* Equivalence of Inference tasks. (MARG, MAX, SAMP, compute Z)

· Suppose we have black-box tool $Z(G)$

then we can use it to compute marginal $P(X_1)$

$$P(x_1) = \sum_{x_{-1}} \frac{1}{Z} \boxed{\prod_c f_c(x_c)}$$

$$= \frac{1}{\underset{\overset{\uparrow}{b.b.}}{Z(G)}} \boxed{\sum_{x_{-1}} \underbrace{\boxed{\prod_c f_c(x_c)}}_{G_{-1}}} \quad \underset{\overset{\uparrow}{b.b.}}{Z(G_{-1}(x_1=a))}$$

$$= \frac{Z(G_{-1})}{Z(G)}$$

* We focus on the task of computing Z

for $P(x) = \frac{1}{Z} \underbrace{\prod_{(i,j)\in E} f_{ij}(x_i, x_j)}_{\triangleq f_{total}(x)}$

· Given $f_{total}(x) \longrightarrow$ find $Z \triangleq \sum_x f_{total}(x)$

· focus on approximating log Partition function

$$\Phi \triangleq \log Z$$

· Def. Variational characterization of Φ as solution of an optimization problem.

$$\Phi = \max F(b)$$

b : distribution over $\mathcal{X}^n$, $b \in \Delta_{|\mathcal{X}|^n - 1}$

P.M.F

· Def. Gibbs Free Energy $G_{f_{total}}(b)$.

$$G_{f_{total}}(b) \triangleq \sum_{x \in \mathcal{X}^n} b(x) \cdot \log(f_{total}(x)) - \sum_{x \in \mathcal{X}^n} b(x) \log b(x)$$

$$= - \underbrace{\mathbb{E}_b[\underbrace{-\log f_{total}(x)}_{H(x) = \text{energy at state } x}]}_{\text{expected energy}} + \underbrace{\mathbb{E}_b[-\log b(x)]}_{\text{Entropy of } b(\cdot)}$$

· $\mathbb{P}(x) = \frac{1}{Z} \exp(-H(x))$:

· claim
$$\begin{cases} G(b) \text{ is strictly concave} \\ \max_{b} G(b) = \Phi \\ P(x) = \arg\max_{b} G(b) \end{cases}$$

· proof>

max $$G_{f_{total}}(b) = \sum_{x} b(x) \log f_{total}(x) - \sum_{x} b(x) \log b(x)$$

$$= \sum_{x} b(x) \left\{ \log Z + \log \underbrace{\frac{f_{total}(x)}{Z}}_{P(x)} \right\} - \sum_{x} b(x) \log b(x)$$

$$= \log Z + \sum_{x} \left(b(x) \cdot \log P(x) - b(x) \log b(x) \right)$$

$$= \Phi - D_{KL}(b \| p) \leq \Phi$$ & max achieved $b = p$

↳ Kullback-Leibler divergence

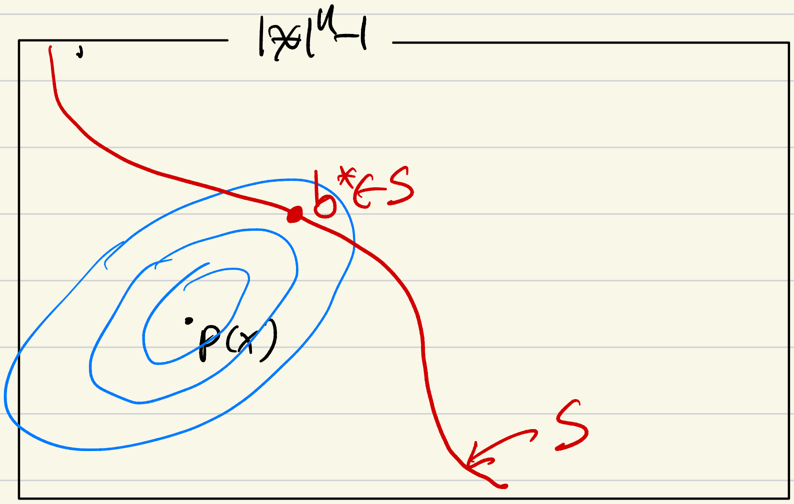
· From Information Theory that

$D_{KL}(b \| p) \geq 0$

$D_{KL}(b \| p) = 0$ iff $b = p$

$D_{KL}(b \| p)$ is convex in b.

$$\max_{b \in \Delta_{|\chi|^n - 1}} G_{f_{total}}(b)$$



strategy

$$\Phi \geq \max_{b \in S} G_{f_{total}}(b)$$

Consider Naive Mean Field factorization

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

$$b = \{b_i(\cdot)\}_{i=1}^{n}$$

dim: $(|\mathcal{X}|-1)\,n$

- plug-in this test dist b.

$$\mathbb{F}_{MF}(b) = \mathbb{G}_{T,free}(b_1 \times b_2 \times \cdots \times b_n) = \sum_{x} b(x) \prod_{ij} \log f_{ij}(x_i, x_j)$$

$$- \sum_{x} b(x) \log b(x)$$

$b = b_1 \times \cdots \times b_n$

$$\boxed{\begin{aligned} &= \sum_{(i,j)\in E} \sum_{x_i, x_j} b_i(x_i) b_j(x_j) \log f_{ij}(x_i, x_j) \\ &\qquad - \sum_{i \in V} \sum_{x_i} b_i(x_i) \log b_i(x_i) \end{aligned}}$$

- Mean Field variational inference Problem

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

$$\text{s.t.} \quad \sum_{x_i} b_i(x_i) = 1 \quad , \forall i \in V$$

I omitted $b_i(x_i) \geq 0$ as log enforces it.

- $b_i(x_i)$'s are approximation $P(x_i)$.
- objective in b is not concave: $b_i(x_i) \cdot b_j(x_j)$ bilinear.

$y = x_1 \cdot x_2$

- Search for a Stationary Point (b^*, λ^*) s.t. $\nabla \mathcal{L}(b,\lambda) = 0$

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

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

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

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

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

Naive Mean Field equation

$$\log b_i^*(x_i) = C + \sum_{j\in\partial i}\sum_{x_j} b_j^*(x_j)\log f_{ij}(x_i,x_j) \quad , \forall i\in V$$

$$b_i^*(x_i) \propto \exp\Big(\sum_{j\in\partial i}\sum_{x_j} b_j^*(x_j)\log f_{ij}(x_i,x_j)\Big) \quad , \forall i\in V$$

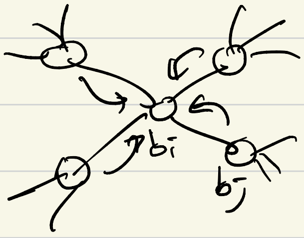
$$b^* = \tilde{F}(b^*)$$

$$b \in \mathbb{R}^{n|\mathcal{X}|}$$

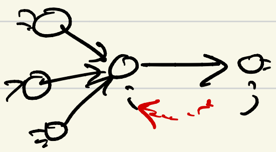
$$b^{(t+1)} \leftarrow \tilde{F}(b^{(t)})$$

$b(x_i) \in \Delta_{|\mathcal{X}|-1}$ is Message/belief

Gossip algorithm.



M.F.



B.P.

n messages:

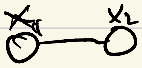
$$b_i(x_i) \leftarrow \exp\Big(\sum_{j\in\partial i}\sum_{x_j} b_j(x_j)\log f_{ij}(x_i,x_j)\Big)$$

$2|E|$ messages:

$$b_{i\to j}(x_j) \leftarrow \prod_{k\in\partial i\setminus \underline{\boxed{j}}}\Big(\sum_{x_k} f_{ik}(x_i,x_k)\, b_{k\to i}(x_k)\Big)$$

* Naive MF is bad.

ex) $x_1, x_2 \in \{0,1\}$



$$P(x) = \frac{1}{2}\mathbb{I}(x_1 = x_2) \qquad\qquad P(x) = \frac{1}{2}\mathbb{I}(x_1 \neq x_2)$$

$$\begin{array}{c|cc|c} x_1 \backslash x_2 & 0 & 1 & \\ \hline 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} & \end{array} \qquad\qquad \begin{array}{c|cc|c} & 0 & 1 & \\ \hline 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & \frac{1}{2} & 0 & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} & \end{array}$$

Naive M.F. [accounts for marginals ($P(x_i)$)
exact only when $P(x) = P(x_1) \cdot P(x_2) \cdots P(x_n)$

Bethe free energy [account for pairwise marginals ($P(x_i, x_j)$)
exact on trees $P(x)$.

simple ✓
Naive Mean Field
Mean field equation

Bethe free energy
Belief Propagation.

Complex ✓
Gibbs free energy

- Parameters: $b_i(x_i) \sim P(x_i)$

$\mathbb{R}^{n(|\mathcal{X}|-1)+|E|(|\mathcal{X}|^2-1)}$ $b_{ij}(x_i, x_j) \sim P(x_i, x_j)$ on $(i,j) \in E$

$\ni b = \{\{b_i\}, \{b_{ij}\}\}$

- Def. Ideally we would like to search over Globally Consistent Marginals w.r.t G

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

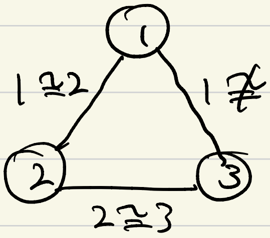
- Checking if $b \in MARG(G)$. is NP-hard.

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

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

- Local consistency can result in a non-distribution (not globally consistent)

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



$$b_{12}(x_1, x_2) = \begin{bmatrix} 0.49 & 0.01 \\ 0.01 & 0.49 \end{bmatrix}$$

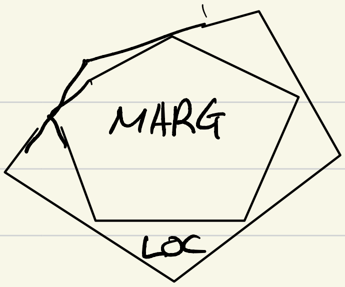
$$b_{13}(x_1, x_3) = \begin{bmatrix} 0.01 & 0.49 \\ 0.49 & 0.01 \end{bmatrix}$$

$$b_{23}(x_2, x_3) = \begin{bmatrix} 0.49 & 0.01 \\ 0.01 & 0.49 \end{bmatrix}$$

b is Locally Consistent, not globally consistent

- for general graph G.

$$b = \{b_i\}_{i\in V}, \{b_{ij}\}_{(i,j)\in E}$$



- for G that is a tree

$\forall$ Locally consistent b, $\exists$ P(x) that is globally consistent.

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