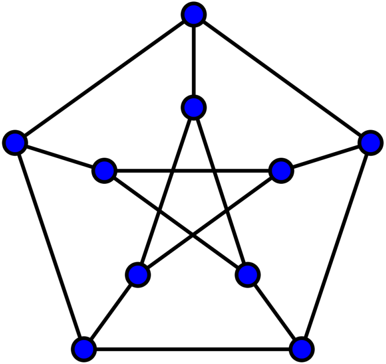
## Definition (Neighborhood)

Two vertices $u, v \in V$ are *adjacent*, denoted by $u \sim v$, if $\{u, v\} \in E$. The set of adjacent vertices of $v$, denoted by $N(v)$, is called the *neighborhood* of $v$ and its size, i.e. $\delta(v) := |N(v)|$, is the *degree* of $v$. A vertex $v$ with $\delta(v) = 0$ is called *isolated*.

## Example

All vertices in the Petersen graph have degree 3. In particular there is no isolated vertex.

# An example: the Petersen graph



# Paths and connectivity

## Definition (Paths)

A *path* in $G$ is a sequence of vertices $x_0 \sim x_1 \sim \cdots \sim x_k$. The number of edges, $k$, is called the *length* of the path. If $x_0 = x_k$, we call it a *cycle*. We write $u \leftrightarrow v$ if there is a path between $u$ and $v$. The equivalence classes of $\leftrightarrow$ are called *connected components*. The length of the shortest path between two vertices $u, v$ is their *graph distance*, denoted $d_G(u, v)$.

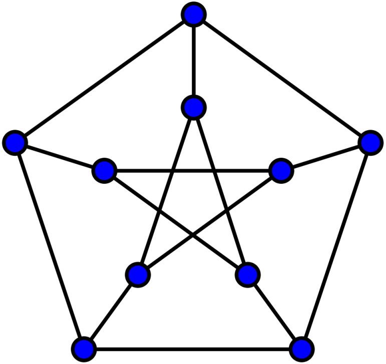
## Definition (Connectivity)

A graph is *connected* if any two vertices are linked by a path, i.e., if $u \leftrightarrow v$ for all $u, v \in V$.

## Example

The Petersen graph is connected.

# An example: the Petersen graph



# Adjacency matrix

## Definition

Let $G = (V, E)$ be a graph with $n = |V|$. The *adjacency matrix* $A$ of $G$ is the $n \times n$ matrix defined as $A_{xy} = 1$ if $\{x, y\} \in E$ and 0 otherwise.

## Example

The adjacency matrix of a *triangle* (i.e. 3 vertices with all edges) is

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

# Examples of finite graphs

- $K_n$: clique with $n$ vertices, i.e., graph with all edges present
- $C_n$: cycle with $n$ non-repeated vertices
- $\mathbb{H}^n$: $n$-dimensional hypercube, i.e., $V = \{0,1\}^n$ and $u \sim v$ if $u$ and $v$ differ at one coordinate

1. Graph terminology

2. Basic examples of stochastic processes on graphs

# Erdös-Rényi random graph

## Definition

Let $V = [n]$ and $p \in [0,1]$. The *Erdös-Rényi graph* $G = (V,E)$ on $n$ vertices with density $p$ is defined as follows: for each pair $x \neq y$ in $V$, the edge $\{x,y\}$ is in $E$ with probability $p$ independently of all other edges. We write $G \sim \mathbb{G}_{n,p}$ and we denote the corresponding measure by $\mathbb{P}_{n,p}$.

Questions:

- What is the probability of observing a triangle?
- Is $G$ connected?
- What is the typical chromatic number (i.e., the smallest number of colors needed to color the vertices so that no two adjacent vertices share the same color)?

# Other random graph models

- Preferential attachement
- Small world
- Fixed degree distribution

# 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.

Questions:

- How often does the walk return to its starting point?
- How long does it take to visit all vertices once or a particular subset of vertices for the first time?
- How fast does it approach stationarity?

# Other sampling schemes

- Random walks with restarts
- Branching random walks
- Random sample of vertices and their neighbors

# 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 $\mu$ 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.

# Undirected graphical models II

## Example

For $\beta > 0$, the *ferromagnetic Ising model* with inverse temperature $\beta$ 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*.

Questions:

- How fast is correlation decaying?
- How to sample efficiently?
- How to reconstruct the graph from samples?

# Other graphical models

- Gaussian graphical models
- Bayes nets
- Latent graphical models

# Go deeper

More details and examples on basic models at:

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

For more on probability on graphs in general, see e.g. (available online):

- *Probability on Graphs* by Grimmett
- *Probability on Trees and Networks* by Lyons with Peres

# Probability on Graphs: Techniques and Applications to Data Science

## *1 - First and second moment methods*

Sébastien Roch
*UW–Madison*
*Mathematics*

July 25, 2018

1 Markov's inequality

2 First and second moment methods

3 Illustration: Erdös-Rényi connectivity threshold

# Markov's inequality

## Theorem (Markov's inequality)

*Let $X$ be a non-negative random variable. Then, for all $b > 0$,*

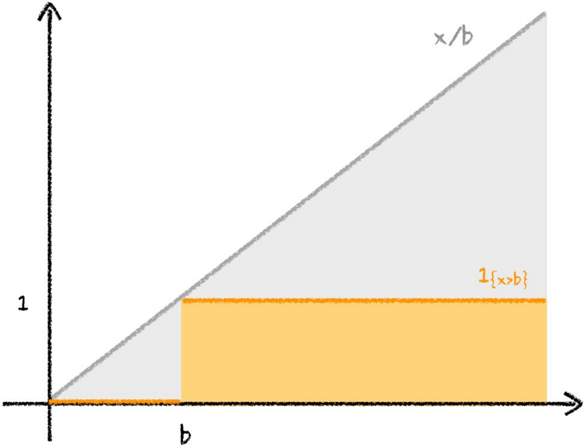
$$\mathbb{P}[X \geq b] \leq \frac{\mathbb{E}X}{b}.$$

*Proof:*

$$\mathbb{E}X \geq \mathbb{E}[X; X \geq b] \geq \mathbb{E}[b; X \geq b] = b\,\mathbb{P}[X \geq b].$$

■

# Markov's inequality: Proof by picture



# Chebyshev's inequality

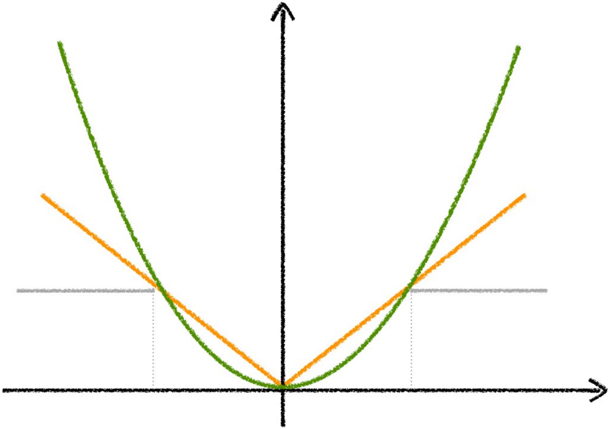
## Theorem (Chebyshev's inequality)

*Let $X$ be a random variable with $\mathbb{E}X^2 < +\infty$. Then, for all $\beta > 0$,*

$$\mathbb{P}[|X - \mathbb{E}X| > \beta] \leq \frac{\mathrm{Var}[X]}{\beta^2}.$$

*Proof:* This follows immediately by applying Markov's inequality to $|X - \mathbb{E}X|^2$ with $b = \beta^2$. ■

# Chebyshev's inequality: Proof by picture



1. Markov's inequality
2. First and second moment methods
3. Illustration: Erdös-Rényi connectivity threshold

# First moment method

## Theorem (First moment method)

*If $X$ is a non-negative, integer-valued random variable, then*

$$\mathbb{P}[X > 0] \leq \mathbb{E}X.$$

*Proof:* Take $b = 1$ in Markov's inequality. ■

That is: if $X$ has "small" expectation, then its value is 0 with "large" probability. Typically used in the following way: one wants to show that a "bad event" does not occur with high probability; the random variable $X$ counts the number of such "bad events." In that case, $X$ is a sum of indicators and the theorem reduces to the *union bound*.

# Going in the other direction

The first moment method gives an *upper bound* on the probability that a non-negative, integer-valued random variable is positive—provided its expectation is small. Suppose we want a *lower bound*. Note that a large expectation does not suffice.

## Example

Say $X_n$ is $n^2$ with probability $1/n$, and 0 otherwise. Then $\mathbb{E}X_n = n \to +\infty$, yet $\mathbb{P}[X_n > 0] \to 0$.

# Second moment method

### Theorem (Second moment method)

*If $X$ is a non-negative, integer-valued random variable, then*

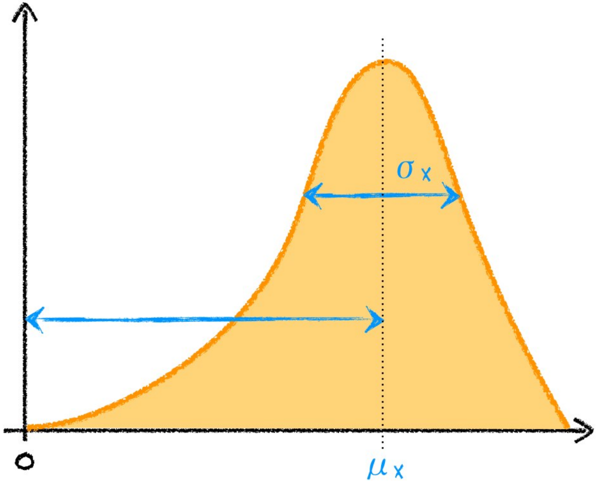
$$\mathbb{P}[X > 0] \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]} \left( = 1 - \frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2 + \mathrm{Var}[X]} \right).$$

*Proof (of weaker version):* By Chebyshev's inequality,

$$\mathbb{P}[X = 0] \leq \mathbb{P}[|X - \mathbb{E}X| \geq \mathbb{E}X] \leq \frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2}.$$

■

# Second moment method: Proof by picture



# First and second moment methods: summary

If $X$ is a non-negative, integer-valued random variable, then

$$\mathbb{P}[X > 0] \leq \mathbb{E}X,$$

and

$$\mathbb{P}[X > 0] \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}.$$

1. Markov's inequality

2. First and second moment methods

3. Illustration: Erdös-Rényi connectivity threshold

# Threshold phenomena

Consider the Erdös-Rényi random graph. A *threshold function* for a graph property $P$ is a function $r(n)$ such that

$$\lim_n \mathbb{P}_{n,p_n}[G_n \text{ has property } P] = \begin{cases} 0, & \text{if } p_n \ll r(n) \\ 1, & \text{if } p_n \gg r(n), \end{cases}$$

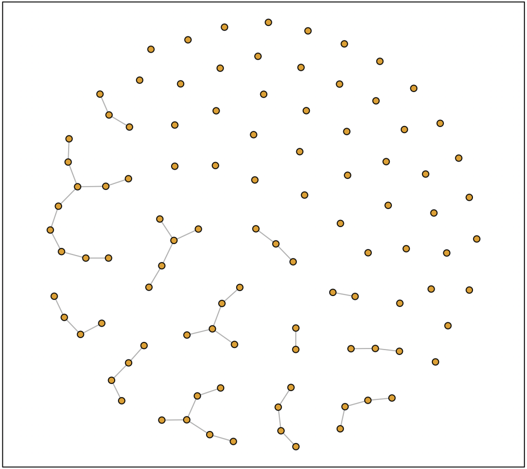
where $G_n \sim \mathbb{G}_{n,p_n}$ is an Erdös-Rényi graph with $n$ vertices and density $p_n$.

# Connectivity via isolated vertices

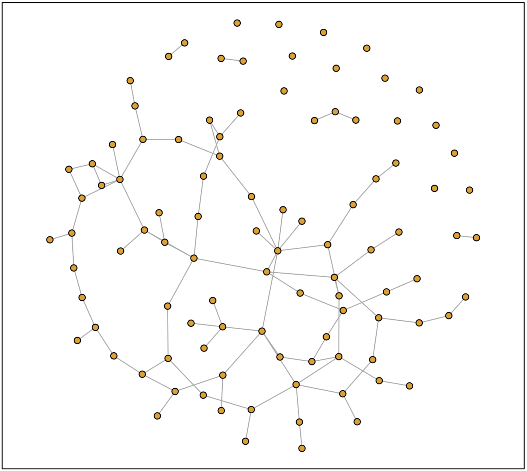
We use the first and second moment methods to show that the threshold function for connectivity in the Erdös-Rényi random graph is $\frac{\log n}{n}$.

We prove this result by deriving the threshold function for the presence of isolated vertices. Of course isolated vertices imply a disconnected graph. What is less obvious: the two thresholds actually *coincide*.

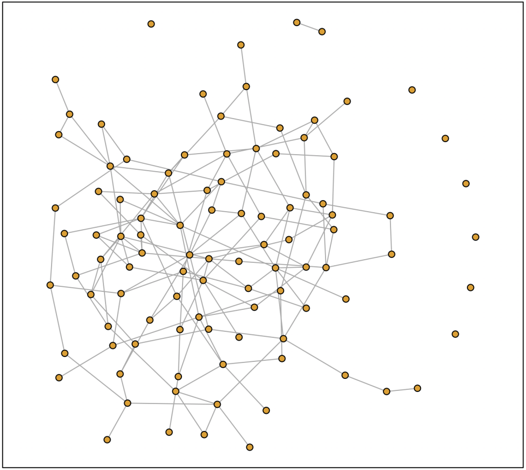
# Erdös-Rényi with $n = 100$ and $p_n = 1/100$



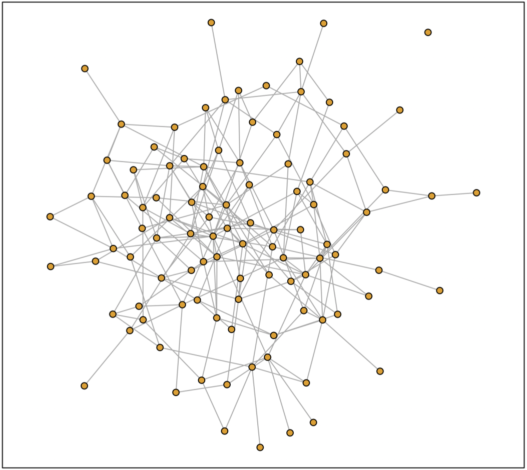
# Erdös-Rényi with $n = 100$ and $p_n = 2/100$



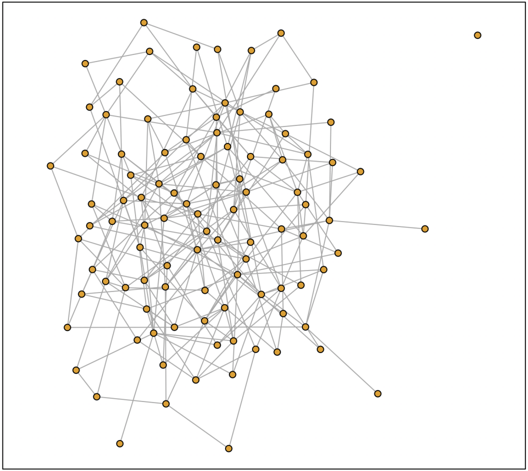
# Erdös-Rényi with $n = 100$ and $p_n = 3/100$



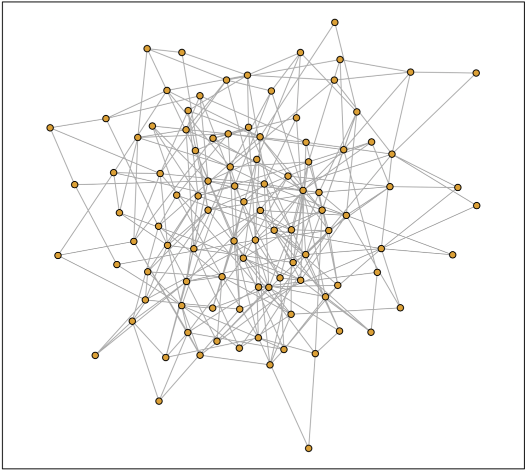
# Erdös-Rényi with $n = 100$ and $p_n = 4/100$



# Erdös-Rényi with $n = 100$ and $p_n = 5/100$



# Erdös-Rényi with $n = 100$ and $p_n = 6/100$



# Threshold for isolated vertices I

## Theorem

*"Not having an isolated vertex" has threshold function* $\frac{\log n}{n}$.

*Proof:* Let $X_n$ be the number of isolated vertices in the Erdös-Rényi graph $G_n \sim \mathbb{G}_{n,p_n}$. Using $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$,

$$\mathbb{E}_{n,p_n}[X_n] = n(1-p_n)^{n-1} \leq e^{\log n - (n-1)p_n} \to 0,$$

when $p_n \gg \frac{\log n}{n}$. So the first moment method gives one direction: $\mathbb{P}_{n,p_n}[X_n > 0] \to 0$ when $p_n \gg \frac{\log n}{n}$.

# Threshold for isolated vertices II

*Proof (continued):* Let $A_j$ be the event that vertex $j$ is isolated and $X_n = \sum_j \mathbf{1}_{A_j}$. By the computation above, using $1 - x \geq e^{-x-x^2}$ for $x \in [0, 1/2]$,

$$\mu_n = \mathbb{E}_{n,p_n}[X_n] = \sum_i \mathbb{P}_{n,p_n}[A_i] = n(1-p_n)^{n-1} \geq e^{\log n - np_n - np_n^2},$$

which goes to $+\infty$ when $p_n \ll \frac{\log n}{n}$.

Note that for all $i \neq j$

$$\mathbb{P}_{n,p_n}[A_i \cap A_j] = (1-p_n)^{2(n-2)+1},$$

so that

$$\gamma_n = \mathbb{E}_{n,p_n}[X_n^2] - \mathbb{E}_{n,p_n}[X_n] = \sum_{i \neq j} \mathbb{P}_{n,p_n}[A_i \cap A_j] = n(n-1)(1-p_n)^{2n-3}.$$

# Threshold for isolated vertices III

*Proof (continued):* We have

$$\begin{aligned}\frac{\mathbb{E}_{n,p_n}[X_n^2]}{(\mathbb{E}_{n,p_n}[X_n])^2} &= \frac{\mu_n + \gamma_n}{\mu_n^2} \\ &\leq \frac{n(1-p_n)^{n-1} + n^2(1-p_n)^{2n-3}}{n^2(1-p_n)^{2n-2}} \\ &\leq \frac{1}{n(1-p_n)^{n-1}} + \frac{1}{1-p_n},\end{aligned}$$

which is $1 + o(1)$ when $p_n \ll \frac{\log n}{n}$. ■

# Threshold for connectivity I

## Theorem

*Connectivity has threshold function $\frac{\log n}{n}$.*

*Proof:* We start with the easy direction. If $p_n \ll \frac{\log n}{n}$, the previous result implies that the graph has isolated vertices, and therefore is disconnected, with probability going to 1 as $n \to +\infty$.

Now assume that $p_n \gg \frac{\log n}{n}$. Let $\mathcal{D}_n$ be the event that $G_n$ is disconnected. To bound $\mathbb{P}_{n,p_n}[\mathcal{D}_n]$, for $k \in \{1, \ldots, n/2\}$ we let $Y_k$ be the number of subsets of $k$ vertices that are disconnected from all other vertices in the graph. Then, by the first moment method,

$$\mathbb{P}_{n,p_n}[\mathcal{D}_n] \leq \mathbb{P}_{n,p_n}\left[\sum_{k=1}^{n/2} Y_k > 0\right] \leq \sum_{k=1}^{n/2} \mathbb{E}_{n,p_n}[Y_k].$$

# Threshold for connectivity II

*Proof (continued):* Using that $k \leq n/2$ and $\binom{n}{k} \leq n^k$,

$$\mathbb{E}_{n,p_n}[Y_k] = \binom{n}{k}(1-p_n)^{k(n-k)} \leq \left(n(1-p_n)^{n/2}\right)^k.$$

The expression in parentheses is $o(1)$ when $p_n \gg \frac{\log n}{n}$. Summing over $k$,

$$\mathbb{P}_{n,p_n}[\mathcal{D}_n] \leq \sum_{k=1}^{+\infty}\left(n(1-p_n)^{n/2}\right)^k = O(n(1-p_n)^{n/2}) = o(1),$$

where we used that the geometric series (started at $k=1$) is dominated asymptotically by its first term. ■

# Go deeper

More details and examples on the first and second moment methods at:

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

For more on random graphs in general, see e.g. (available online):

- *Random Graphs and Complex Networks. Vol. I and II* by van der Hofstad
- *Random Graph Dynamics* by Durrett

# Probability on Graphs: Techniques and Applications to Data Science

## *2 - Exponential tail bounds*

Sébastien Roch
*UW–Madison*
*Mathematics*

July 26, 2018

1. Chernoff-Cramér method

2. Epsilon-net arguments

3. Application: Community detection

# Moment-generating function

## Definition

The *moment-generating function* of $X$ is the function

$$M_X(s) = \mathbb{E}\left[e^{sX}\right],$$

defined for all $s \in \mathbb{R}$ where it is finite, which includes $s = 0$.

# Chernoff-Cramér bound

Assume $X$ is a centered (i.e. mean 0) random variable such that $M_X(s) < +\infty$ for $s \in (-s_0, s_0)$ for some $s_0 > 0$. Exponentiating within Markov's inequality gives, for any $\beta > 0$ and $s > 0$,

$$\mathbb{P}[X \geq \beta] = \mathbb{P}[e^{sX} \geq e^{s\beta}] \leq \frac{M_X(s)}{e^{s\beta}} = \exp\left[-\left\{s\beta - \Psi_X(s)\right\}\right],$$

where $\Psi_X(s) = \log M_X(s)$. The best exponent is

$$\Psi_X^*(\beta) = \sup_{s \in \mathbb{R}_+} (s\beta - \Psi_X(s)).$$

# Chernoff-Cramér for sums of independent variables

Let $S_n = \sum_{i \leq n} X_i$, where the $X_i$s are i.i.d. centered random variables. Then

$$\Psi_{S_n}(s) = \log \mathbb{E}[e^{s \sum_{i \leq n} X_i}] = \log \prod_{i \leq n} \mathbb{E}[e^{sX_i}] = n\Psi_{X_1}(s)$$

## Theorem

*Assume $M_{X_1}(s) < +\infty$ on $s \in (-s_0, s_0)$ for some $s_0 > 0$. For any $\beta > 0$,*

$$\mathbb{P}[S_n \geq \beta] \leq \exp\left(-n\Psi^*_{X_1}\left(\frac{\beta}{n}\right)\right).$$

# Example: Binomial

Let $Z_n$ be a binomial random variable with parameters $n$ and $p$. Recall that $Z_n$ is a sum of i.i.d. indicators $Y_1, \ldots, Y_n$ and, letting $X_i = Y_i - p$ and $S_n = Z_n - np$,

$$\Psi_{X_1}(s) = \log \mathbb{E}[e^{s(Y_1 - p)}] = \log\left(pe^s + (1-p)\right) - ps.$$

For $b \in (0, 1-p)$, letting $a = b + p$, direct calculation gives

$$\begin{aligned}\Psi^*_{X_1}(b) &= \sup_{s>0}(sb - (\log\left[pe^s + (1-p)\right] - ps)) \\ &= (1-a)\log\frac{1-a}{1-p} + a\log\frac{a}{p} =: D(a\|p),\end{aligned}$$

achieved at $s_b = \log\frac{(1-p)a}{p(1-a)}$. By the previous result, for $\beta > 0$,

$$\mathbb{P}[Z_n \geq np + \beta] \leq \exp\left(-n\,D\left(p + \beta/n \| p\right)\right).$$

# Sub-Gaussian variables I

Let $X \sim N(0,\nu)$ where $\nu > 0$ and note that

$$\begin{aligned} M_X(s) &= \int_{-\infty}^{+\infty} e^{sx} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{x^2}{2\nu}}\,\mathrm{d}x = \int_{-\infty}^{+\infty} e^{\frac{s^2\nu}{2}} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{(x-s\nu)^2}{2\nu}}\,\mathrm{d}x \\ &= \exp\left(\frac{s^2\nu}{2}\right), \end{aligned}$$

so that straightforward calculus gives for $\beta > 0$

$$\Psi_X^*(\beta) = \sup_{s>0}(s\beta - s^2\nu/2) = \frac{\beta^2}{2\nu},$$

achieved at $s_\beta = \beta/\nu$. Plugging $\Psi_X^*(\beta)$ into Theorem 2 leads for $\beta > 0$ to the bound

$$\mathbb{P}[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right).$$

# Sub-Gaussian variables II

We say that a centered random variable $X$ is *sub-Gaussian with variance factor* $\nu > 0$ if for all $s \in \mathbb{R}$

$$\Psi_X(s) \leq \frac{s^2 \nu}{2},$$

which is denoted by $X \in \mathcal{G}(\nu)$. By the Chernoff-Cramér bound

$$\mathbb{P}[X \leq -\beta] \vee \mathbb{P}[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right),$$

where we used that $X \in \mathcal{G}(\nu)$ implies $-X \in \mathcal{G}(\nu)$.

# Example: Back to the binomial

## Theorem (Case $p = 1/2$)

*Let $X_1, \ldots, X_n$ be independent $\{-1, 1\}$-valued random variables with $\mathbb{P}[X_i = 1] = \mathbb{P}[X_i = -1] = 1/2$. Let $S_n = \sum_{i \leq n} X_i$. Then, for any $\beta > 0$,*

$$\mathbb{P}[S_n \geq \beta] \leq e^{-\beta^2/2n}.$$

*Proof:* The moment-generating function of $X_1$ can be bounded as follows

$$M_{X_1}(s) = \frac{e^s + e^{-s}}{2} = \sum_{j \geq 0} \frac{s^{2j}}{(2j)!} \leq \sum_{j \geq 0} \frac{(s^2/2)^j}{j!} = e^{s^2/2}. \qquad (1)$$

So $\Psi_{S_n}(s) = n\Psi_{X_1}(s) \leq s^2 n/2$ and $S_n \in \mathcal{G}(n)$. ■

# Sub-Gaussian variables III

## Theorem (General Hoeffding inequality)

*Let $X_1, \ldots, X_n$ be independent centered random variables with $X_i \in \mathcal{G}(\nu_i)$ for $0 < \nu_i < +\infty$ and let $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. Let $S_n = \sum_{i\leq n} \alpha_i X_i$. Then $S_n \in \mathcal{G}(\sum_{i=1}^n \alpha_i^2 \nu_i)$ and for all $\beta > 0$,*

$$\mathbb{P}\left[S_n \geq \beta\right] \leq \exp\left(-\frac{\beta^2}{2\sum_{i=1}^n \alpha_i^2 \nu_i}\right).$$

*Proof:* By independence,

$$\Psi_{S_n}(s) = \sum_{i\leq n} \Psi_{\alpha_i X_i}(s) = \sum_{i\leq n} \Psi_{X_i}(s\alpha_i) \leq \sum_{i\leq n} \frac{(s\alpha_i)^2 \nu_i}{2} = \frac{s^2 \sum_{i\leq n} \alpha_i^2 \nu_i}{2}.$$

■

# Example: Bounded variables I

For bounded random variables, the previous inequality reduces to a standard bound.

## Theorem (Hoeffding's inequality)

*Let $X_1, \ldots, X_n$ be independent random variables where, for each i, $X_i$ takes values in $[a_i, b_i]$ with $-\infty < a_i \leq b_i < +\infty$. Let $S_n = \sum_{i \leq n} X_i$. For all $\beta > 0$,*

$$\mathbb{P}[S_n - \mathbb{E}S_n \geq \beta] \leq \exp\left(-\frac{2\beta^2}{\sum_{i \leq n}(b_i - a_i)^2}\right).$$

# Illustration: Maximum degree of Erdös-Rényi

Let $G_n \sim \mathbb{G}_{n,p}$ be an Erdös-Rényi graph with $n$ vertices and density $p_n = p \in (0,1)$. Let $D_i$ be the degree of vertex $i$ and let $D^* = \max_i D_i$. Note that $D_i$ is $\mathrm{Bin}(n-1,p)$, i.e. a sum of independent $[0,1]$-variables, so by Hoeffding's inequality

$$\mathbb{P}_{n,p}[D_i - (n-1)p \geq \sqrt{(1+\varepsilon) n \log(n)/2}] \leq e^{-(1+\varepsilon)\log n}.$$

By a union bound

$$\begin{aligned}
&\mathbb{P}_{n,p}[D^* \geq (n-1)p + \sqrt{(1+\varepsilon) n \log(n)/2}] \\
&\quad \leq \sum_i \mathbb{P}_{n,p}[D_i - (n-1)p \geq \sqrt{(1+\varepsilon) n \log(n)/2}] \\
&\quad \leq n \times n^{-(1+\varepsilon)} \to 0.
\end{aligned}$$

# Example: Bounded variables II

*Proof:* By the general Hoeffding inequality, it suffices to show that $X_i - \mathbb{E}X_i \in \mathcal{G}(\nu_i)$ with $\nu_i = \frac{1}{4}(b_i - a_i)^2$. We give a quick proof of a weaker version that uses a trick called *symmetrization*. Suppose the $X_i$s are centered and satisfy $|X_i| \leq c_i$ for some $c_i > 0$. Let $X_i'$ be an independent copy of $X_i$ and let $Z_i$ be an independent uniform in $\{-1, 1\}$. By Jensen's inequality

$$\mathbb{E}\left[e^{sX_i}\right] = \mathbb{E}\left[e^{s\mathbb{E}[X_i - X_i' \mid X_i]}\right] \leq \mathbb{E}\left[\mathbb{E}\left[e^{s(X_i - X_i')} \,\middle|\, X_i\right]\right] = \mathbb{E}\left[e^{s(X_i - X_i')}\right].$$

By the symmetry of $X_i - X_i'$, we then get

$$\begin{aligned}\mathbb{E}\left[e^{s(X_i - X_i')}\right] &= \mathbb{E}\left[e^{sZ_i(X_i - X_i')}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{sZ_i(X_i - X_i')} \,\middle|\, X_i, X_i'\right]\right] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[e^{(s(X_i - X_i'))^2/2} \,\middle|\, X_i, X_i'\right]\right] \leq \mathbb{E}\left[e^{(s(X_i - X_i'))^2/2}\right] \leq e^{-2c_i^2 s^2}.\end{aligned}$$

■

## Many more concentration inequalities

- Bernstein's inequality
- Azuma's inequality
- Matrix inequalities

1 Chernoff-Cramér method

2 Epsilon-net arguments

3 Application: Community detection

# Epsilon-nets I

Exponential tail inequalities are useful, among other things, to study the deviations of suprema of random variables. When the supremum is over an *infinite* index set, one way to proceed is to apply a tail inequality to a sufficiently dense finite subset of the index set, and then extend the resulting bound by continuity. This is referred to as an $\varepsilon$*-net argument*.

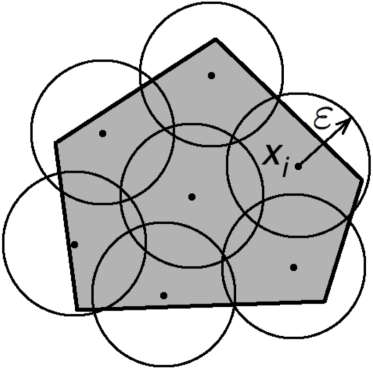
# Epsilon-nets II

**Definition ($\varepsilon$-net)**

Let $S$ be a subset of a metric space $(M, \rho)$ and let $\varepsilon > 0$. A collection of points $N \subseteq S$ is called an *$\varepsilon$-net of $S$* if all pairs of points in $N$ are at distance greater than $\varepsilon$ and $N$ is maximal by inclusion in $S$. In particular for all $z \in S$, $\inf_{y \in N} \rho(z, y) \leq \varepsilon$. The *covering number* of $S$, denoted by $\mathcal{N}(S, \rho, \varepsilon)$, is the smallest cardinality of an $\varepsilon$-net of $S$.

The definition of an $\varepsilon$-net immediately suggests an algorithm for constructing one. Start with $N = \emptyset$ and successively add a point to $N$ at distance at least $\varepsilon$ from all other previous points until that is not possible to do so anymore. (Provided $S$ is compact, this procedure will terminate after a finite number of steps.)

# Epsilon-nets by picture



(a) This covering of a pentagon $K$ by seven $\varepsilon$-balls shows that $\mathcal{N}(K, \varepsilon) \leq 7$.

# Illustration: Spectral norm of random matrix I

For a $m \times n$ matrix $A \in \mathbb{R}^{m \times n}$, recall that the spectral norm is defined as

$$\|A\| := \sup_{\mathbf{x} \in \mathbb{R}^n \setminus \{0\}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x} \in \mathbb{S}^{n-1}} \|A\mathbf{x}\|_2 = \sup_{\substack{\mathbf{x} \in \mathbb{S}^{n-1} \\ \mathbf{y} \in \mathbb{S}^{m-1}}} \langle A\mathbf{x}, \mathbf{y} \rangle,$$

where $\mathbb{S}^{n-1}$ is the sphere of radius 1 around the origin in $\mathbb{R}^n$.

(To see the rightmost equality above, note that Cauchy-Schwarz implies $\langle A\mathbf{x}, \mathbf{y} \rangle \leq \|A\mathbf{x}\|_2 \|\mathbf{y}\|_2$ and that one can take $\mathbf{y} = A\mathbf{x}/\|A\mathbf{x}\|_2$ for any $\mathbf{x}$ such that $A\mathbf{x} \neq 0$ in the rightmost expression.)

# Illustration: Spectral norm of random matrix II

## Theorem

*Let $A \in \mathbb{R}^{m \times n}$ be a random matrix whose entries are centered, independent and sub-Gaussian with variance factor $\nu$. Then there exist a constant $0 < C < +\infty$ such that, for all $t > 0$,*

$$\|A\| \leq C\sqrt{\nu}(\sqrt{m} + \sqrt{n} + t),$$

*with probability at least $1 - e^{-t^2}$.*

Without independence of the entries, the spectral norm can be much larger. Say $A$ is all-$(+1)$ or all-$(-1)$ with equal probability. Taking the vector $\mathbf{x} = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$ shows that $\|A\| \geq n$ with probability 1.

# Illustration: Spectral norm of random matrix III

*Proof:* We seek to bound

$$\|A\| = \sup_{\substack{\mathbf{x}\in\mathbb{S}^{n-1}\\ \mathbf{y}\in\mathbb{S}^{m-1}}} \langle A\mathbf{x}, \mathbf{y}\rangle = \sup_{\substack{\mathbf{x}\in\mathbb{S}^{n-1}\\ \mathbf{y}\in\mathbb{S}^{m-1}}} \sum_{i,j} x_i y_j A_{ij},$$

where we note that the last quantity is a linear combination of independent variables. Fix $\varepsilon = 1/4$. We proceed in two steps:

1. We first apply the general Hoeffding inequality to control the deviations of the supremum *restricted to $\varepsilon$-nets $N$ and $M$ of $\mathbb{S}^{n-1}$ and $\mathbb{S}^{m-1}$*.
2. We then extend the bound to the full supremum by continuity.

# Back to $\varepsilon$-nets: Sphere

Let $\mathbb{S}^{k-1}$ be the sphere of radius 1 centered around the origin in $\mathbb{R}^k$ with the Euclidean metric. Let $0 < \varepsilon < 1$. We claim that

$$\mathcal{N}(S, \rho, \varepsilon) \leq \left(\frac{3}{\varepsilon}\right)^k .$$

Let $N$ be any $\varepsilon$-net of $S$. The balls of radius $\varepsilon/2$ around points in $N$, $\{\mathbb{B}^k(x_i, \varepsilon/2) \,:\, x_i \in N\}$, satisfy two properties:

1. Pairwise disjoint: if $z \in \mathbb{B}^k(x_i, \varepsilon/2) \cap \mathbb{B}^k(x_j, \varepsilon/2)$, then $\|x_i - x_j\|_2 \leq \|x_i - z\|_2 + \|x_j - z\|_2 \leq \varepsilon$, a contradiction.
2. Contained in $\mathbb{B}^k(0, 3/2)$: if $z \in \mathbb{B}^k(x_i, \varepsilon/2)$, then $\|z\|_2 \leq \|z - x_i\|_2 + \|x_i\| \leq \varepsilon/2 + 1 \leq 3/2$.

The volume of a ball of radius is $\varepsilon/2$ is $\frac{\pi^{k/2}(\varepsilon/2)^k}{\Gamma(k/2+1)}$ and that of a ball of radius $3/2$ is $\frac{\pi^{k/2}(3/2)^k}{\Gamma(k/2+1)}$. Divide one by the other.

# Illustration: Spectral norm of random matrix IV

**Lemma**

*Let N and M be as above. For C large enough, for all $t > 0$,*

$$\mathbb{P}\left[\max_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2} C\sqrt{\nu}(\sqrt{m} + \sqrt{n} + t)\right] \leq e^{-t^2}.$$

*Proof:* By the general Hoeffding inequality, $\langle A\mathbf{x}, \mathbf{y}\rangle$ is sub-Gaussian with variance factor

$$\sum_{i,j}(x_i y_j)^2 \nu = \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \nu = \nu,$$

for all $\mathbf{x} \in N$ and $\mathbf{y} \in M$. In particular, for all $\beta > 0$,

$$\mathbb{P}\left[\langle A\mathbf{x}, \mathbf{y}\rangle \geq \beta\right] \leq \exp\left(-\frac{\beta^2}{2\nu}\right).$$

# Illustration: Spectral norm of random matrix V

*Proof of lemma (continued):* Hence, by a union bound over $N$ and $M$,

$$\begin{aligned}
\mathbb{P}&\left[\max_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2} C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right]\\
&\leq \sum_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \mathbb{P}\left[\langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2} C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right]\\
&\leq |N||M| \exp\left(-\frac{1}{2\nu}\left\{\frac{1}{2} C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right\}^2\right)\\
&\leq 12^{n+m} \exp\left(-\frac{C^2}{8}\left\{m+n+t^2)\right\}\right)\\
&\leq e^{-t^2},
\end{aligned}$$

for $C^2/8 = \log 12 \geq 1$, where in the third inequality we ignored all cross-products since they are non-negative. ∎

# Illustration: Spectral norm of random matrix VI

## Lemma

*For any $\varepsilon$-nets N and M of $\mathbb{S}^{n-1}$ and $\mathbb{S}^{m-1}$ respectively, the following inequalities hold*

$$\sup_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \leq \|A\| \leq \frac{1}{1-2\varepsilon} \sup_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle.$$

*Proof:* The first inequality is immediate. For the second inequality, we will use the following observation

$$\langle A\mathbf{x}, \mathbf{y}\rangle - \langle A\mathbf{x}_0, \mathbf{y}_0\rangle = \langle A\mathbf{x}, \mathbf{y} - \mathbf{y}_0\rangle + \langle A(\mathbf{x} - \mathbf{x}_0), \mathbf{y}_0\rangle.$$

Fix $x \in \mathbb{S}^{n-1}$ and $y \in \mathbb{S}^{m-1}$ such that $\langle A\mathbf{x}, \mathbf{y}\rangle = \|A\|$, and let $\mathbf{x}_0 \in N$ and $\mathbf{y}_0 \in M$ such that

$$\|\mathbf{x} - \mathbf{x}_0\|_2 \leq \varepsilon \qquad \text{and} \qquad \|\mathbf{y} - \mathbf{y}_0\|_2 \leq \varepsilon.$$

# Illustration: Spectral norm of random matrix VII

*Proof of lemma (continued):* Then the inequality above, Cauchy-Schwarz and the definition of the spectral norm imply

$$\|A\| - \langle A\mathbf{x}_0, \mathbf{y}_0\rangle \leq \|A\|\|\mathbf{x}\|_2\|\mathbf{y} - \mathbf{y}_0\|_2 + \|A\|\|\mathbf{x} - \mathbf{x}_0\|_2\|\mathbf{y}_0\|_2 \leq 2\varepsilon\|A\|.$$

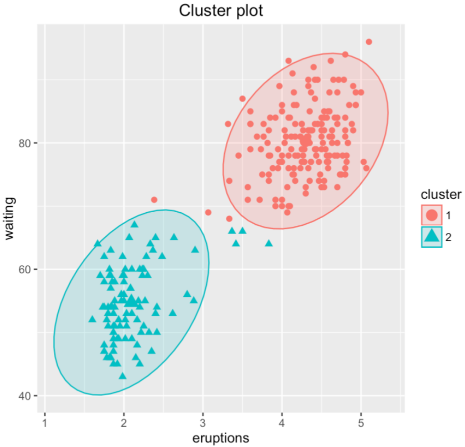
Rearranging gives the claim. ■

1 Chernoff-Cramér method

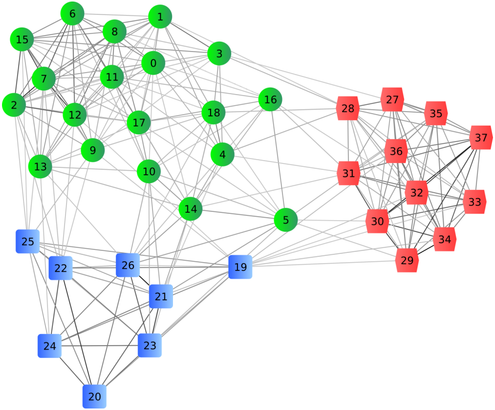
2 Epsilon-net arguments

3 Application: Community detection

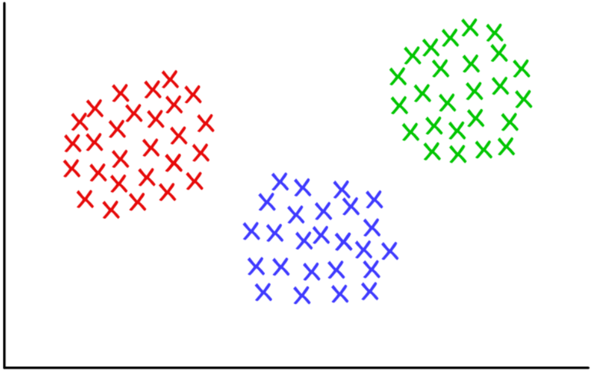
# Clustering in Euclidean space



# Clustering in graphs



# Reducing the second problem to the first one



# Stochastic blockmodel with two balanced blocks

## Definition

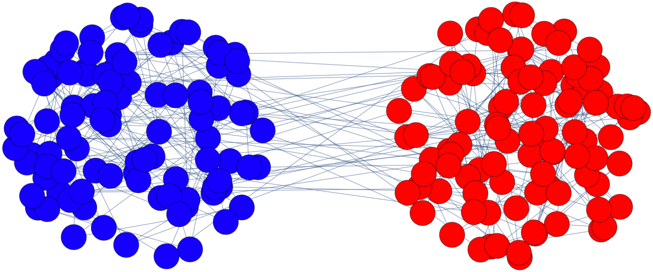
Let $V = [n]$ with $n$ even, let $V_1 = \{1, \ldots, n/2\}$ and $V_2 = \{n/2 + 1, \ldots, n\}$, and let $0 < q < p < 1$. We draw a graph $G = (V, E)$ at random as follows. For each pair $x \neq y$ in $V$, the edge $\{x, y\}$ is in $E$ with probability:

- $p$ if $x, y \in V_1$, or $x, y \in V_2$;
- $q$ if $x \in V_1$ and $y \in V_2$, or $x \in V_2$ and $y \in V_1$;

independently of all other edges. We write $G \sim \mathrm{SBM}_{n,p,q}$ and we denote the corresponding measure by $\mathbb{P}_{n,p,q}$.

**Community detection problem:** Given $G$ (without the node labels), output $V_1$, $V_2$ (possibly approximately).

# Stochastic blockmodel by picture



# Expected adjacency matrix

Let $G \sim \mathrm{SBM}_{n,p,q}$ and let $A$ be the adjacency matrix of $G$.

## Theorem

*Let $D = \mathbb{E}_{n,p,q}[A]$. Then*

$$D = n\frac{p+q}{2}\mathbf{u}_1\mathbf{u}_1^T + n\frac{p-q}{2}\mathbf{u}_2\mathbf{u}_2^T - p\,I,$$

*where* $\mathbf{u}_1 = \frac{1}{\sqrt{n}}(1,\ldots,1)^T$ *and* $\mathbf{u}_2 = \frac{1}{\sqrt{n}}(1,\ldots,1,-1,\ldots,-1)^T$.

*Proof:* Note that $D$ is a block matrix with diagonal blocks all-$p$ and off-diagonal blocks all-$q$, all of size $n/2 \times n/2$, with the exception of the diagonal which is all-0. ■

**Idea:** Compute the second eigenvector of $A$ and cluster by sign.

# Spectral clustering: a positive result

### Theorem

*Let $G \sim \mathrm{SBM}_{n,p,q}$ and let $A$ be the adjacency matrix of $G$. Let $\mu = \min\left\{q, \frac{p-q}{2}\right\} > 0$. Clustering according to the sign of the second eigenvector of $A$ identifies the two communities of $G$ with probability at least $1 - e^{-n}$, except for $C/\mu^2$ misclassified nodes for some constant $C > 0$.*

# Matrix perturbation

## Theorem (A version of Davis-Kahan)

*Let $S$ and $T$ be symmetric $n \times n$ matrices. Let $\lambda_i(S)$ be the $i$-th largest eigenvalue of $S$ with corresponding unit eigenvector $\mathbf{v}_i(S)$ (and similarly for $T$). If*

$$\delta := \min_{j \neq i} |\lambda_i(S) - \lambda_j(S)| > 0,$$

*then there is $\theta \in \{+1, -1\}$ such that*

$$\|\mathbf{v}_i(S) - \theta\, \mathbf{v}_i(T)\|_2 \leq \frac{4\|S - T\|}{\delta}.$$

# Bounding the spectral norm

## Lemma

*Let $G \sim \mathrm{SBM}_{n,p,q}$, let $A$ be the adjacency matrix of $G$ and let $D = \mathbb{E}_{n,p,q}[A]$. Then, there is a constant $C > 0$ such that*

$$\|A - D\| \leq C\sqrt{n},$$

*with probability at least $1 - e^{-n}$.*

*Proof:* The entries of $R$ are centered, independent and sub-Gaussian with variance factor $1/4$. ■

# Spectral clustering: proof I

*Proof of spectral clustering theorem:* The eigenvalues of $D$ are

$$n\,\frac{p+q}{2} - p, \qquad n\,\frac{p-q}{2} - p, \qquad -p,$$

so $\lambda_2(D) = n\,\frac{p-q}{2} - p$ and

$$\delta = \min_{j \neq 2} |\lambda_2(D) - \lambda_j(D)| = \min\left\{n\,\frac{p-q}{2}, n\,q\right\} =: n\,\mu > 0.$$

By Davis-Kahan and the previous lemma, with probability at least $1 - e^{-n}$, there is $\theta \in \{+1, -1\}$ such that

$$\|\mathbf{v}_2(D) - \theta\,\mathbf{v}_2(A)\|_2 \leq \frac{4C\sqrt{n}}{n\,\mu} \leq \frac{C'}{\sqrt{n}\,\mu}.$$

# Spectral clustering: proof II

*Proof of spectral clustering theorem (continued):* Put differently,

$$\sum_i \left|\sqrt{n}\,(\mathbf{v}_2(D))_i - \sqrt{n}\,\theta\,(\mathbf{v}_2(A))_i\right|^2 \leq \frac{(C')^2}{\mu^2}.$$

If the signs of $(\mathbf{v}_2(D))_i$ and $\theta\,(\mathbf{v}_2(A))_i$ disagree, then the $i$-th term in the sum above is $\geq 1$. So there can be at most $(C')^2/\mu^2$ of those. That establishes the desired bound on the number of misclassified nodes. ■

# Go deeper

More details and examples on tail bounds at:

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

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

- *High-dimensional probability: An introduction with applications in data science* by Vershynin
- *Probability in High Dimension* by van Handel

# Probability on Graphs: Techniques and Applications to Data Science

## *3 - Spectral Techniques*

Sébastien Roch
*UW–Madison*
*Mathematics*

July 26, 2018

1. Review of Markov chains
2. Bounding the mixing time via the spectral gap
3. Bottleneck ratio and Cheeger's inequality
4. Application: Gibbs sampling at low temperature

# 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* $\pi$ 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 $\pi$ is a *stationary distribution* if in addition $\pi$ 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 $\eta$ 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 $\pi$. Then, for all $x, y$, $P^t(x,y) \to \pi(y)$ as $t \to +\infty$.*

# Convergence II

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

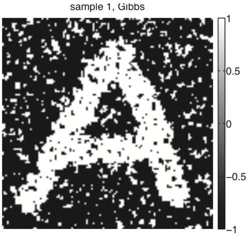
### Definition (Mixing time)

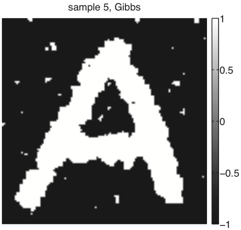
The *mixing time* is $\mathrm{t}_{\mathrm{mix}}(\varepsilon) := \min\{t \geq 0 \,:\, d(t) \leq \varepsilon\}$, where $d(t) := \max_{x \in V} \|P^t(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}}$.

# Other useful random walk quantities

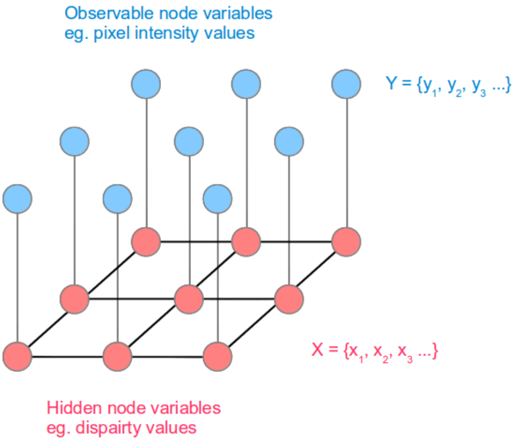
- Hitting times
- Cover times
- Heat kernels

# Application: Bayesian image analysis I





# Application: Bayesian image analysis II



Observable node variables
eg. pixel intensity values

$Y = \{y_1, y_2, y_3 ...\}$

$X = \{x_1, x_2, x_3 ...\}$

Hidden node variables
eg. dispairty values

# 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 $\mu$ 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

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

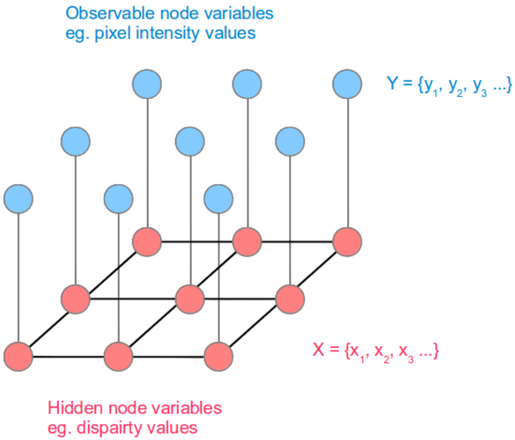
where $x_K$ is $x$ restricted to the vertices of $K$ and $\mathcal{Z}$ is a normalizing constant.

# Application: Undirected graphical models II

## Example

For $\beta > 0$, the *ferromagnetic Ising model* with inverse temperature $\beta$ 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


Observable node variables
eg. pixel intensity values
$Y = \{y_1, y_2, y_3 \ldots\}$
$X = \{x_1, x_2, x_3 \ldots\}$
Hidden node variables
eg. dispairty values

# 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 $\tau$ 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 $\mu_\beta$ 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 $\mu_\beta$ 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 $\sigma$ with the spin at $i$ being set to $\gamma$. 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. $\mu_\beta$. How quickly does the chain approach $\mu_\beta$?

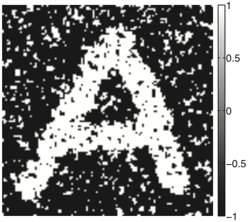
# Application: Gibbs sampling III

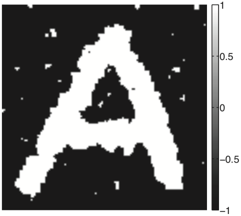
*Proof of reversibility:* This chain is clearly irreducible. For all $\sigma \in \mathcal{X}$ and $i \in V$, let $S_{\neq i}(\sigma) := \mathcal{H}(\sigma^{i,+}) + S_i(\sigma) = \mathcal{H}(\sigma^{i,-}) - S_i(\sigma)$. We have

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

■

# Application: Back to Bayesian image analysis


sample 1, Gibbs


sample 5, Gibbs

1. Review of Markov chains
2. Bounding the mixing time via the spectral gap
3. Bottleneck ratio and Cheeger's inequality
4. Application: Gibbs sampling at low temperature

# 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 $D_\pi$ be the diagonal matrix with $\pi$ 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

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

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 $D_\lambda$ 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 $\lambda$ 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 $\mathrm{t}_{\mathrm{rel}} := \gamma_*^{-1}$.

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 $\pi$. Let $\pi_{\min} = \min_x \pi(x)$. For all $\varepsilon > 0$,*

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

# 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*

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

*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 $\mathrm{t_{mix}}(\varepsilon) = O(n^2 \log n)$ and $\mathrm{t_{mix}}(\varepsilon) = \Omega(n^2)$.

1. Review of Markov chains
2. Bounding the mixing time via the spectral gap
3. **Bottleneck ratio and Cheeger's inequality**
4. Application: Gibbs sampling at low temperature

# Back to eigenvalues I

## Theorem (Rayleigh's quotient)

*Let $P$ be irreducible and reversible with respect to $\pi$. 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 $\pi$. So the variational characterization of $\lambda_2$ translates into

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

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)

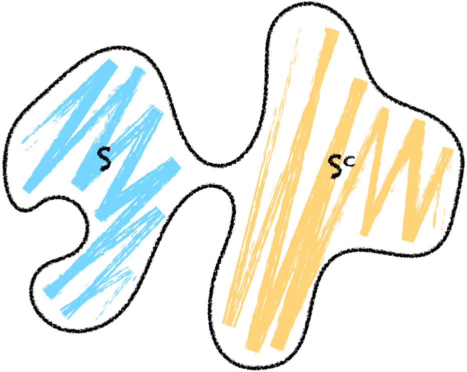
For a subset of states $S \subseteq V$, the *bottleneck ratio* of $S$ is

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

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 $t_{\mathrm{rel}} = \gamma^{-1}$, 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,

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

and

$$\frac{n}{2} \leq \mathrm{t}_{\mathrm{rel}} = \gamma^{-1} \leq 2n^2.$$

Thus

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

and

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

1. Review of Markov chains
2. Bounding the mixing time via the spectral gap
3. Bottleneck ratio and Cheeger's inequality
4. Application: Gibbs sampling at low temperature

# 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* $\beta$ 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 $\sigma_i$ according to $\mu_\beta(\sigma)$ conditioned on agreeing with $\sigma$ 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 $\sigma$ with the state at $i$ being set to $\gamma$. 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 $\beta$ with $n$. We define $\alpha := \beta(n-1)$.

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

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

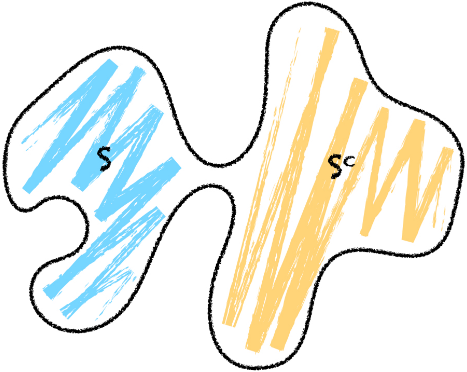
# Curie-Weiss model II

*Proof:* We only prove exponential mixing when $\alpha$ 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 $-1$s 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

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

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

# Curie-Weiss model VI

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

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

■

# 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

1. Definitions and basic properties
2. Couplings of Markov chains
3. Path coupling
4. Back to the application: Gibbs sampling at high temperature

# Coupling

## Definition

Let $\mu$ and $\nu$ be probability measures on the same measurable space $(S, \mathcal{S})$. A *coupling* of $\mu$ and $\nu$ is a probability measure $\gamma$ on the product space $(S \times S, \mathcal{S} \times \mathcal{S})$ such that the *marginals* of $\gamma$ coincide with $\mu$ and $\nu$, 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 $\mu$ and $\nu$ 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 $\mu$ and $\nu$ be probability measures on $(S, \mathcal{S})$. For any coupling $(X, Y)$ of $\mu$ and $\nu$,*

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

# Coupling inequality II

*Proof:*

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

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

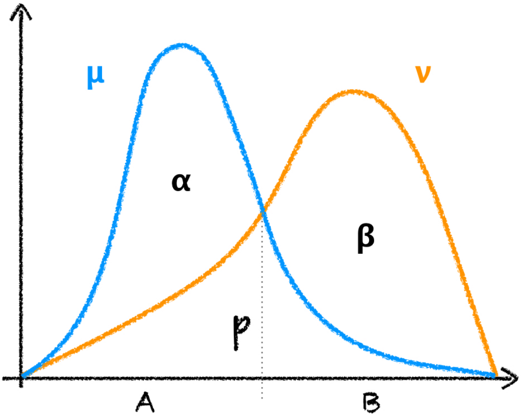
In fact, the inequality is tight.

**Lemma**

*Assume $S$ is finite and let $\mathcal{S} = 2^S$. Let $\mu$ and $\nu$ 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 $\mu$ and $\nu$ 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

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

which coincides with the monotone coupling.

1. Definitions and basic properties
2. Couplings of Markov chains
3. Path coupling
4. Back to the application: Gibbs sampling at high temperature

# Bounding the mixing time via coupling I

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

$$\mathrm{t}_{\mathrm{mix}}(\varepsilon) := \min\{t \,:\, d(t) \leq \varepsilon\},$$

where

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

It will be easier to work with

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

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 $\tau_{\mathrm{coal}}$ be the *coalescence time* (also called coupling time), i.e.,

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

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

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

# 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 $\pi$. Then*

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

*In particular*

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

# 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 $\tau_{\mathrm{coll}}$ be the time it takes to update each coordinate at least once. Then, for any $c > 0$,*

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

*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

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

■

# 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:

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

1. Definitions and basic properties

2. Couplings of Markov chains

3. Path coupling

4. Back to the application: Gibbs sampling at high temperature

# 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 $\pi$. 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*

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

# 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

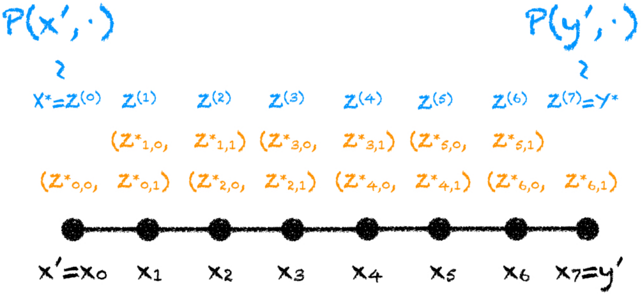
$$\begin{aligned}
\mathbb{E}_{(x,y)}[w_0(X_t, Y_t)] &= \mathbb{E}_{(x,y)}\left[\mathbb{E}[w_0(X_t, Y_t) \,|\, X_{t-1}, Y_{t-1}]\right] \\
&\leq \mathbb{E}_{(x,y)}\left[\kappa\, w_0(X_{t-1}, Y_{t-1})\right] \leq \cdots \leq \kappa^t\, \mathbb{E}_{(x,y)}[w_0(X_0, Y_0)] \\
&= \kappa^t\, w_0(x, y) \leq \kappa^t\,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

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

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

which implies the theorem.

# Global coupling lemma: proof by picture


$P(x',\cdot)$
$\wr$
$P(y',\cdot)$
$\wr$
$X^*=Z^{(0)}$
$Z^{(1)}$
$Z^{(2)}$
$Z^{(3)}$
$Z^{(4)}$
$Z^{(5)}$
$Z^{(6)}$
$Z^{(7)}=Y^*$
$(Z^*_{1,0},\ Z^*_{1,1})$
$(Z^*_{3,0},\ Z^*_{3,1})$
$(Z^*_{5,0},\ Z^*_{5,1})$
$(Z^*_{0,0},\ Z^*_{0,1})$
$(Z^*_{2,0},\ Z^*_{2,1})$
$(Z^*_{4,0},\ Z^*_{4,1})$
$(Z^*_{6,0},\ Z^*_{6,1})$
$x'=x_0$
$x_1$
$x_2$
$x_3$
$x_4$
$x_5$
$x_6$
$x_7=y'$

1. Definitions and basic properties
2. Couplings of Markov chains
3. Path coupling
4. Back to the application: Gibbs sampling at high temperature

# 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* $\beta$ 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 $\sigma_i$ according to $\mu_\beta(\sigma)$ conditioned on agreeing with $\sigma$ 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 $\sigma$ with the state at $i$ being set to $\gamma$. 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 $\mu_\beta$.

# Fast mixing at high temperature I

We show that the Glauber dynamics of the Ising model is fast mixing when the inverse temperature $\beta$ 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 $\sigma$ and $\omega$, i.e., if $\sigma_j = \omega_j$ for all $j \sim i_*$, we update $X^*$ from $\sigma$ 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 $\sigma$, 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 $\omega$. 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

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

and

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

# 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

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

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,

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

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

# Probability on Graphs: Techniques and Applications to Data Science

## *5 - Correlation decay*

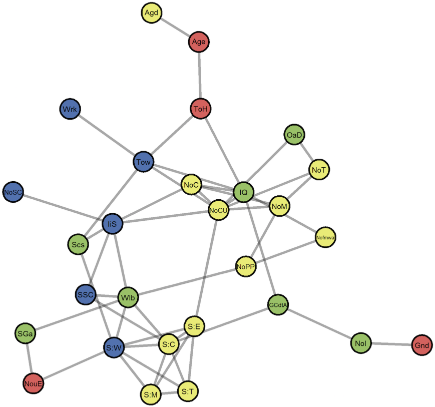
Sébastien Roch
*UW–Madison*
*Mathematics*

July 26, 2018

1 Correlation decay

2 Application: Reconstructing Markov random fields

# An undirected graphical model



# Recall: Ising model

Let $G = (V, E)$ be a finite, connected graph with maximum degree $\bar{\delta} = d$. Define $\mathcal{X} := A^V$ where $A = \{-1, +1\}$. Recall that the (ferromagnetic) Ising model on $V$ with *inverse temperature* $\beta$ 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 $\mathcal{Z}(\beta) := \sum_{\sigma \in \mathcal{X}} e^{-\beta \mathcal{H}(\sigma)}$.

# Correlation decay

**Spatial mixing:** How much does the state at one vertex “influence” the state at a vertex far away?

There are many ways to measure this. Let $X \sim \mu_\beta$ on $G = (V, E)$. For $u, v \in V$, define

$$d_C(u, v) = \sum_{x_u, x_v \in S} |\mathbb{P}[X_u = x_u, X_v = x_v] - \mathbb{P}[X_u = x_u]\mathbb{P}[X_v = x_v]|$$

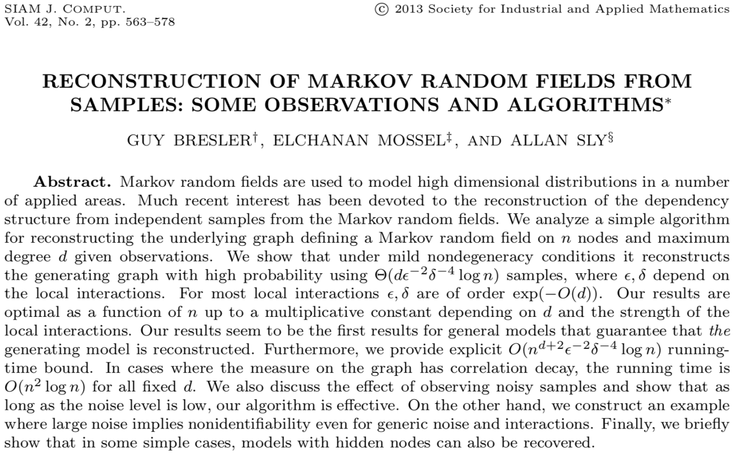
It can be shown in some cases that, when $\beta$ is large enough, the measure above decays exponentially with the graph distance. Such a statement can be useful to analyze the behavior of Ising models. This is easier seen on an example.

1. Correlation decay
2. Application: Reconstructing Markov random fields

# Structure learning

**Problem:** Let $X^{(1)}, \ldots, X^{(k)}$ be i.i.d. $\sim \mu_\beta$ on an unknown graph $G = (V, E)$ with maximal degree $\bar{\delta} = d$. How to recover $G$ from the samples $X^{(1)}, \ldots, X^{(k)}$?

# Bresler et al. (2013)



SIAM J. COMPUT.
Vol. 42, No. 2, pp. 563–578

## RECONSTRUCTION OF MARKOV RANDOM FIELDS FROM SAMPLES: SOME OBSERVATIONS AND ALGORITHMS*

GUY BRESLER[†], ELCHANAN MOSSEL[‡], AND ALLAN SLY[§]

**Abstract.** Markov random fields are used to model high dimensional distributions in a number of applied areas. Much recent interest has been devoted to the reconstruction of the dependency structure from independent samples from the Markov random fields. We analyze a simple algorithm for reconstructing the underlying graph defining a Markov random field on $n$ nodes and maximum degree $d$ given observations. We show that under mild nondegeneracy conditions it reconstructs the generating graph with high probability using $\Theta(d\epsilon^{-2}\delta^{-4}\log n)$ samples, where $\epsilon, \delta$ depend on the local interactions. For most local interactions $\epsilon, \delta$ are of order $\exp(-O(d))$. Our results are optimal as a function of $n$ up to a multiplicative constant depending on $d$ and the strength of the local interactions. Our results seem to be the first results for general models that guarantee that *the* generating model is reconstructed. Furthermore, we provide explicit $O(n^{d+2}\epsilon^{-2}\delta^{-4}\log n)$ running-time bound. In cases where the measure on the graph has correlation decay, the running time is $O(n^2 \log n)$ for all fixed $d$. We also discuss the effect of observing noisy samples and show that as long as the noise level is low, our algorithm is effective. On the other hand, we construct an example where large noise implies nonidentifiability even for generic noise and interactions. Finally, we briefly show that in some simple cases, models with hidden nodes can also be recovered.

# Recall: Gibbs sampling

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 $\sigma_i$ according to $\mu_\beta(\sigma)$ conditioned on agreeing with $\sigma$ 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 $\sigma$ with the state at $i$ being set to $\gamma$. Then the transition matrix is

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

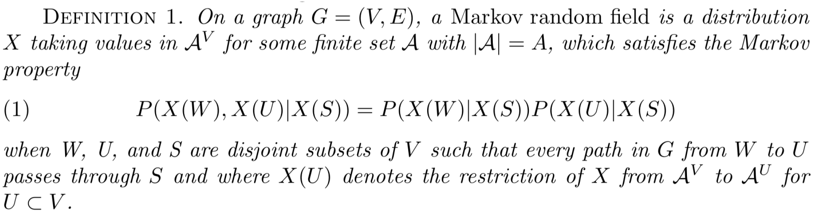
where

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

All other transitions have probability 0. Recall that this chain is irreducible and reversible with respect to $\mu_\beta$.

# Markov property

Let $X \sim \mu_\beta$ on $G = (V, E)$. Then $X$ satisfies the following.



DEFINITION 1. *On a graph* $G = (V, E)$, *a* Markov random field *is a distribution* $X$ *taking values in* $\mathcal{A}^V$ *for some finite set* $\mathcal{A}$ *with* $|\mathcal{A}| = A$, *which satisfies the Markov property*

$$P(X(W), X(U)|X(S)) = P(X(W)|X(S))P(X(U)|X(S)) \tag{1}$$

*when* $W$, $U$, *and* $S$ *are disjoint subsets of* $V$ *such that every path in* $G$ *from* $W$ *to* $U$ *passes through* $S$ *and where* $X(U)$ *denotes the restriction of* $X$ *from* $\mathcal{A}^V$ *to* $\mathcal{A}^U$ *for* $U \subset V$.

# One of BMS's Main Results

THEOREM 3. *For an assignment $x_U = (x_{u_1}, \ldots, x_{u_l})$ and $y \in \mathcal{A}$, define*

$$x_U^i(y) = (x_{u_1}, \ldots, y, \ldots, x_{u_l})$$

*to be the assignment obtained from $x_U$ by replacing the ith element by $y$. Suppose there exist $\epsilon, \delta > 0$ such that the following condition holds: for all $v \in V$, if $N(v) = \{u_1, \ldots, u_l\}$, then for each $i, 1 \le i \le l$, and for any set $W \subset V - (\{v\} \cup N(v))$ with $|W| \le d$, there exist values $x_v, x_{u_1}, \ldots, x_{u_i}, \ldots, x_{u_l}, y \in \mathcal{A}$, and $x_W \in \mathcal{A}^{|W|}$ such that*

$$\begin{aligned} &\big|P(X(v) = x_v | X(N(v)) = x_{N(v)}) \\ &\quad - P(X(v) = x_v | X(N(v)) = x^i_{N(v)}(y))\big| > \epsilon \end{aligned} \tag{16}$$

*and*

$$\begin{aligned} &P(X(N(v)) = x_{N(v)}, X(W) = x_W) > \delta, \\ &P(X(N(v)) = x^i_{N(v)}(y), X(W) = x_W) > \delta. \end{aligned} \tag{17}$$

*Then for some constant $C = C(\epsilon, \delta) > 0$, if $k > Cd \log n$, then there exists an estimator $\widehat{G}(\underline{X})$ such that the probability of correct reconstruction is $P(G = \widehat{G}(\underline{X})) = 1 - o(1)$. The estimator $\widehat{G}$ is computable in time $O(n^{2d+1} \log n)$.*

# Reconstructing neighborhoods

*Proof.* As in Theorem 2, we can assume that with high probability we have

$$\left|\widehat{P}(X(U) = x_U) - P(X(U) = x_U)\right| \leq \gamma \tag{18}$$

for all $U = \{u_i\}_{i=1}^{l} \subset V$ and $\{x_i\}_{i=1}^{l}$ when $l \leq 2d + 1$ and $k \geq C(\gamma) d \log n$, so we assume that (18) holds. For each vertex $v \in V$ we consider all candidate neighborhoods for $v$, subsets $U = \{u_1, \ldots, u_l\} \subset V - \{v\}$ with $0 \leq l \leq d$. For each candidate neighborhood $U$, the algorithm computes a score

$$\begin{aligned} f(v;U) = \min_{W,i} \max_{x_v, x_W, x_U, y} \Big|&\widehat{P}(X(v) = x_v | X(W) = x_W, X(U) = x_U) \\ &- \widehat{P}(X(v) = x_v | X(W) = x_W, X(U) = x_U^i(y))\Big|, \end{aligned}$$

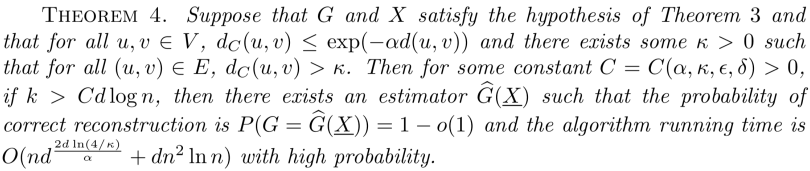
where for each $W, i$, the maximum is taken over all $x_v, x_W, x_U, y$, such that

$$\begin{aligned} \widehat{P}(X(W) = x_W, X(U) = x_U) &> \delta/2, \\ \widehat{P}(X(W) = x_W, X(U) = x_U^i(y)) &> \delta/2, \end{aligned} \tag{19}$$

and $W \subset V - (\{v\} \cup U)$ is an arbitrary set of nodes of size $d$, $x_W \in \mathcal{A}^d$ is an arbitrary assignment of values to the nodes in $W$, and $1 \leq i \leq l$.

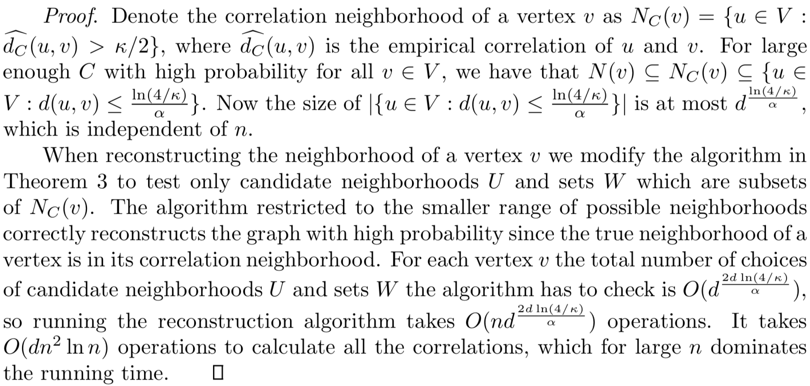
The algorithm selects as the neighborhood of $v$ the largest set $U \subset V - \{v\}$ with $f(v;U) > \epsilon/2$. It is necessary to check that if $U$ is the true neighborhood of $v$, then the algorithm accepts $U$, and otherwise the algorithm rejects $U$.

# A faster method under correlation decay



THEOREM 4. *Suppose that $G$ and $X$ satisfy the hypothesis of Theorem 3 and that for all $u, v \in V$, $d_C(u, v) \leq \exp(-\alpha d(u, v))$ and there exists some $\kappa > 0$ such that for all $(u, v) \in E$, $d_C(u, v) > \kappa$. Then for some constant $C = C(\alpha, \kappa, \epsilon, \delta) > 0$, if $k > Cd \log n$, then there exists an estimator $\widehat{G}(\underline{X})$ such that the probability of correct reconstruction is $P(G = \widehat{G}(\underline{X})) = 1 - o(1)$ and the algorithm running time is $O(nd^{\frac{2d \ln(4/\kappa)}{\alpha}} + dn^2 \ln n)$ with high probability.*

# Cutting down the number of potential neighborhoods



*Proof.* Denote the correlation neighborhood of a vertex $v$ as $N_C(v) = \{u \in V : \widehat{d_C}(u,v) > \kappa/2\}$, where $\widehat{d_C}(u,v)$ is the empirical correlation of $u$ and $v$. For large enough $C$ with high probability for all $v \in V$, we have that $N(v) \subseteq N_C(v) \subseteq \{u \in V : d(u,v) \leq \frac{\ln(4/\kappa)}{\alpha}\}$. Now the size of $|\{u \in V : d(u,v) \leq \frac{\ln(4/\kappa)}{\alpha}\}|$ is at most $d^{\frac{\ln(4/\kappa)}{\alpha}}$, which is independent of $n$.

When reconstructing the neighborhood of a vertex $v$ we modify the algorithm in Theorem 3 to test only candidate neighborhoods $U$ and sets $W$ which are subsets of $N_C(v)$. The algorithm restricted to the smaller range of possible neighborhoods correctly reconstructs the graph with high probability since the true neighborhood of a vertex is in its correlation neighborhood. For each vertex $v$ the total number of choices of candidate neighborhoods $U$ and sets $W$ the algorithm has to check is $O(d^{\frac{2d\ln(4/\kappa)}{\alpha}})$, so running the reconstruction algorithm takes $O(nd^{\frac{2d\ln(4/\kappa)}{\alpha}})$ operations. It takes $O(dn^2 \ln n)$ operations to calculate all the correlations, which for large $n$ dominates the running time. □

# Go deeper

More details and results in:

- Bresler, Mossel, Sly, *Reconstruction of Markov Random Fields from Samples: Some Observations and Algorithms*, SIAM J. Comput., 42(2):563–578.
- Bresler, *Efficiently Learning Ising Models on Arbitrary Graphs*, STOC 2015.