✓ What are GMs?
✓ How to do inference on GMs?
- Sampling
How to learn GMs?

We're talking about (1) how to view BP as optimization and (2) how optimization can help us appx perform inference + sampling

(1) Loopy BP

- G : Send messages on directed edges $\vec{E}$
  - $\nu^{(t+1)}_{i\to j} \propto \prod_{k \in \partial i \setminus \{j\}} \left\{ \sum_{x_k \in \mathcal{X}} \psi_{ik}(x_i, x_k)\, \nu^{t}_{k\to i}(x_k) \right\}$

$$F : M(\mathcal{X})^{\vec{E}} \to M(\mathcal{X})^{\vec{E}} \quad \text{(prob measures on } \mathcal{X}\text{)}$$
$$F(\nu^t) = \nu^{t+1}$$

- If $F(F(F \cdots (F(\nu^1))$ converges, it's to a fixed point of F
$$F(\nu^*) = \nu^*$$
  - Does F have a fixed pt?
  - More than 1? What are they?
  - Does BP find a/the FP of F if it has one?

Existence: (Hadamard 1910, Brouwer 1912)

Any continuous function mapping from a convex, compact (closed + bounded) set to itself has a fixed pt.

(normalized msgs are convex + compact
F is continuous).

- What is/are the fixed pts of F? And does BP find one?

Variational methods treat BP/inference as optimization to analyze.

① [Accurate but complex] ———— Belief Prop ———— [simple/inaccurate]

Gibbs Free Energy — Bethe Free Energy — Näive mean Field

① Gibbs variational principle

- Start w. a hard opt problem
- Add a bunch of constraints & search smaller feasible set
- compare ↑ to BP

- Actual prob: $\mu(x) = \frac{1}{Z} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j) = \frac{1}{Z} \psi_{tot}(x)$

If we had an oracle that comp. Z_G ∀ gms G, we could compute $P(x_i=0) \propto Z_{G(x_i=0)}$, $G_{x_i=0} \triangleq$ the GM / $x_i=0$

- We know ψ_{tot} but not Z
- $b(x) \in M(\mathcal{X}^{|V|})$ can be a "trial" probability/belief
- We'll focus on computing (apx) the log partition fn:

$$\Phi \triangleq \log Z = \log\left(\sum_{x\in\mathcal{X}^n} f_{total}(x)\right)$$

[Def] The Gibbs free energy

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

$$\underbrace{\mathbb{E}_b}_{\text{Expected energy}}\Big[\underbrace{-\log(f_{total}(x))}_{H(x),\ \text{energy of state } x}\Big] + \underbrace{\mathbb{E}_b[-\log(b(x))]}_{\text{Entropy of } b}$$

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

low energy state is more likely

Def
The variational characterization of Φ is $\quad \Phi = \sup_{\substack{b\,:\,\text{dist over } \mathcal{X}^n \\ \text{PMF}}} G(b)$

Claim

- $G_{f\,total}(b)$ is strictly concave
- $\sup_{P_b} G_{f\,total}(b) = \Phi$
- $P(x) = \arg\max_b G(b)$

Pf

[max] $G_{f\,total}(b) = \sum_x b(x) \log\left[f_{total}(x)\right] - \sum_x b(x) \log b(x)$

$= \sum_x b(x) \log\left[\frac{f_{total}(x)}{Z} \cdot Z\right] - \sum_x b(x) \log b(x)$

$= \sum_x b(x)\left[\log Z + \log \underbrace{\frac{f_t(x)}{Z}}_{P(x)}\right] - \sum_x b(x) \log b(x)$

$= \log Z + \sum_x b(x) \log P(x) - b(x) \log(b(x))$

$= \Phi - D_{KL}(b \| P) \leq \Phi$, $\boxed{\text{achieved for } b = P \quad \text{max}}$

$\hookrightarrow$ $\boxed{\text{Kullbeck - Leibler divergence}}$

Info Theory

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

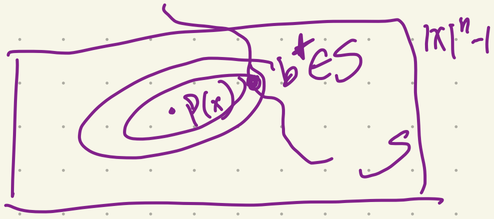
$= 0 \iff b = P$

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

$\max_{b \in \Delta_{|\mathcal{X}|^n - 1}} G_{f\,total}(b)$

Strategy

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



Consider Naïve Mean Field factorization

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

  $b = \{b_i(\cdot)\}_{i=1}^{n}$ $\quad \mathbb{R}^{(X-1)n}$

  dim $(X-1)\cdot n$

- Plug into test distrib b

$$F_{MF}(b) = G_{total}(b_1 \times \cdots \times b_n)$$

$$= \sum_x b(x) \prod_{i,j} \log(f_{ij}(x_i, x_j)) - \sum_x b(x)\log(b(x))$$

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

MF Variational inference prob

$$\max_{b\in S_{MF}} F_{MF}(b)$$

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

$(b_i(x_i) \geq 0$, but log $\Rightarrow$ it$)$

- Dimension $|X|(n-1)$

  b_i's $\approx p_i(x_i)$

- Not concave, bilinear $b_i(x_i)\cdot b_j(x_j)$
- what are local maxima?

Stationary pts of Naive MF are characterized by Lagrangian

$$L(b,\lambda) = F_{MF}(b) - \sum_{i \in V} \lambda_i \Big(\sum_{x_i} b_i(x_i) - 1 \Big)$$

$$= \sum_{\substack{i \in V \\ x_i \in X}} b_i(x_i) \Big\{ \sum_{\substack{j \in \partial i \\ x_j \in X}} b_j(x_j) \log f_{ij}(x_i x_j) \Big\}$$

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

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

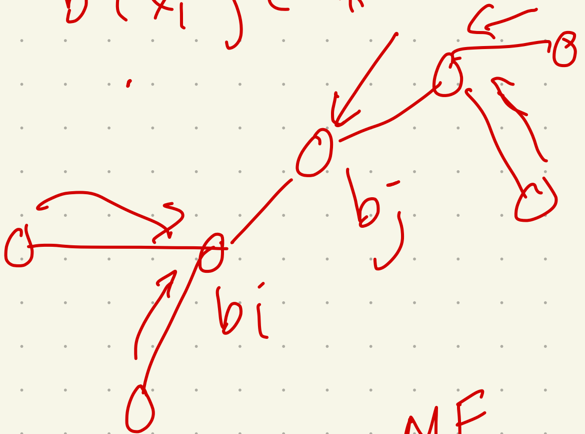
$$\forall i \quad \log b^*_i(x_i) = \overset{(1+\lambda_i)}{C} + \sum_{j \in \delta i} \sum_{x_j} b^*_j(x_j) \log f_{ij}(x_i, x_j)$$

$$\Rightarrow \forall i \quad b^*_i(x_i) \propto \exp[\qquad]$$

$$b^* = \tilde{F}(b^*) \qquad b \in \mathbb{R}^{n|X|}$$

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

$b(x_i) \in \mathbb{R}^{n|x|}$ is a belief/msg like a gossip alg ($\perp$ of j)



n messages $b_i(x_i) \leftarrow \exp\left(\sum_{j \in S_i} \sum_{x_j} b_j(x_j) \log f_{ij}\right)$

MF

vs messages on edges (BP), $2|E|$

$$b_{i \to j}(x_j) = \prod_{k \in S_i \setminus j} f_{ik}() \, b_{k \to i}(x_k)$$

Naïve Mean Field is a v poor apx. Ex: $x_1, x_2 \in \{0,1\}$ $x_1 - x_2$

$$P(x) = \tfrac{1}{2}\mathbb{I}[x_1 = x_2] \quad \text{or} \quad P(x) = \tfrac{1}{2}\mathbb{I}[x_1 \neq x_2]$$

Not well approx by $b(x_1, x_2) = b_1(x_1) b_2(x_2) \Rightarrow \perp$

Is there some parametrization that
- accts for pairwise corr?
- Exact y on trees?

Bethe Free Energy

accts for pairwise corr

- Exact $\checkmark$ on trees

Params $b_i(x_i) \cong P(x_i)$

$b_{ij}(x_i, x_j) \cong P(x_i, x_j)$

Abuse notation + $b = \{b_i, b_{ij}\}$ wrt G

Ideally, wanna search over globally consistent marg wrt G

Def $\text{Marg}(G) \triangleq \{ b : \{b_i\}_{i \in V}, \{b_{ij}\}_{(i,j) \in E}$ s.t. $\exists P(x)$ w.

$$b_i(x_i) = \sum_{x_{-i}} P(x) \quad \forall i \in V$$

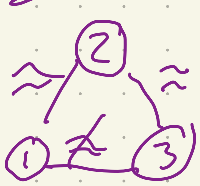
$$b_{ij}[x_i, x_j] = \sum_{x_{-i,j}} P(x) \quad \forall (i,j) \in E$$

but checking if $b \in \text{Marg}$ is NP-hard :(

Instead,

Def $\text{Loc}(G) \triangleq \{ b : \sum_{x_i} b_i(x_i) = 1 \quad \forall i \in V$

$$\sum_{x_j} b_{ij}(x_i, x_j) = b_i(x_i) \quad \forall (i,j) \in E \}$$



Local $\not\Rightarrow$ Global

$X = \{0, 1\}$ $\quad b_i = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$ $\quad b_{12} = b_{23} \begin{bmatrix} .49 & .01 \\ .01 & .49 \end{bmatrix}$ $\quad b_{13} = \begin{bmatrix} .01 & .49 \\ .49 & .01 \end{bmatrix}$

For general G

When G is a tree, $\forall$ locally consistent b $\exists$ $P(x)$ which is globally consistent for b

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

Claim: $\tilde{P}$ is consistent if G is a tree

Pf By induction on n.

$n=1$ $\quad \tilde{P}(x_1) = b_1(x_1)$ ✓

suppose true $\forall$ trees of size $\leq n$

Fix a tree of size $n+1$. It has a leaf.

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

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

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

$$\overset{IH}{=} \cancel{b_n(x_n)} \frac{b_{n,n+1}}{\cancel{b_n} \cdot \cancel{b_{n+1}}} \cdot \cancel{b_{n+1}}$$ ✓

Def Bethe Free Energy on a tree

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

$$= \sum_{(i,j)\in E} \sum_{x_i,x_j} b_{ij}(x_i,x_j) \log(f_{ij}(x_i,x_j)) - \log[b_{ij}(x_i,x_j)] +$$

$$\sum_{i\in V} (\deg(i)-1) \sum_{x_i} b_i(x_i) \log b_i(x_i)$$

Claim: if G is a tree, then

$$\sup_{b\in LOC(G)} \mathbb{F}(b) = \sup_{b\in Marg(G)} G(b) = \Phi$$

This is the Bethe variational prob

(opt on a non-tree is called Bethe appx)

Bethe free energy & BP

Claim: (1) fixed pts of BP are one-to-one w. stationary pts of Beth.

(2) BP messages $\{m_{i\to j}(x_i)\}$ are the exponentials of lagrangians $\{\lambda^*_{ij}(x_i)\}$

Gibbs Free Energy for gen b is

$$G_{F total}(b) = -\underbrace{\mathbb{E}_b[-\log(f\,total(b))]}_{\text{energy}} + \underbrace{\mathbb{E}_b[-\log b(x)]}_{\text{entropy}}$$

Evaluate on

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

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

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

$$= \sum_i \sum_{x_i} -b_i \log b_i(x_i)$$

$$- \sum_{\substack{i,j\in E\\ x_i,x_j}} b_{ij} (\log b_{ij} - \log b_i - \log b_j)$$

$$\sum_{\substack{(i,j)\in E\\ x_i,x_j}} b_{ij} \log b_{ij} - \sum_{i\in V}\sum_{x_i} \deg(i)\, b_i \log b_i$$

Pf Lagrangian mults: λ_i for $\sum_{x_i} b_i(x_i) = 1$

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

$$\mathrm{Lag}(b,\lambda) = F_b - \lambda_i\Big[\sum_{x_i} b_i(x_i) - 1\Big] - \sum_{i,j}\sum_{x_i} \lambda_{i\to j}(x_i)\Big[\sum_{x_j} b_{ij}(x_i\, x_j) - b_i(x_i)\Big]$$

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

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

setting $=0$ & solving $\Rightarrow$

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

$$b^*_i(x_i) \propto \exp\Big\{-\frac{1}{d-1}\sum_{j\in\partial i} \lambda_{i\to j}(x_i)\Big\}$$

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

Change variables $m_{i\to j}(x_i) \propto e^{-\lambda_{i\to j}(x_i)}$

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

$$b^*_i \propto \prod_{j\in\partial i} (m_{i\to j}(x_j))^{\deg(i)-1}$$

To show $\equiv$

$$\prod_{k\in\delta i\setminus j}\Big\{\sum_{x_k} b^*_{ik}(x_i\, x_k)\Big\} = \prod_{k\in\delta i\setminus j} b_i^*(x_i) = b_i^*(x_i)^{\deg(i)-1}$$

Local consistency

$\| \leftarrow$ stationarity of F_b

$\} \| \leftarrow$ stationarity

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

cancelling

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

The BP update!