* Recap

$$\underset{\text{Bethe free energy}}{\max_{b\in LOC(T)} G_T(b)} \le \underset{\text{Gibbs free energy}}{\max_{b\in \Delta_{|X|^{|V|}}} G(b)} \le \underset{\text{Tree reweighted BP}}{\sum_{T\in\mathcal{T}} \max_{b\in LOC(T)} G_T(b)}$$

Tight if the Graphical model is a Tree

$$\equiv$$

Bethe free energy

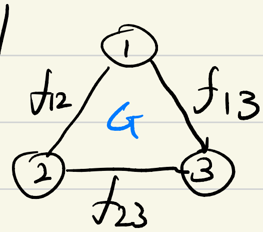
$$\max_{b\in LOC(G)} G(b)$$

$\downarrow$

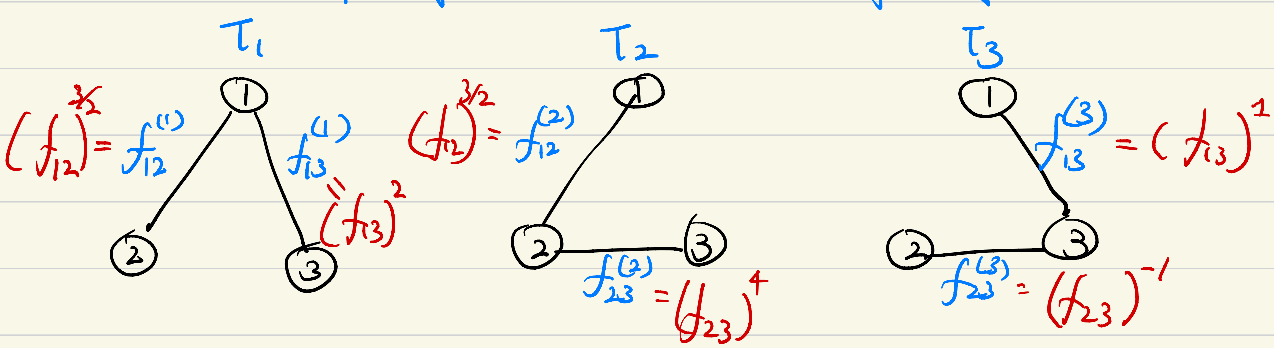
B.P.

Tree reweighted belief propagation [more details on slides
"A new class of upper bounds on the log partition function." 2005 Wainwright, Jaakkola, Willsky

example> given Graphical Model



Consider all spanning trees on G, and assign weights on those trees.



Rule #1: $\sum_K c_K = 1$

$c_1 = \frac{1}{3}$ $\quad c_2 = \frac{1}{3}$ $\quad c_3 = \frac{1}{3}$

Rule #2: for each $(i,j)\in E$ $\quad \prod_K \left(f_{ij}^{(K)}\right)^{c_K} = f_{ij}$

Def. Tree reweighted belief propagation.

given $\mathcal{T} = \{T_k\}_{k=1}^{K}$ set of spanning trees, and corresponding weights $\{c_k\}$ s.t.

$$\sum c_k = 1$$

$$\log f_{ij}(x_i, x_j) = \sum_k c_k \log f_{ij}^{(k)}(x_i, x_j)$$

then the energy term in Gibbs free energy decomposes as

$$\mathbb{E}_b\Big[-\sum_{ij \in E} \log f_{ij}(x_i, x_j)\Big] = \mathbb{E}_b\Big[-\sum_{(i,j)\in E} \sum_k c_k \log f_{ij}^{(k)}(x_i, x_j)\Big]$$

$$= \sum_k c_k \underbrace{\mathbb{E}_b\Big[-\sum_{(i,j)\in E} \log f_{ij}^{(k)}(x_i, x_j)\Big]}_{\text{energy of a one tree model.}}$$

* Claim: $\log Z \leq \sum_k c_k \log Z_k$

easy to compute with B.P. as it is a tree.

proof:

$$\log Z = \max_{b \in \Delta_{|X|^n - 1}} \mathbb{E}_b\Big[\sum_{(i,j)\in E} \log f_{ij}(x_i, x_j)\Big] + \text{Entropy}(b)$$

$\sum c_k = 1$ $\rightarrow$ $= \max_{b \in \Delta_{|X|^n - 1}} \sum_k c_k \Big\{\mathbb{E}_b\Big[\sum_{(i,j)\in E} \log f_{ij}^{(k)}(x_i, x_j)\Big] + \text{Entropy}(b)\Big\}$

exchange sum & max $\rightarrow$ $\leq \sum_k c_k \max_{b \in \Delta_{|X|^n - 1}} \Big\{\mathbb{E}_b\Big[\sum_{(i,j)\in E} \log f_{ij}^{(k)}(x_i, x_j)\Big] + \text{Entropy}(b)\Big\}$

Graphical model is on Tree $\rightarrow$ $= \sum_k c_k \underbrace{\max_{b \in LOC(T_k)} \Big\{\mathbb{E}_b\Big[\sum_{(i,j)\in E} \log f_{ij}^{(k)}(x_i, x_j)\Big] + \text{Entropy}(b)\Big\}}_{\text{this can be solved exactly with B.P.}}$

- Caveat
  - we want to select c_k, $f_{ij}^{(k)}$ that minimize $\sum c_k \log Z_k$
  - # of spanning trees can explode

* To solve inference problems $P(x_i)$

Variational Methods	Sampling
↓	↓
Belief Propagation	Gibbs Sampling
↓	↓
· Deterministic	· Randomized
· fast	· slower
· approximation	· exact in the limit $N \to \infty$, but difficult to decide when to stop.

Sampling: $\hat{P}(x_i = a) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(x_i = a)$

Def. Markov Chain Monte Carlo methods.

- construct a Markov chain $X^{(t)} \in \mathcal{X}^n \to X^{(t+1)} \in \mathcal{X}^n$ with transition matrix Q whose stationary distribution $P(x)$
- start with an arbitrary realization $X^{(0)}$ and run the Markov Chain until it converges close to its stationary distribution
- Repeat.

Q_1: How do we construct Q? → Metropolis-Hastings algorithm → Gibbs sampling

Q_2: How long does it take for the markov chain to converge?

→ spectral analysis.

→ path coupling.

Strategy >

given a graphical model $G \Rightarrow P(x) = \frac{1}{Z} \prod f_{ij}(x_i, x_j)$.

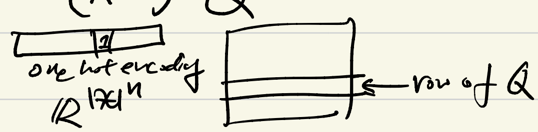
state $X^{(t)} \in \mathcal{X}^n$: realization

initial state $X^{(0)}$

Construct transition matrix $Q \in \mathbb{R}^{|\mathcal{X}|^n \times |\mathcal{X}|^n}$

we will not write it in a memory, but conceptually define it + sparse

Repeat: sample: $X^{(t+1)} \sim (X^{(t)})^T \cdot Q$



eventually $X^{(t)} \sim \pi$: stationary distribution of Q.

$\pi = P(x)$

Def. time-homogeneous finite-state Markov Chains.

Markov chain
- State space $\mathcal{X}^n$
- Transition matrix $Q \in \mathbb{R}^{|\mathcal{X}|^n \times |\mathcal{X}|^n}$

$$Q_{xy} = \mathbb{P}(X_{t+1} = y \mid X_t = x)$$

Def. stationary distribution π of M.C.

$$\pi^T \cdot Q = \pi^T \qquad \pi = \Big[\ \Big]\Big\} |\mathcal{X}|^n$$

- might not be unique
- might not exist.

we do not get into detailed conditions on Q that ensures existence & uniqueness, but construct Q s.t. it is U&E.

Def. a Markov chain is Reversible if $\exists \pi$ s.t detailed balance equation is satisfied

$$\pi_x \cdot Q_{xy} = \pi_y Q_{yx} \quad \text{for all } x, y. \qquad (*)$$

Claim. π satisfying $(*)$ is a stationary distribution of Q.

proof>

$$(\pi^T Q)_y = \sum_x \pi_x Q_{xy} \underset{(*)}{=} \sum_x \pi_y Q_{yx} \underset{Q \text{ is stochastic, i.e. row-sum to one}}{=} \pi_y.$$

$$\Rightarrow \pi^T Q = \pi^T$$

we will construct a Markov Chain Q that is reversible.
- we can use $(*)$ to ensure $\pi = P(x)$
- spectral analysis can be applied.

Def. Metropolis-Hastings Algorithm

- Start with a candidate transition matrix K
- To ensure unique stationary distribution, it is sufficient to have
  - $K_{xx} > 0 \quad , \forall x \in \mathcal{X}^n$ [aperiodic]
  - for any $x, y \in \mathcal{X}^n$, $\exists$ a path $(x_1 = x, x_2, \cdots, x_m = y)$ of positive probability transitions [irreducibility]

$$K_{x_i x_{i+1}} > 0, \quad \forall i \in \{1, \cdots, m-1\}$$

what we want

$$(*) \quad Q_{xy} \cdot P(x) = Q_{yx} P(y)$$

what we have

$$K_{xy} P(x) \underset{\text{w.l.o.g}}{>} K_{yx} P(y)$$

the main trick is to remove some "probability mass" from the larger one.

Define: $R_{xy} \triangleq \min\left\{ 1 , \frac{P(y) K_{yx}}{P(x) K_{xy}} \right\}$ for each x, y

$$Q_{xy} = K_{xy} \cdot R_{xy}$$

$$Q_{xx} \underset{\substack{\text{s.t.} \\ \text{probability sums to one}}}{=} 1 - \sum_{y \neq x} Q_{xy}$$

with this rejection sampling $(1 - R_{xy})$

claim: $(Q, P(x))$ satisfy $(*)$.

proof> $p(x) \cdot Q_{xy} = P(x) \cdot R_{xy} \cdot K_{xy}$

$= P(y) K_{yx}$ (if $R_{xy} < 1$ & $R_{yx} = 1$)

$= P(y) R_{yx} \cdot K_{yx}$.

* Remark: as $P(x)$ is only used in $R_{xy} = \min\left\{1, \frac{P(y)K_{yx}}{P(x)K_{xy}}\right\}$

we only need $\frac{P(y)}{P(x)} = \prod_{(i,j)\in E} \frac{f_{ij}(y_i,y_j)}{f_{ij}(x_i,x_j)}$ ← takes $O(|E|)$ computation

does not require Z.

* Remark: Do we need to store K and $Q \in \mathbb{R}^{|\chi|^n \times |\chi|^n}$

- We can choose K to be simple, such as $K = \frac{1}{|\chi|^n} \cdot \mathbb{1}\mathbb{1}^T$

Step1. - At time t first generate candidate $x^{t+1/2}$ from $K(x^t, x^{t+1/2})$

Step2. - Accept $x^{t+1/2}$ with probability $R_{x^t, x^{t+1/2}}$: $x^{t+1} = x^{t+1/2}$

Step3. - Otherwise reject and keep current state : $x^{t+1} = x^t$

*Theorem: Metropolis-Hastings Algorithm finds l_1-projection of K onto the space of reversible Markov Chains with stationary distribution $P(x)$.

$$Q_{MH} = \underset{\tilde{Q} \text{ s.t. } P^T\tilde{Q} = P^T}{\arg\min} \sum_x \sum_{y \neq x} \left| P(x) \cdot K_{xy} - P(x)\tilde{Q}_{xy} \right|$$

*Are we done? - the art is in choosing K.

if the spread of K is too large, then acceptance rate is low

if the spread of K is too narrow, then mixing time can be large

example > $K = \frac{1}{|\chi|^n} \mathbb{1}\mathbb{1}^T$, $R_{xy} = \min\left(1, \prod_{(i,j)\in E} \frac{f_{ij}(y_i,y_j)}{f_{ij}(x_i,x_j)}\right)$

- all pairs first sampled with equal probability (as per K)
- but many candidates might be unlikely and be rejected.

* Def. Gibbs Sampling.

$$Q\begin{cases} \text{Step 1. sample } i \in \{1, \cdots, n\} \text{ uniformly at random.} \\ \text{Step 2. set } \underset{[n]\setminus\{i\}}{\overset{}{y_{-i}}} = x_{-i}^{(t)} \\ \text{Step 3. sample } y_i \text{ from } P(y_i \mid x_{-i}^{(t)}) \end{cases}$$

* Remark. $P(y_i \mid x_{-i}^{(t)}) \propto \prod_{j \in \partial i} f_{ij}(y_i, x_j^{(t)})$ is efficient

* claim. $(Q, P(x))$ satisfy $(*)$.

proof. for y that differ at only i-th coordinate from x,

$$P(x) \cdot Q_{xy} = P(x) \cdot \frac{1}{n} \cdot P(y_i \mid x_{-i})$$

$$\text{Bayes} \rightarrow = P(x_i \mid x_{-i}) P(x_{-i}) \cdot \frac{1}{n} P(y_i \mid x_{-i})$$

$$= \underbrace{P(x_i \mid x_{-i}) \frac{1}{n}}_{Q_{yx}} \underbrace{P(x_{-i}) P(y_i \mid x_{-i})}_{P(y)}$$

Otherwise $Q_{xy} = 0$ if x & y differ more than one coordinate.

* the resulting dynamics of the Markov chain is called Glauber Dynamics.