9. Approximate inference by sampling

- Markov Chain Monte Carlo methods
- Metropolis-Hastings algorithm
- Gibbs sampling
- Bounding mixing time via spectral analysis
- Bounding mixing time via coupling
Approximate inference with samples

- inference problem in graphical model

\[ \mu(x) = \frac{1}{Z} \prod_{(i,j) \in E} \psi_{ij}(x_i, x_j) \]

- belief propagation
  - fast (especially on sparse graphs) and very popular
  - deterministic
  - computes (incorrect) marginals on the computation tree

- approximate inference with samples
  given samples \( \{x^{(1)}, \ldots, x^{(N)}\} \) from distribution \( \mu(x) \)

\[ \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(x_{i}^{(i)} = x_i) \rightarrow \mu(x_i) \]

gives an approximate marginal
  - slower and difficult to decide when to stop
  - randomized
Generating samples from a distribution

<table>
<thead>
<tr>
<th>generating samples from $\mu(x)$</th>
<th>generating samples from $\mu(x_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Markov Chain Monte Carlo methods</td>
<td>sequential Monte Carlo methods</td>
</tr>
<tr>
<td>Metropolis-Hastings algorithm</td>
<td>particle filtering</td>
</tr>
</tbody>
</table>

- **Markov Chain Monte Carlo methods** work as follows
  - construct a Markov chain $P$ whose stationary distribution is equal to $\mu$
  - start from an arbitrary realization $x^{(0)}$ and run the Markov chain until it converges to its stationary distribution
  - this gives a sample from $\mu(x)$

- how do we construct such a Markov chain $P$?
- how long does it take for the Markov chain to converge?
Metropolis-Hastings algorithm

- Markov chain with a finite state space
  - a Markov chain is defined by a state space $\mathcal{X}^n$ and a $|\mathcal{X}|^n \times |\mathcal{X}|^n$ dimensional transition matrix $P$ such that
    \[ P_{xy} = \mathbb{P}(x_{t+1} = y | x_t = x) \]
  - stationary distribution of a Markov chain is a $|\mathcal{X}|^n$-dim row vector of distribution such that
    \[ \pi^T P = \pi^T \]
  - a Markov chain is **reversible** if there exists a probability distribution $\pi$ such that the **detailed balance equation** is satisfied:
    \[ \pi_x P_{xy} = \pi_y P_{yx} \quad \text{for all } x, y \]
  - further, the corresponding $\pi$ is a stationary distribution
    \[ (\pi^T P)_x = \sum_y \pi_y P_{yx} = \sum_y \pi_x P_{xy} = \pi_x \]
- the strategy is to construct a Markov chain $P$ such that it is reversible, so that we can apply spectral analysis techniques, and has the desired stationary distribution $\pi_x = \mu(x)$
Metropolis-Hastings algorithm

- start with a candidate transition matrix $K$, which we will modify to create $P$
- to ensure unique stationary distribution, it is sufficient to have
  - $K_{xx} > 0$ for all $x \in \mathcal{X}^n$, and [aperiodic]
  - the undirected graph $G(K) = (\mathcal{X}^n, E(K))$ is connected, where $E(K) \equiv \{(x, y) : K_{xy}K_{yx} > 0\}$ [irreducible]
- we want the transition matrix to satisfy the detailed balance equation with $\mu$, but instead for each pair $(x, y)$, suppose the following holds without loss of generality, i.e. instead of $\mu(x)K_{xy} = \mu(y)K_{yx}$ we have
  \[ \mu(x)K_{xy} > \mu(y)K_{yx} \]
- the trick is to remove some ‘probability mass’ from the larger one
  - define $R_{xy} \equiv \min\left(1, \frac{\mu(y)K_{yx}}{\mu(x)K_{xy}}\right)$
  - let
    \[ P_{xy} \equiv \begin{cases} K_{xy}R_{xy} & \text{if } y \neq x \\ 1 - \sum_{y \neq x} P_{xy} & \text{if } y = x \end{cases} \]
  - then, $P$ satisfies the detailed balance equations w.r.t $\mu$, and hence $\mu$ is a stationary distribution of $P$
    \[ \mu(x)K_{xy}R_{xy} = \mu(x)K_{xy} = \mu(x)K_{yx} \frac{\mu(y)K_{yx}}{\mu(y)K_{yx}} = \mu(y)K_{yx}R_{yx} \]
challenges with **Metropolis-Hastings algorithm**

- do we need $\mu$ to construct $P$?
  we only need
  $$\frac{\mu(x)}{\mu(y)} = \prod_{(i,j) \in E} \frac{\psi_{ij}(x_i, x_j)}{\psi_{ij}(y_i, y_j)}$$
  which can be evaluated efficiently. In particular, we do not need to compute the partition function $Z$.

- how do we store $K$ and $P$ with dimensions $|\mathcal{X}|^n \times |\mathcal{X}|^n$?
  consider this construction as describing a sampling process
  - at time $t$ generate a candidate sample $x'$ according to $K(x(t), x')$, which possibly has a simple structure
  - *accept* the candidate state with probability $R_{x(t), x'}$
  - *otherwise* *reject* and keep current state

**Theorem.** Metropolis-Hastings algorithm finds $\ell_1$-projection of $K$ onto
the space of reversible Markov chains with stationary distribution $\mu$

$$P = \min_{Q \in R(\mu)} \sum_x \sum_{y \neq x} |\mu(x) K_{xy} - \mu(x) Q_{xy}|$$
the ‘art’ is in choosing appropriate $K$, since bad choice of $K$ results in a Markov chain with slower convergence

if ‘spread’ is too narrow, we are not exploring

if ‘spread’ is too large, acceptance rate can be low

example.

$$K = \frac{1}{|\mathcal{X}|n} 11^T, \quad R_{xy} = \min \left( 1, \prod_{(i,j) \in E} \frac{\psi_{ij}(y_i, y_j)}{\psi_{ij}(x_i, x_j)} \right)$$

all pairs are sampled with equal probability (as per $K$), but many of them might be unlikely and be rejected with high probability
Gibbs sampling

- **Gibbs sampling** defines $P_{xy}$ as
  
  - at each time step, first select $i \in \{1, \ldots, n\}$ from a uniform distribution
  - set $y_{[n]\setminus i} = x_{[n]\setminus i}^{(t)}$ and sample $y_i$ from $\mu(y_i|x_{[n]\setminus i})$

- for sparse graphs, it is easy to evaluate $\mu(y_i|x_{[n]\setminus i}) \propto \prod_{j \in \partial i} \psi_{ij}(y_i, x_j)$

- thus generated $P$ satisfy the detailed balance with $\mu$
  
  - suppose $x$ and $y$ only differ in exactly one position $i$

  $\mu(x)P_{xy} = \mu(x)\frac{1}{n}\mu(y_i|x_{[n]\setminus i})$

  $= \mu(x_i|x_{[n]\setminus i})\mu(x_{[n]\setminus i})\frac{1}{n}\mu(y_i|x_{[n]\setminus i})$

  $= \underbrace{\mu(x_{[n]\setminus i})\mu(y_i|x_{[n]\setminus i})\frac{1}{n}\mu(x_i|x_{[n]\setminus i})}_{\mu(y)} \cdot \underbrace{\mu(y_i|x_{[n]\setminus i})}_{P_{yx}}$

  - otherwise, $P_{xy} = 0$ if $x$ and $y$ differ in more than one position

- the resulting dynamics of the Markov chain is called **Glauber dynamics**
Gibbs sampling and the analysis of Galuber dynamics is used in:

- Noisy best response in coordination games
  [L. Blume, Games Econ. Behav., 1995]
- Learning Boltzmann machines (contrastive divergence)
  [G. Hinton, Neural Computation, 2002]
- ...
Mixing time

- two common ways to analyze the mixing time of a (reversible) Markov chain is spectral analysis and coupling.

Define. \( \epsilon \)-mixing time of \( P \) is the smallest time such that for all \( t > T_{\text{mix}}(\epsilon) \):

\[
|((p^{(0)})^T P^t - \pi^T)|_{\text{TV}} \leq \epsilon
\]

for any initial distribution \( p^{(0)} \), where \( |x - y|_{\text{TV}} = \sum_i |x_i - y_i| \) is the total variation distance.

Theorem. we can show that \( |((p^{(0)})^T P^t - \pi^T)|_{\text{TV}} \leq |\lambda_2|^t \left( \frac{2}{\sqrt{\pi_{\text{min}}}} \right) \),

where \( |\lambda_2| < 1 \) is the second largest eigenvalue of \( P \).

This implies

\[
T_{\text{mix}}(\epsilon) \leq \frac{\log \frac{2}{\epsilon \sqrt{\pi_{\text{min}}}}}{\log(1/|\lambda_2|)} \leq \frac{\log \frac{2}{\epsilon \sqrt{\pi_{\text{min}}}}}{1 - |\lambda_2|}
\]

\( \frac{1}{1 - |\lambda_2|} \) is called the relaxation time of a Markov chain.
spectral properties of Markov chains

Property 1. \( \pi P = \pi \) and \( P \mathbb{1} = \mathbb{1} \) corresponding to \( \lambda_1 = 1 \)

Property 2. \( \pi^T = \pi^T P = \cdots = \pi^T P^t \)

spectral properties of reversible Markov chains

Property 3. \( P = \Pi^{-1/2} S \Pi^{1/2} \) for some symmetric matrix \( S \) and \( \Pi = \text{diag}(\pi) \)

Proof.

Property 4. \( P \) and \( S \) have the same (set of) eigen values

Property 5. \( \lambda_1(S) = 1 \) with \( \begin{bmatrix} \sqrt{\pi_1} \\ \vdots \\ \sqrt{\pi_n} \end{bmatrix} \) as the eigen vector

such that

\[
S = U \Lambda U^T \\
= \begin{bmatrix} \sqrt{\pi_1} \\ \vdots \\ \sqrt{\pi_n} \end{bmatrix} \begin{bmatrix} \sqrt{\pi_1} & \cdots & \sqrt{\pi_n} \end{bmatrix} + \begin{bmatrix} u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} \lambda_2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} u_2^T \\ \vdots \\ u_n^T \end{bmatrix}
\]

Approximate inference by sampling
Proof. of the spectral bound

\[ |(p^{(0)})^T P^t - \pi^T|_{TV} = \sum_i |((p^{(0)})^T P^t - \pi^T)_i| \]

\[ = \sum_i \frac{|((p^{(0)})^T P^t - \pi^T)_i| \pi_i^{1/2}}{\pi_i^{1/2}} \]

\[ \leq \|((p^{(0)})^T P^t - \pi^T)\pi^{-1/2}\| \|\pi^{1/2}\| \quad \text{[Cauchy-Schwarz]} \]

\[ = \|((p^{(0)})^T P^t - \pi^T P^t)\pi^{-1/2}\| \]

\[ = \|((p^{(0)} - \pi)^T \pi^{-1/2} S^t)\| \]

\[ \leq \|(p^{(0)} - \pi)^T \pi^{-1/2}\| \|\lambda_2\|^t \quad \text{[Spectral analysis]} \]

\[ \leq (1 + \frac{1}{\sqrt{\pi_{\min}}}) \|\lambda_2\|^t \quad \text{[Triangular ineq.]} \]

\[ \|((p^{(0)} - \pi)^T \pi^{-1/2}\| \leq \left\{ \begin{array}{l} \|\pi^T \pi^{-1/2}\| \quad \text{[1]} \\ \|p^{(0)}\| \|\pi^{-1/2}\|_2 \quad \leq 1/\sqrt{\pi_{\min}} \end{array} \right. \]

Approximate inference by sampling
\[ \| (p^{(0)} - \pi) \Pi^{-1/2} S^t \| \leq \| (p^{(0)} - \pi) \Pi^{-1/2} \| |\lambda_2|^t \]

1. \( (p^{(0)} - \pi)^T \Pi^{-1/2} \) is orthogonal to the first singular vector of \( S \)
   - recall \( P = \Pi^{-1/2} S \Pi^{1/2} \)
   - largest eigenvalue of \( P \) is one with left and right eigen vectors \( \pi \) and \( 1 \)
   - let \( \pi^{1/2} = \Pi^{1/2} 1 \)
   - \( S \pi^{1/2} = \pi^{1/2} \), since \( S \pi^{1/2} = \Pi^{1/2} P \Pi^{-1/2} \Pi^{1/2} 1 = \Pi^{1/2} 1 \)
   - hence, \( \pi^{1/2} = \Pi^{1/2} 1 \) is the eigenvector corresponding to the largest eigenvalue of \( S \) which is also one
     \[ (p^{(0)} - \pi)^T \Pi^{-1/2} \cdot \Pi^{1/2} 1 = 0 \]

2. if \( a \) is orthogonal to the first singular left vector of \( S \), then
   \[ \| a^T S^t \| \leq \| a \| \sigma_2(S)^t \]
   - eigen value decomposition: \( S = U \Lambda U^T \), where \( U U^T = U^T U = I \)
   - \( S_1 \equiv U_1 \lambda_1 U_1^T \), and \( a^T S^t = a^T (S - S_1)^t \)
   - \( \| a^T S^t \| = \| a^T (S - S_1)^t \| \leq \| a \| \| S - S_1 \|_2^{t} = \lambda_2^t \| a \| \)
the spectral properties of some simple random walks on graphs

- complete graph:

\[
P = \frac{1}{4} \begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
\end{bmatrix}, \text{ with } |\lambda_2| = 0, \ T_{\text{mix}} \propto \frac{1}{\log(1/0)}
\]

- cycle:

\[
P = \frac{1}{2} \begin{bmatrix}
0 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 \\
\end{bmatrix}, \text{ with } |\lambda_2| = 1 - O(1/n^2), \ T_{\text{mix}} \propto d^2
\]

- star:

\[
P = \begin{bmatrix}
0 & 1/4 & 1/4 & 1/4 & 1/4 \\
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
\end{bmatrix}, \text{ with } \lambda_2 = -1, \ T_{\text{mix}} = \infty
\]
Bounding mixing time via conductance

- spectral analysis, and in particular the second largest eigen value of $P$, gives a means to bound the mixing time
- however, computing the spectral gap can be challenging
- **Cheeger’s inequality** provides a bound on the spectral gap:

\[
\frac{1}{1 - \lambda_2} \leq \frac{2}{\Phi^2}
\]

where conductance $\Phi$ of $P$ is defined as

\[
\Phi \triangleq \min_{S \subset \mathcal{X}} \frac{\sum_{x \in S, y \in S^c} \pi_x P_{xy}}{\pi(S)\pi(S^c)}
\]

- direct computation of $\Phi$ is possible in some cases

\[
T_{\text{mix}}(\epsilon) \leq \frac{2 \log \frac{2}{\epsilon \sqrt{\pi_{\min}}} \Phi^2}{\Phi^2}
\]

Approximate inference by sampling
Bounding mixing time via coupling

- **Define.** A coupling of two random variables $X$ and $Y$ with distributions $\mu_X(x)$ and $\mu_Y(y)$ is a construction of a joint probability distribution over $(X, Y)$, i.e. $\mu(x, y)$ such that the marginals are preserved: $\sum_y \mu(x, y) = \mu_X(x)$ and $\sum_x \mu(x, y) = \mu_Y(y)$

- **Example.** Two (marginal) Gaussians $\mu(x) \sim \mathcal{N}(0, 1)$ and $\mu(y) \sim \mathcal{N}(0, 2)$
  - Independent
  - $Y = 2X$
example. two (marginal) Bernoulli \( X \sim \text{Bern}(p) \) and \( Y \sim \text{Bern}(q) \)

* independent
* construction from \( U[0, 1] \)

how closely can we couple \( X \) and \( Y \)?

in other words, what is

\[
\min_{\text{coupling of } \mu_X, \mu_Y} \mathbb{P}(X \neq Y)
\]
**Coupling lemma.** for two (continuous or discrete) random variables $X$ and $Y$ in the same domain,

$$|\mu_X - \mu_Y|_{TV} = \min_{\text{couplings of } \mu_X, \mu_Y} \mathbb{P}(X \neq Y)$$

**proof.**

$$\mathbb{P}(X \neq Y) = 1 - \sum_x \mu_{X,Y}(x, x)$$

$$\geq \sum_x \left\{ \mu_X(x) - \min\{\mu_X(x), \mu_Y(x)\} \right\}$$

$$= \sum_x \max\{0, \mu_X(x) - \mu_Y(x)\}$$

$$= \frac{1}{2} \sum_x |\mu_X(x) - \mu_Y(x)|$$

Further, exists $\mu(x, y)$ such that $\mu(x, x) = \min\{\mu_1(x), \mu_2(x)\}$, and

$$\mu(x, y) = \frac{(\mu_X(x) - \mu(x, x))(\mu_Y(y) - \mu(y, y))}{1 - \sum_x \mu(x, x)}$$
example of an optimal coupling

\[ X = \begin{cases} 
0 & \text{w.p. } p \\
1 & \text{w.p. } 1 - p 
\end{cases} \quad Y = \begin{cases} 
0 & \text{w.p. } q \\
1 & \text{w.p. } 1 - q 
\end{cases} \]

need to construct a probability distribution over \( X \) and \( Y \)

\[
\begin{array}{c|c|c}
\min\{p, q\} & \max\{0, p - q\} & p \\
\hline
\max\{0, q - p\} & \min\{1 - p, 1 - q\} & 1 - p \\
\hline
q & 1 - q & \\
\end{array}
\]

this naturally extends to larger alphabet. Equivalently, one could draw \( Z \sim \text{Uniform}[0,1] \), then coupling is nothing but determining intervals in \([0, 1]\) for each output of \( X \) and \( Y \). For example, the optimal coupling is

\[ X = \begin{cases} 
0 & \text{if } Z \in [0, p] \\
1 & \text{otherwise} 
\end{cases} \quad Y = \begin{cases} 
0 & \text{if } Z \in [0, q] \\
1 & \text{otherwise} 
\end{cases} \]

**Corollary of the coupling lemma.** total variation can be upper bounded by any coupling,

\[ |\mu_X - \mu_Y|_{TV} \leq \mathbb{P}(X \neq Y) \]
Coupling for bounding $T_{\text{mix}}$

- let $X_t$ and $Y_t$ be random states after $t$ transitions according to $P$ with initial state $X_0$ and $Y_0$
- **Corollary of the coupling lemma.** for any coupling of $X_t$ and $Y_t$,

\[
|\mu_{X_t} - \mu_{Y_t}|_{\text{TV}} \leq \mathbb{P}(X_t, Y_t)(X_t \neq Y_t)
\]

- **Strategy.** to get a tight bound on the total variation, we need to construct good coupling.

we consider a particular coupling of two Gibbs sampling chains for $x, y \in \{0, 1\}^n$

1. draw uniform $I \in [n]$
2. draw $x'_I$ from $\mu(x'_I|x_{\partial I})$ and $y'_I$ from $\mu(y'_I|y_{\partial I})$ using the optimal coupling
Bounding \( \mathbb{P}(X_t, Y_t)(X_t \neq Y_t) \) by \textbf{path coupling}

[R. Bubley and M. Dyer, FOCS 1997]

- **Define.** \( D(x, y) \) is the minimal number of allowed moves in the transition matrix \( P \) to go from \( x \) to \( y \) (e.g. Hamming distance for Gibbs sampling)

- **Idea.** if we can construct a coupling such that

\[
\mathbb{E}[D(x_{t+1}, y_{t+1})|x_t, y_t] \leq \alpha D(x_t, y_t) \tag{1}
\]

for some \( 0 < \alpha < 1 \), then

\[
|\mu_{X_t} - \mu_{Y_t}|_{TV} \leq \mathbb{P}(X_t \neq Y_t) \leq \mathbb{E}[D(x_t, y_t)] \leq \alpha^t D(x_0, y_0) \leq \frac{\log \frac{D(x_0, y_0)}{\epsilon}}{\log \frac{1}{\alpha}}
\]

\[
\Rightarrow T_{\text{mix}}(\epsilon) \leq \frac{\log \frac{D(x_0, y_0)}{\epsilon}}{\log \frac{1}{\alpha}}
\]
Path coupling for Gibbs sampling

two Markov chains start at a distance as measured by $D(x^{(1,0)}, x^{(2,0)})$, and with the right coupling two sample paths eventually converge and follow the same sample path after some (random) time.

- **Path coupling.** to prove that $\mathbb{E}[D(x_{t+1}, y_{t+1})|x_t, y_t] \leq \alpha D(x_t, y_t)$ it is sufficient to prove it for $x_t$ and $y_t$ that only differ in one vertex.
Claim. If $\mathbb{E}[D(\hat{x}, \hat{y})|D(x, y) = 1] \leq \alpha$ then Eq. (1) follows.

Proof sketch. consider a minimum length path from $x$ to $y$:

$$p = (x, p_1, \ldots, p_{D(x,y)-1}, y)$$

which are, after one step of the Markov chain, mapped to

$$(\hat{x}, \hat{p}_1, \ldots, \hat{p}_{D(x,y)-1}, \hat{y})$$

by triangular inequality,

$$\mathbb{E}[D(\hat{x}, \hat{y})|x, y] \leq \mathbb{E}[D(\hat{x}, \hat{p}_1) + D(\hat{p}_1, \hat{p}_2) + \cdots + D(\hat{p}_{D(x,y)-1}, \hat{y})] \leq \alpha \mathbb{E}[D(x, y)]$$

Approximate inference by sampling
for some graphical models, path coupling constant $\alpha$ can be bounded, e.g.

$$
\mu(x) = \frac{1}{Z} \exp \left\{ \sum_{i,j \in E} \theta_{ij} x_i x_j \right\}
$$

- **Claim.** for Gibbs sampling on Ising models,

$$
\mathbb{E}[D(x_{t+1}, y_{t+1})|D(x_t, y_t) = 1] \leq 1 - \frac{1 - d_{\text{max}} \tanh(\theta_{\text{max}})}{n}
$$

- hence, Gibbs sampling mixes fast when $d_{\text{max}} \tanh(\theta_{\text{max}}) < 1$

- **Step 1. Construction of a good coupling.** to prove the claim, we consider a particular coupling of two Gibbs sampling chains

  1. draw uniform $I \in [n]$
  2. draw $x'_I$ from $\mu(x'_I|x_\partial I)$ and $y'_I$ from $\mu(y'_I|y_\partial I)$ coupled in the following way
    2-1. draw a random $Z \sim \text{Uniform}[0, 1]$
    2-2. let

$$
x'_I = \begin{cases} 
  +1 & \text{if } Z \in [0, \mu(x'_I = +1|x_\partial I)] \\
-1 & \text{otherwise}
\end{cases}
$$

$$
y'_I = \begin{cases} 
  +1 & \text{if } Z \in [0, \mu(y'_I = +1|y_\partial I)] \\
-1 & \text{otherwise}
\end{cases}
$$

Approximate inference by sampling
Step 2. Analysis of the distance. We are left to show that

$$\mathbb{E}[D(x', y') | x \text{ and } y \text{ differ only at } i] \leq 1 + \frac{1}{n} \left\{ -1 + \sum_{j \in \partial i} |\tanh(\theta_{ij})| \right\}$$

Case 1. If $I = i$, $D(x', y')$ reduces to 0

$$\mathbb{E}[D(x', y') | x \text{ and } y \text{ differ only at } i, I = i] = 0$$

This happens with probability $1/n$
case 2. if $I \notin \{i\} \cup \partial i$, $D(x', y')$ remains at 1

$$\mathbb{E}[D(x', y') | x \text{ and } y \text{ differ only at } i, I \notin \{i\} \cup \partial i] = 1$$

this happens with probability $1 - \frac{1+|\partial i|}{n}$
case 3. if \( I \in \partial \cdot \), \( D(x', y') \) can increase with probability

\[
|\mu(x_I = + | x_{\partial I}) - \mu(y_I = + | y_{\partial I})| = \\
\left| \frac{A^{(+)\psi_{iI}(+, +)} - A^{(+)\psi_{iI}(-, +)}}{A^{(+)\psi_{iI}(+, +)} + A^{(-)\psi_{iI}(+, -)}} \right|
\]

where \( A^{(+)} = \prod_{j \in \partial I \setminus \{i\}} \psi_{ji}(x_j, +) \), and \( A^{(-)} = \prod_{j \in \partial I \setminus \{i\}} \psi_{ji}(x_j, -) \)
Claim. for Ising model with $\psi(x_i, x_I) = e^{\theta_{iI} x_i x_I}$, the probability is bounded by $| \tanh(\theta_{iI}) |$

**Proof.** in the case of $\theta_{iI} > 0$, we want to show that

\[
\frac{A(+) e^{\theta_{iI}}}{A(+) e^{\theta_{iI}} + A(-) e^{-\theta_{iI}}} - \frac{A(+) e^{-\theta_{iI}}}{A(+) e^{-\theta_{iI}} + A(-) e^{\theta_{iI}}} = \frac{A(+) A(-) (e^{2\theta_{iI}} - e^{-2\theta_{iI}})}{(A(+))^2 + (A(-))^2 + A(+) A(-) (e^{2\theta_{iI}} + e^{-2\theta_{iI}})} = \frac{(e^{2\theta_{iI}} - e^{-2\theta_{iI}})}{(A(+))^2 + (A(-))^2 + (e^{2\theta_{iI}} + e^{-2\theta_{iI}})} = \tanh(\theta_{iI})
\]

where we used the fact that $A(+) A(-) = 1$ and it also follows that $(A(+))^2 + (A(-))^2 \geq 2$. 
For Ising model,
\[ \mu_{G, \theta}(x) = \frac{1}{Z_G(\theta)} \exp \left\{ \theta \sum_{(i,j) \in E} x_i x_j \right\}. \]

we showed that Gibbs sampling mixed fast if \( \tanh(\theta_{am}) \deg_{\text{max}} < 1 \).

Experiment with \( G \) uniformly random with \( N \) vertices and \( 2N \) edges (average degree 4).

\[
C(t) = \frac{1}{|V|} \sum_{i \in V} x_i(0)x_i(t),
\]

\[
t = \frac{1}{|V|} [\text{number of steps}]
\]
**Theorem.** [Mossel, Sly, 2010] Assume $\theta_{ij} = \theta > 0$. Then the Glauber Markov chain mixes rapidly provided

$$(k - 1) \tanh(\theta) < 1$$


$$(k - 1) \tanh(\theta) > 1$$

then there exists a sequence of $k$-regular graphs $G_n = ([n], E_n)$ for which the Glauber Markov chain mixes in time $\exp\{\Theta(n)\}$. 
Is \((k - 1) \tanh(\theta) = 1\) fundamental?

- Recall computation tree \(T^{(t,i)}\) is formed from a graphical model by considering a root node \(x_i\) and a tree of all non-backtracking (non-reversing) paths for length \(t\).

- **Proposition.** Let \(\nu_i(x_i)\) be the BP estimate after \(t\) iterations, \(\nu_{i \rightarrow j}^{(t)}(x_i)\) be the BP message, and \(\mu^{(t,i)}(x_i)\) be the marginal of the root \(x_i\) on the computation tree \(T^{(t,i)}\), with some boundary conditions to be specified with the model. Then,

\[
\nu_i^{(t_0 + t_1)}(x_i) = \mu^{(t_1,i)}(x_i)
\]

with the boundary condition of the computation tree set to \(\nu_j^{(t_0)}(x_j)\) for a node \(x_j\) in the boundary with parent node \(x_k\).

- **Proof.** proof by induction.

- **Corollary.** Let \(\partial T^{(t,i)}\) denote the boundary nodes of the tree. If

\[
\max_{x_{\partial T^{(t,i)}}, x'_{\partial T^{(t,i)}}} \left| \mu^{(t,i)}(x_i | x_{\partial T^{(t,i)}}) - \mu^{(t,i)}(x_i | x'_{\partial T^{(t,i)}}) \right|_{TV} \leq \delta(t), \quad (2)
\]

then, for all \(t_1, t_2 \geq t\),

\[
\left| \nu_i^{(t_1)}(x_i) - \nu_i^{(t_2)}(x_i) \right| \leq \delta(t).
\]

In particular, if \(\delta(t) \to 0\) as \(t\) grows, then BP converges.
Define. $B_i(t)$ as the subgraph of $G$ that includes all nodes at most distance $t$ from node $x_i$.

Corollary. If $B_i(t)$ is a tree, and Equation (2) holds, then

$$\left| \mu(x_i) - \nu_i^{(t)}(x_i) \right| \leq \delta(t).$$

In particular, if $g$ if the girth (the length of the shortest cycle) of $G$, then we have

$$\left| \mu(x_i) - \nu_i(x_i) \right| \leq \delta((g - 1)/2)$$

Proof. observe that $\mu(x_i) = \sum_{x(t)} \mu(x_i | x^{(t)}) \mu(x^{(t)})$ where $x^{(t)}$ are the nodes at distance $t$ from $x_i$.

the condition (2) is known as correlation decay and we established that correlation decay implies convergence of BP in general graphs and correctness of BP on locally tree-like graphs, but checking condition (2) can be challenging.
Dobrushin’s uniqueness criterion

- Dobrushin’s criterion measures the strengths of interactions, and provides a sufficient condition for Condition (2).

- **Define.** Influence of \( j \) on \( i \) as

\[
C_{ij} \triangleq \max_{x, x'} \left| \mu(x_i = j | x_{V \setminus i}) - \mu(x_i = j' | x_{V \setminus i}) \right|_{TV}
\]

- \( 0 \leq C_{ij} \leq 1 \)
- \( C_{ij} = 0 \) unless \( (i, j) \in E \)

- **Theorem.** [Dobrushin, 1968] Small influence implies correlation decay. Let

\[
\gamma \triangleq \max_{i \in V} \left\{ \sum_{j \in \partial i} C_{ij} \right\}.
\]

Then,

\[
\max_{x, x'} \left| \mu(x_i = j | x_{V \setminus B_i(t)}) - \mu(x_i = j' | x_{V \setminus B_i(t)}) \right|_{TV} \leq \frac{\gamma^t}{1 - \gamma}
\]
Proof strategy

- bound influence on vertex $j$ from those outside a ball of radius $\ell$
• assume neighborhood of \( j \) is a \( k \)-regular tree

• a graphical model satisfies **uniqueness condition** if

\[
\sup_{y_{\partial B}, z_{\partial B}} \left| \mu(x_j | x_{\partial B} = y_{\partial B}) - \mu(x_j | x_{\partial B} = z_{\partial B}) \right| \leq \varepsilon(\ell) \downarrow 0
\]

[In reality slightly stronger condition needed for proof]
Checking for uniqueness

\[ \text{Uniqueness: } h_{i \rightarrow j} \equiv \operatorname{atanh} \mathbb{E}_{\mu, T(i \rightarrow j)} \{ x_i \}. \]

Uniqueness: \( h_{i \rightarrow j} \) asymptotically independent of boundary condition
Checking for uniqueness

Exercise:

\[ h_{i \rightarrow j} = \theta_i + \sum_{v \in \text{children}(i)} \text{atanh}\{ \text{tanh} \theta_{iv} \text{tanh} h_{v \rightarrow i} \} . \]

- \( \theta_{ij} = \beta, \theta_i = 0, \)
- \( x_{\partial B(j, \ell)} = +1, \ x_{\partial B(j, \ell)} = -1 \) (monotonicity)

\[ h_{\ell+1} = (k - 1) \text{atanh}\{ \text{tanh} \beta \text{tanh} h_{\ell} \} . \]
A one-dimensional recursion

\[ h_{\ell+1} \]

\[ (k - 1) \tanh \beta < 1 \quad (k - 1) \tanh \beta = 1 \quad (k - 1) \tanh \beta > 1 \]

- who cares about regular trees?
- regular trees are the worst case for decay of correlations
What about the lower bound?

**Theorem (Gerschenfeld, Montanari, FOCS 2007)**

Assume \((k - 1) \tanh \beta > 1\).

Then there exists a sequence of \(k\)-regular graphs \(G_n = (V_n = [n], E_n)\) for which the Glauber Markov chain mixes in time \(\exp\{\Theta(n)\}\).

**Proof.**

Take \(G_n\) a uniformly random \(k\)-regular graph and prove that w.h.p.

\[
P_{\mu}\left\{ \sum_{i \in V} x_i = 0 \right\} = e^{-\Theta(n)},
\]

\[
P_{\mu}\left\{ \sum_{i \in V} x_i > 0 \right\} = P_{\mu}\left\{ \sum_{i \in V} x_i < 0 \right\} = \frac{1}{2} - e^{-\Theta(n)}.\]

Bottleneck!
Are random graphs a curiosity?

No! Used as gadgets in

- Sly, *Computational transition at the uniqueness threshold*, 2010
- ...

**Theorem**

For antiferromagnetic Ising models $\theta_{ij} = -\theta < 0$, $\theta_i = 0$, the partition function cannot be approximated unless $\text{RP} = \text{NP}$. 

Approximate inference by sampling
$Q_n(\beta) \equiv \mathbb{P}_\mu \{ \sum_{i \in V} x_i = 0 \}$

$$\mu_{G,\beta}(x) = \frac{1}{Z_{G}(\beta)} \exp \left\{ \beta \sum_{(i,j) \in E} x_i x_j \right\}$$

$$Q_n(\beta) = \frac{Z^*_G(\beta)}{Z_G(\beta)}, \quad Z^*_G(\beta) \equiv \sum_{x: \langle x, 1 \rangle = 0} e^\beta \sum_{(i,j) \in E} x_i x_j$$

- Upper bound $Z^*_G(\beta)$ by $n^{10} \mathbb{E}_G Z^*_G(\beta)$.

- Lower bound $Z_G(\beta)$ by …
Estimating $Z_G$


Let $\{G_n = (V_n, E_n)\}_{n \geq 1}$ be a sequence of graphs that (i) Is uniformly sparse; (ii) Converges locally to a unimodular Galton-Watson tree. Let $Z_n(\beta, B)$ be the Ising model partition function with $\theta_{ij} = \beta$, $\theta_i = B$. Then

$$\lim_{n \to \infty} \frac{1}{n} \log Z_n(\beta, B) = \text{[explicit expression]} = \text{[Bethe free energy]}$$