* Equivalence of inference tasks.
  - Suppose we have a black-box tool that computes the partition function Z_G of any graphical model G.
  then we can compute marginal $\mathbb{P}(X_i=0) \propto Z_{G(X_i=0)}$
  $\triangleq$ is the graphical model conditioned on x_i being 0.

* We will focus on the task of computing Z given a graphical model.
  for
$$P(x) = \frac{1}{Z} \cdot \underbrace{\prod_{(i,j)\in E} f_{ij}(x_i, x_j)}_{f_{total}(x)}$$
  - We are given $f_{total}(x)$ but not Z.
  - We focus on computing (approximately) the log partition function
$$\Phi \triangleq \log Z = \log\left\{\sum_{x\in\mathcal{X}^n} f_{total}(x)\right\}$$

  - Def. <u>Variational characterization</u> of log partition function.
  is to represent as a solution of an optimization problem.
$$\Phi = \sup_{b:\text{ distribution over } \mathcal{X}^n \text{ PMF}} G_f(b)$$
  - Def. Gibbs free energy $G_{f_{total}}(b)$.
$$G_{f_{total}}(b) = \sum_{x\in\mathcal{X}^n} b(x)\cdot \log f_{total}(x) - \sum_{x\in\mathcal{X}^n} b(x)\log b(x)$$
$$= \underbrace{-\mathbb{E}_b[\underbrace{-\log f_{total}(X)}_{H(x)\text{ energy of state } x}]}_{\text{expected energy}} + \underbrace{\mathbb{E}_b[-\log b(x)]}_{\text{entropy of } b}$$

* $\mathbb{P}(x) = \frac{1}{Z}\exp(-\underbrace{H(x)}_{\text{energy}})$: low energy state is more likely.

claim: $G_{f_{\text{teel}}}(b)$ is strictly Concave

$\sup_b G_{f_{\text{teel}}}(b) = \Phi$

$P(x) = \arg\max_b G_{f_{\text{teel}}}(b)$.

Interpretation: $b^*(x) = P(x)$ minimizes average energy (= more likely for each state) while maximizing entropy. (= more # of states)

Proof:
$$G_{f_{\text{teel}}}(b) = \sum_x b(x) \log f_{\text{teel}}(x) - \sum_x b(x) \log b(x)$$
$$= \sum_x b(x) \{ \log Z + \log \underbrace{\tfrac{1}{Z} f_{\text{teel}}(x)}_{P(x)} \} - \sum_x b(x) \log b(x)$$
$$= \log Z + \sum_x \left(b(x) \log P(x) - b(x) \log b(x) \right)$$
$$= \Phi - \underbrace{D_{KL}(b \,\|\, P)}_{\text{Kullback-Leibler Divergence}}.$$

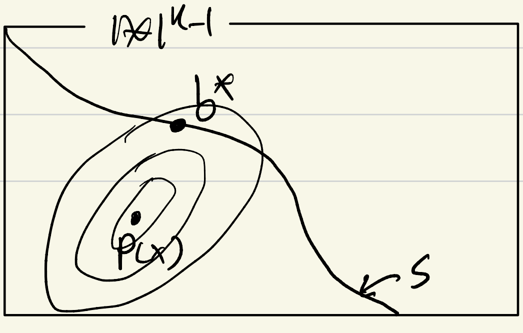
$\leq \Phi$ & equal iff $b = P$.

- From information theory, we know
  - $D_{KL}(b\|p) \geq 0$
  - $D_{KL}(b\|p) = 0$ iff $b = p$.
  - $D_{KL}(b\|p)$ is convex in b.

then, $$\Phi = \sup_{b: \Delta_{|\mathcal{X}|^n - 1}} G_{f_{\text{teel}}}(b)$$

is a concave maximization. but in $|\mathcal{X}|^n - 1$ dimensions.

- strategy



* we instead search over a small dimensional space, and find approximate solution.

$$\Phi \geq \sup_{b \in S} G_{f_{\text{teel}}}(b)$$

this provides a lower bounds

* Def. Naive Mean Field approach.

Consider Naive Mean Field factorization

$$S_{MF} = \{ b \in \Delta_{|\mathcal{X}|^n - 1} : b(x) = b_1(x_1) \times b_2(x_2) \times \cdots \times b_n(x_n) \}$$

with a slight abuse of notations, let $b = b_1 \times b_2 \times \cdots \times b_n$,
and plug in-to Gibbs free energy to get

$$\mathbb{F}_{MF}(b) = G_{free}(b_1 \times b_2 \times \cdots \times b_n)$$

$$= \sum_{(i,j) \in E} \sum_{x_i, x_j} b_i(x_i) b_j(x_j) \log \psi_{ij}(x_i, x_j) - \sum_{i \in V, x_i} b_i(x_i) \log b_i(x_i)$$

Mean Field Variational Inference problem.

$$\max_{b \in S_{MF}} \mathbb{F}_{MF}(b)$$

$$\text{s.t.} \quad b_i(x_i) \geq 0 \quad \forall i, x_i$$

$$\sum_{x_i} b_i(x_i) = 1. \quad \forall i$$

* $b_i(\cdot)$'s play the role of approximate marginal distribution
$b_i(x_i) \cong P(x_i)$

* Dimension $b \in \mathbb{R}^{n(|\mathcal{X}|-1)}$, $b_i \in \Delta_{|\mathcal{X}|-1}$

* but now the objective is no longer concave: $b_i(x_i) \cdot b_j(x_j)$: bilinear

* we look for a local maxima

Stationary Point of Naive Mean Field is characterized by Lagrangian

$$\mathcal{L}(b, \lambda) = \mathbb{F}_{MF}(b) - \sum_{i \in V} \lambda_i \{ \sum_{x_i \in \mathcal{X}} b_i(x_i) - 1 \}$$

(we don't include positivity as they appear in $\log(\cdot)$ in the objective)

$$= \frac{1}{2} \sum_{i \in V, x_i \in \mathcal{X}} b_i(x_i) \Big\{ \sum_{j \in \partial i, x_j \in \mathcal{X}} b_j(x_j) \cdot \log \psi_{ij}(x_i, x_j) \Big\}$$

$$- \sum_{i \in V, x_i \in \mathcal{X}} b_i(x_i) \log b_i(x_i) - \sum_{i \in V} \lambda_i \sum_{x_i \in \mathcal{X}} (b_i(x_i) - 1)$$

taking the derivative, w.r.t $b_i(x_i)$

$$\frac{\partial L(b,\lambda)}{\partial b_i(x_i)} = \sum_{j \in \partial i} \sum_{x_j \in \mathcal{X}} b_j(x_j) \log \psi_{ij}(x_i, x_j) - 1 - \log b_i(x_i) - \lambda_i = 0$$

Def. Naive Mean field equation

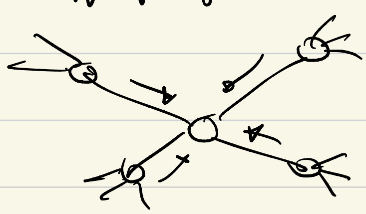
$$b_i(x_i) \propto \exp\left\{ \sum_{j \in \partial i} \sum_{x_j \in \mathcal{X}} \log \psi_{ij}(x_i, x_j)\, b_j(x_j) \right\}$$

is a fixed point of: $b = F_{MF}(b)$

which can be ~~solved~~ searched by

$$b^{(t+1)} = F_{MF}(b^{(t)})$$

One can think of this as having a belief at each node (as opposed to edges) and update by aggregating its neighbors.



similar to Gossip algorithms

* this gives a very poor approximation.

ex) $X_1, X_2 \in \{0,1\}$.

$P(x) = \frac{1}{2} \mathbb{I}(X_1 = X_2)$ vs. $P(x) = \frac{1}{2} \mathbb{I}(X_1 \neq X_2)$

$$\begin{array}{c|cc|c} & & X_2 & \\ X_1 & \frac{1}{2} & 0 & \frac{1}{2} \\ & 0 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} & \end{array} \qquad \begin{array}{cc|cc|c} & & 0 & 1 & \\ X_1 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ & 1 & \frac{1}{2} & 0 & \frac{1}{2} \\ \hline & & \frac{1}{2} & \frac{1}{2} & \end{array}$$

[account for marginal $P(X_i)$
[exact only if $P(x) = P(x_1) P(x_2) \cdots P(x_n)$.

* Bethe free energy

[account for pairwise correlations induced by edges/factors
[exact on distribution P on trees

· Parameters: $b_i(x_i)$: approximates $P(x_i)$
$b_{ij}(x_i, x_j)$: approximates $P(x_i, x_j)$
use $b = \{b_i, b_{ij}\}$ notation

Def. Ideally we want to search over Globally Consistent Marginals w.r.t G.

$$MARG(G) \triangleq \left\{ b = \left[\{b_i\}_{i\in V}, \{b_{ij}\}_{i,j\in E}\right] \text{ s.t. } \exists P(x) \text{ with } \begin{array}{l} b_i(x_i) = \sum_{x_{-i}} P(x), \ \forall i \in V \\ b_{ij}(x_i, x_j) = \sum_{x \setminus \{i,j\}} P(x), \ \forall i,j \in E \end{array} \right\}$$

but checking if $b \in MARG$ is NP-hard.

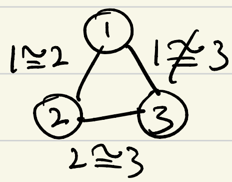
Def. Instead we propose searching over Locally Consistent Marginals w.r.t G

$$LOC(G) \triangleq \left\{ b \text{ s.t. } \begin{array}{l} \sum_{x_i} b_i(x_i) = 1, \ \forall i \in V \\ \sum_{x_j} b_{ij}(x_i, x_j) = b_i(x_i), \ \forall i,j \in E \end{array} \right\}$$

· Local consistency does not imply Global consistency

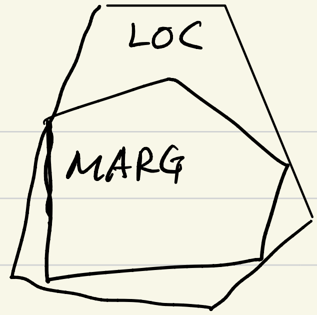
<counter example> $\mathcal{X} = \{0,1\}$, $b_1 = b_2 = b_3 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$

$$b_{12} = b_{23} = \begin{bmatrix} 0.49 & 0.01 \\ 0.01 & 0.49 \end{bmatrix}, \quad b_{13} = \begin{bmatrix} 0.01 & 0.49 \\ 0.49 & 0.01 \end{bmatrix}$$



we can check that $b_i(x_i) = \sum_{x_j} b_{ij}(x_i, x_j)$

- for general graph G.



- for G that is a tree.

$\forall$ Locally consistent $b = \{b_i, b_{ij}\}$, $\exists$ $P(x)$ that is globally consistent for b.

$$\tilde{P}(x) = \prod_{i \in V} b_i(x_i) \cdot \prod_{(i,j) \in E} \frac{b_{ij}(x_i, x_j)}{b_i(x_i)\, b_j(x_j)}$$

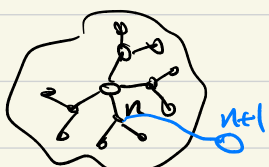
claim: $\tilde{P}(x)$ is globally consistent if G = tree.

that is $\begin{cases} b_i(x_i) = \tilde{P}(x_i), & \forall i \in V \\ b_{ij}(x_i, x_j) = \tilde{P}(x_i, x_j), & \forall (i,j) \in E \end{cases}$

proof: by induction

$n = 1$ $\quad \tilde{P}(x_1) = b(x_1)$ is globally consistent.

suppose it is true for a tree with n nodes, then we consider a tree with 1 more node that is connected to node n.



$$\tilde{P}(x_1^n, x_{n+1}) = \tilde{P}(x_1^n) \cdot \frac{b_{n,n+1}(x_n, x_{n+1})}{b_n(x_n)\, b_{n+1}(x_{n+1})}\, b_{n+1}(x_{n+1})$$

$$\tilde{P}(x_n, x_{n+1}) = \sum_{x_{V \setminus \{n,n+1\}}} \tilde{P}(x_1^n, x_{n+1}) = \tilde{P}(x_n) \cdot \frac{b_{n,n+1}}{b_n \cdot b_{n+1}} \cdot b_{n+1}$$

$$\xrightarrow{\text{induction}} = \cancel{b(x_n)} \cdot \frac{b_{n,n+1}}{\cancel{b_n} \cdot \cancel{b_{n+1}}} \cdot \cancel{b_{n+1}}$$

$$= b_{n,n+1}(x_n, x_{n+1}).$$

Def. Bethe free energy on a tree.

Recall, Gibbs free energy for general $b(x)$ is defined as

$$G_{f_{total}}(b) = -\underbrace{\mathbb{E}_b[-\log f_{total}(X)]}_{\text{Energy}} + \underbrace{\mathbb{E}_b[-\log b(x)]}_{\text{Entropy.}}$$

we evaluate it on $\quad b(x) = \prod_{i\in V} b_i(x_i) \prod_{(i,j)\in E} \frac{b_{ij}(x_i,x_j)}{b_i(x_i)\, b_j(x_j)}$

$$\text{Energy} = -\sum_{(i,j)\in E} \sum_{x_i,x_j} b_{ij}(x_i,x_j) \cdot \log f_{ij}(x_i,x_j)$$

$$\text{Entropy} = \mathbb{E}_b\left[-\log\left(\prod_i b_i(x_i) \prod_{(i,j)} \frac{b_{ij}(x_i,x_j)}{b_i(x_i)\, b_j(x_j)}\right)\right]$$

$$= \sum_i \sum_{x_i}\left(-b_i(x_i)\log b_i(x_i)\right) - \underbrace{\sum_{(i,j)\in E} \sum_{x_i,x_j}\left(\log b_{ij} - \log b_i - \log b_j\right) b_{ij}}_{\sum_{(i,j)\in E}\sum_{x_i,x_j} b_{ij}\log b_{ij} - \sum_{i\in V}\sum_{x_i} \deg(i)\, b_i \log b_i}$$

Bethe free energy

$$\mathbb{F}(b) \triangleq G_{\underset{\text{Tree}}{f_{tree}}}(b) = -\text{Energy} + \text{Entropy}$$

$$= \sum_{(i,j)\in E} \sum_{x_i,x_j} b_{ij}(x_i,x_j)\left(\log f_{ij}(x_i,x_j) - \log b_{ij}(x_i,x_j)\right) + \sum_{i\in V}(\deg(i)-1)\sum_{x_i} b_i(x_i)\log b_i(x_i)$$

claim. if G is a tree then,

$$\sup_{b\in LOC(G)} \mathbb{F}(b) = \sup_{b\in MARG(G)} G(b) = \Phi \quad : \log \text{ partition function.}$$

this is called Bethe variational problem.

if you apply this formula $\mathbb{F}(b)$ and optimize for a non-tree G, then it is called Bethe approximation.

* From Bethe free energy to belief propagation.

claim. fixed points of BP are one-to-one correspondence with stationary points of Bethe variational problem.

Further, BP messages $\{ m_{i\to j}(x_i) \}$ are simple exponentials of Lagrangian variables $\{ \lambda^*_{i,j}(x_i) \}$.

proof.

Define. Lagrangian multipliers λ_i for $\sum_{x_i} b_i(x_i) = 1$

$\lambda_{i\to j}(x_i)$ for $\sum_{x_j} b_{ij}(x_i, x_j) = b(x_i)$

$$\text{Lagrangian}(b,\lambda) = \mathbb{F}(b) - \sum_i \lambda_i \left\{ \sum_{x_i} b(x_i) - 1 \right\} - \sum_{(i,j)} \sum_{x_i} \lambda_{i\to j}(x_i) \left\{ \sum_{x_j} b_{ij}(x_i, x_j) - b_i(x_i) \right\}$$

taking the derivative,

$$\nabla_{b_{ij}(x_i,x_j)} L(b,\lambda) = -1 - \log b_{ij}(x_i, x_j) + \log f_{ij}(x_i, x_j) - \lambda_{i\to j}(x_i) - \lambda_{j\to i}(x_j)$$

$$\nabla_{b_i(x_i)} L(b,\lambda) = -(1 - \deg(i)) \log [b_i(x_i) \cdot e] - \lambda_i + \sum_{j \in \partial i} \lambda_{i\to j}(x_i)$$

Setting them to zero,

$$b_{ij}^*(x_i, x_j) = f_{ij}(x_i, x_j) \exp\left\{ -1 - \lambda_{i\to j}(x_i) - \lambda_{j\to i}(x_j) \right\}$$

$$b_i^*(x_i) \propto \exp\left\{ - \frac{1}{d_i - 1} \sum_{j\in\partial i} \lambda_{i\to j}(x_i) \right\}$$

$$\sum_{x_j} b^*_{ij}(x_i, x_j) = b_i^*(x_i)$$

We change variables as: $m_{i\to j}(x_i) \propto e^{-\lambda_{i\to j}(x_i)}$

$$b_{ij}^*(x_i, x_j) \propto m_{i\to j}(x_i) f_{ij}(x_i, x_j) m_{j\to i}(x_j)$$

$$b_i^*(x_i) \propto \prod_{j\in\partial i} \left\{ (m_{i\to j}(x_i))^{\frac{1}{\deg(i)-1}} \right\}$$

$$\sum_{x_j} b_{ij}^*(x_i, x_j) = b_i^*(x_i)$$

To show equivalence btw BP vs. Bethe free

we start with

$$\prod_{k\in\partial i\setminus j}\left\{\sum_{x_k} b_{ik}^*(x_i,x_k)\right\} \underset{\text{Local Consistency}}{=} \prod_{k\in\partial i\setminus j} b_i^*(x_i) = b_i^*(x_i)^{\deg(i)-1}$$

Stationarity of $F(b)$

$$\text{Stationarity} \;\; \| \qquad\qquad\qquad\qquad \|$$

$$\prod_{k\in\partial i\setminus j}\left\{ m_{i\to k}\cancel{(x_i)}\cdot\sum_{x_k} m_{k\to i}(x_k)\cdot f_{ik}(x_i,x_k)\right\} \propto \prod_{k\in\partial i} m_{i\to k}\cancel{(x_i)}$$

cancelling

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

$$\prod_{k\in\partial i\setminus j}\sum_{x_k} m_{k\to i}(x_k)\, f_{ik}(x_i,x_k) \propto m_{i\to j}(x_i)$$

We arrive at BP update