10. Variational methods

- Gibbs free energy
- Naive mean field
- Bethe free energy
- Region-based approximation
- Tree-based convexification

Loopy belief propagation

- directed edges on G: $\vec{E}$
- messages: $\nu^{(t)} \equiv \{\nu_{i\to j}(\cdot)\}_{(i,j)\in\vec{E}}$
- loopy belief propagation: $\nu^{(t+1)} = \mathsf{F}(\nu^{(t)})$

$$\nu_{i\to j}^{(t+1)} \propto \prod_{k\in\partial i\setminus j} \Big\{ \sum_{x_k\in\mathcal{X}} \psi_{ik}(x_i,x_k)\nu_{k\to i}^{(t)}(x_k) \Big\}$$

$$\begin{aligned} \mathsf{F} : \mathsf{M}(\mathcal{X})^{\vec{E}} &\to \mathsf{M}(\mathcal{X})^{\vec{E}} \\ \nu &\mapsto \mathsf{F}(\nu) \end{aligned}$$

where $\mathsf{M}(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$

- if loopy BP converges, it eventually conerges to a **fixed point** of F

$$\nu^* = \mathsf{F}(\nu^*)$$

1. does F have a fixed point?
2. if F has one or more fixed points, what are they?
3. does BP converge to a fixed point?

Existence of a fixed point

- **Theorem.** (Hadamard 1910, Brouwer 1912) Any continuous function mapping from a convex compact set to the same convex compact set has a fixed point.
- existence of at least one fixed point of F follows from
  - F is continuous
  - the set of normalized messages is convex and compact

- but what do these fixed points correspond to?
- and how do they relate to BP?
- **variational approach** tries to answer these questions by formulating the inference problem as an optimization problem

Gibbs variational principle

- start with a hard optimization problem
- approximate the solution by imposing constraints and searching in a smaller feasible set
- relate the solutions of the relaxation to BP

- 'actual' probability

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

- we know ψ_{tot} but not Z
- 'trial' probability ('belief') $b(x) \in \mathsf{M}(\mathcal{X}^{|V|})$

we focus on characterizing **log partition function**

$$\Phi \equiv \log Z = \log \Big\{ \sum_{x \in \mathcal{X}^{|V|}} \prod_{(i,j) \in E} \psi_{ij}(x_i, x_j) \Big\}$$

- **variational characterization** of the log partition function

$$\Phi = \sup_{b \in \mathsf{M}(\mathcal{X}^{|V|})} \mathbb{G}(b)$$

- define **Gibbs free energy** $\mathbb{G}_\psi(b)$

$$\begin{aligned}\mathbb{G}_\psi(b) &= \sum_{x\in\mathcal{X}^{|V|}} \big(b(x)\log\psi_{\text{tot}}(x)\big) - \sum_{x\in\mathcal{X}^{|V|}} \big(b(x)\log b(x)\big) \\ &= -\underbrace{\mathbb{E}_b\big[-\log\psi_{\text{tot}}(X)\big]}_{\text{expected energy w.r.t. } b} + \underbrace{\mathbb{E}_b\big[-\log b(X)\big]}_{\text{entropy of } b}\end{aligned}$$

such that

- strictly concave
- $\sup_{b\in \mathrm{M}(\mathcal{X}^{|V|})} \mathbb{G}_\psi(b) = \Phi$
- $\mu = \arg\max_b \mathbb{G}_\psi(b)$

- interpretation
  - the optimal solution $b^*(x) = \mu(x)$ minimizes average energy while maximizing entropy

Proof of $\Phi = \sup_b \mathbb{G}_\psi(b)$

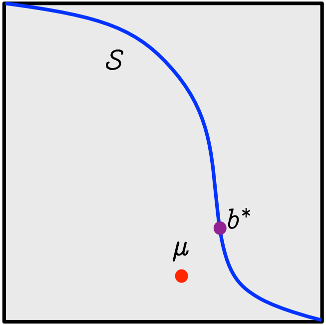
- rearranging terms,

$$\begin{aligned}
\mathbb{G}_\psi(b) &= \sum_{x\in\mathcal{X}^{|V|}} \big(b(x)\log\psi_{\text{tot}}(x)\big) - \sum_{x\in\mathcal{X}^{|V|}} \big(b(x)\log b(x)\big) \\
&= \sum_{x\in\mathcal{X}^{|V|}} b(x)\big(\log Z + \log\frac{1}{Z}\psi_{\text{tot}}(x)\big) - \sum_{x\in\mathcal{X}^{|V|}} \big(b(x)\log b(x)\big) \\
&= \log Z - \sum_{x\in\mathcal{X}^{|V|}} b(x)\big(\log b(x) - \log\mu(x)\big) \\
&= \Phi - D_{\text{KL}}(b||\mu)
\end{aligned}$$

  where $D_{\text{KL}}(\cdot||\cdot)$ is the **Kullback-Leibler divergence**
- from information theory, it is known that
  - $D_{\text{KL}}(b||\mu) \geq 0$
  - $D_{\text{KL}}(b||\mu) = 0$ if and only if $b = \mu$

- good news: we can compute partition function Z by solving a convex optimization
- bad news: $\mathrm{M}(\mathcal{X}^{|V|})$ is $|\mathcal{X}|^{|V|} - 1$ dimensional

- next strategy: solve the optimization over a low-dimensional subset S



- this give a **lower bound** on the log partition function, because we are maximizing over a smaller set

$$\Phi \;\; \geq \;\; \sup_{b \in S} \mathbb{G}_\psi(b)$$

Naive mean field

- define a subset of distributions that can be factorized according to **naive mean field factorization**

$$S_{\mathrm{MF}} = \{ b \in \mathrm{M}(\mathcal{X}^n) \,:\, b(x) = b_1(x_1) \times b_2(x_2) \times \cdots \times b_n(x_n) \}$$

- slight abuse of notation: $b = \{b_i\}_{i \in V}$
- let

$$\begin{aligned} \mathbb{F}_{\mathrm{MF}} \,:\, S_{\mathrm{MF}} &\to \mathbb{R} \\ b &\mapsto \mathbb{G}_{\psi}(b) \end{aligned}$$

- we can compute it explicitly, after some algebra

$$\mathbb{F}_{\mathrm{MF}}(b) \;=\; \sum_{(i,j)\in E} \sum_{x_i, x_j} b_i(x_i) b_j(x_j) \log \psi_{ij}(x_i, x_j) - \sum_{i \in V, x_i} b_i(x_i) \log b_i(x_i)$$

- mean field variational inference problem

$$\begin{aligned} \max_{b \in S_{\mathrm{MF}}} \quad & \mathbb{F}_{\mathrm{MF}}(b) \\ \text{subject to} \quad & b_i(x_i) \geq 0 && \text{for all } i \in V,\, x_i \in \mathcal{X} \\ & \sum_{x_i \in \mathcal{X}} b_i(x_i) = 1 && \text{for all } i \in V \end{aligned}$$

- consider b_i's as approximate node marginals
- although $\mathbb{F}_{\mathrm{MF}}(\cdot)$ is not concave, we can search for local maxima
- characterizing the local maxima
  - the stationary points of a constrained optimization satisfy that the derivatives of the Lagrangian are zero

$$\begin{aligned}
L(b,\lambda) &= \mathbb{F}_{\mathrm{MF}}(b) - \sum_{i\in V}\lambda_i\Big\{\sum_{x_i\in\mathcal{X}} b_i(x_i) - 1\Big\} \\
&= \frac{1}{2}\sum_{i\in V, x_i\in\mathcal{X}} b_i(x_i)\Big\{\sum_{j\in\partial i, x_j\in\mathcal{X}} b_j(x_j)\log\psi_{ij}(x_i,x_j)\Big\} - \sum_{i\in V, x_i} b_i(x_i)\log b_i(x_i) \\
&\qquad - \sum_{i\in V}\lambda_i\Big\{\sum_{x_i\in\mathcal{X}} b_i(x_i) - 1\Big\}
\end{aligned}$$

  - define a **Lagrangian multiplier** λ_i for each constraint $\sum b_i(x_i) = 1$
  - non-negativity constraints are implicit from the log

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

- solving for $b_i(x_i)$ we get **naive mean field equations**:

$$\begin{aligned} b_i(x_i) &\propto \exp\Big\{ \sum_{j\in\partial i} \sum_{x_j \in \mathcal{X}} \log \psi_{ij}(x_i, x_j) b_j(x_j) \Big\} \\ b &= \mathsf{F}_{\mathrm{MF}}(b) \end{aligned}$$

- a fixed point can be searched by iteration:

$$b^{(t+1)} = \mathsf{F}_{\mathrm{MF}}(b^{(t)})$$

Bethe free energy

- one dimensional marginals give a very poor approximation
- **example:** $x_1, x_2 \in \{0, 1\}$

$$\begin{aligned} \mu(x) &= \frac{1}{2}\,\mathbb{I}(x_1 \oplus x_2 = 0) \qquad \text{and} \\ \mu(x) &= \frac{1}{2}\,\mathbb{I}(x_1 \oplus x_2 = 1) \end{aligned}$$

- would like to define a parameterization of $b(x)$ such that
  - account exactly for the pairwise correlations induced by edges
  - exact on distribution μ defined over a tree

Locally consistent marginals

- consider a parametrization
  - $b_i(x_i)$: an approximation of the marginal $\mu(x_i)$
  - $b_{ij}(x_i, x_j)$: as an approximation of the marginal $\mu(x_i, x_j)$
- let $b = \{b_i, b_{ij}\}$
- b is a set of **globally consistent marginals of a distribution on $\mathcal{X}^n$** if there exists a $\mathbb{P}(\cdot) \in \mathsf{M}(\mathcal{X}^{|V|})$ such that

$$
\begin{aligned}
b_i(x_i) &= \sum_{x_{V \setminus \{i\}}} \mathbb{P}(x) && \text{, for all } i \\
b_{ij}(x_i, x_j) &= \sum_{x_{V \setminus \{i,j\}}} \mathbb{P}(x) && \text{, for all } i, j
\end{aligned}
$$

- denote the set of all valid marginals by

$$
\mathrm{MARG}(G) = \left\{ b = \{b_i, b_{ij}\} : \text{ marginals of a distribution on } \mathcal{X}^{|V|} \right\}
$$

- in general, checking $b \in \mathrm{MARG}(G)$ is NP-hard

- $b = \{b_i, b_{ij}\}$ is a set of **locally consistent marginals** if

$$\begin{aligned} \sum_{x_i} b_i(x_i) &= 1 & \text{, for all } i \\ \sum_{x_j} b_{ij}(x_i, x_j) &= b_i(x_i) & \text{, for all } i, j \end{aligned}$$

  - not all locally consistent marginals correspond to a valid joint probability distribution
  - **example.** three nodes with $\mathcal{X} = \{0, 1\}$

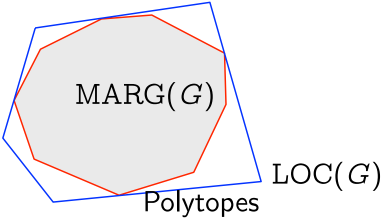
$$b_1 = b_2 = b_3 = (0.5, 0.5)$$

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

$$b_{31} = \begin{bmatrix} 0.01 & 0.49 \\ 0.49 & 0.01 \end{bmatrix}$$

- denote the set of all locally consistent marginals by

$$\mathrm{LOC}(G) \;=\; \Big\{ b = \{b_i, b_{ij}\} \;:\; \text{locally consistent marginals} \Big\}$$



- when G is not a tree
  - locally consistent $\{b_i, b_{ij}\}$ might not be marginals of any distribution
- when G is a tree
  - for any locally consistent $\{b_i, b_{ij}\}$, there exists a unique measure $p \in \mathsf{M}(\mathcal{X}^{|V|})$ whose marginals are given by $\{b_i, b_{ij}\}$
  - the measure $p(x)$ is given by

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

  - (we did not define Bethe free energy $\mathbb{F}(\{b_i, b_{ij}\})$ yet, but) the Gibbs free energy is equal to the Bethe free energy, i.e. $\mathbb{G}(p) = \mathbb{F}(\{b_i, b_{ij}\})$, and hence

$$\log Z \;=\; \max_{\{b_i, b_{ij}\} \in \mathrm{LOC}(G)} \mathbb{F}(\{b_i, b_{ij}\})$$

Locally consistent marginals on a tree

- given a tree $G = (V, E)$ with n nodes and $\{b_i, b_{ij}\} \in \mathrm{LOC}(G)$
- prove (by induction) that

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

  is a unique measure on $\mathcal{X}^n$ with marginals $p(x_i) = b_i(x_i)$ for all i and $p(x_i, x_j) = b_{ij}(x_i, x_j)$ for all $(i, j) \in E$
- for $n = 1$, it is trivial
- assume it is true for n and add a new vertex $i = n + 1$, connected to $j = n$

$$\begin{aligned}
p(x_V, x_{n+1}) &= p(x_V)\, p(x_{n+1}|x_V) \\
&= p(x_V)\, p(x_{n+1}|x_n) && \text{[Markov property]} \\
&= p(x_V)\, \frac{p(x_n, x_{n+1})}{p(x_n) p(x_{n+1})} p(x_{n+1}) && \text{[Bayes rule]} \\
&= \left\{ \prod_{(i,j) \neq (n,n+1)} \frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \prod_{i \in V} b_i(x_i) \right\} \frac{p(x_n, x_{n+1})}{p(x_n) p(x_{n+1})} p(x_{n+1})
\end{aligned}$$

QED

Bethe free energy

- variational inference on locally consistent marginals $b = \{b_i, b_{ij}\}$
  - want to define an objective function

$$
\begin{aligned}
\mathbb{F} \;:\; \mathrm{LOC}(G) \quad & \to \mathbb{R} \\
b = \{b_{ij}, b_\ell\}_{(i,j)\in E, \ell \in V} \quad & \mapsto \mathbb{F}(b)
\end{aligned}
$$

  such that

$$
\begin{aligned}
\arg\max_b \mathbb{F}(b) \quad & \approx \quad \mu\,, \\
\max_b \mathbb{F}(b) \quad & \approx \quad \Phi\,,
\end{aligned}
$$

- recall that for a valid distribution b, Gibbs free energy is defined as

$$
\mathbb{G}_\psi(b) \quad = \quad -\underbrace{\mathbb{E}_b[-\log \psi_{\mathrm{tot}}(X)]}_{\text{energy}} + \underbrace{\mathbb{E}_b[-\log b(X)]}_{\text{entropy}}
$$

- when G is a tree, the first and second order marginals fully describe the joint distribution:

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

- **Bethe free energy** on a tree
  - energy

$$
\begin{aligned}
\mathbb{E}_b[-\log \psi_{\texttt{tot}}(X)] &= -\sum_{(i,j)\in E} \mathbb{E}_b[\log \psi_{ij}(x_i, x_j)] \\
&= -\sum_{(i,j)\in E} \mathbb{E}_{b_{ij}}[\log \psi_{ij}(x_i, x_j)] \\
&= -\sum_{(i,j)\in E} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \log \psi_{ij}(x_i, x_j)
\end{aligned}
$$

  - entropy

$$
\begin{aligned}
H(b) &\equiv \mathbb{E}_b[-\log b(X)] \\
&= \sum_{i\in V} \underbrace{-\mathbb{E}_{b_i}[\log b_i(X_i)]}_{\equiv H(b_i)} - \sum_{(i,j)\in E} \underbrace{-\mathbb{E}_{b_{ij}}[\log b_{ij}(X_i, X_j) - \log b_i(X_i) - \log b_j(X_j)]}_{\equiv I(b_{ij})}
\end{aligned}
$$

- in general, define **Bethe free energy** of $b = \{b_i, b_{ij}\} \in \mathrm{LOC}(G)$ as

$$\begin{aligned}
\mathbb{F}(b) &= -\text{ energy } + \text{ entropy} \\
&= \sum_{(i,j)\in E} \sum_{x_i,x_j} b_{ij}(x_i,x_j) \log \psi_{ij}(x_i,x_j) \\
&\qquad - \sum_{(i,j)\in E} \sum_{x_i,x_j} b_{ij}(x_i,x_j) \log \frac{b_{ij}(x_i,x_j)}{b_i(x_i)b_j(x_j)} - \sum_{i\in V} \sum_{x_i} b_i(x_i) \log b_i(x_i) \\
&= \sum_{(i,j)\in E} \sum_{x_i,x_j} b_{ij}(x_i,x_j) \log \psi_{ij}(x_i,x_j) \\
&\qquad - \sum_{(i,j)\in E} \sum_{x_i,x_j} b_{ij}(x_i,x_j) \log b_{ij}(x_i,x_j) - \sum_{i\in V} (1-\deg(i)) \sum_{x_i} b_i(x_i) \log b_i(x_i)
\end{aligned}$$

- one justification of using $\mathbb{F}(\cdot)$ is that if G is a *tree* then

$$\sup_{\{b_i,b_{ij}\}} \mathbb{F}(\{b_i, b_{ij}\}) = \sup_{b\in \mathsf{M}(G)} \mathbb{G}(b) = \Phi$$

  where $\mathsf{M}(G)$ is the set of distributions that decompose according to G
- the above optimization problem is called **Bethe variational problem**
- for general graphs, the solution to the above maximization approximates the log partition function, and it is known as **Bethe approximation**

Connections between Bethe free energy and belief propagation

- maximizing Bethe free energy

$$\max_{b \in \mathrm{LOC}(G)} \mathbb{F}(b)$$

- **Theorem.** (Yedidia, Freeman, Weiss 2003) Fixed points of BP are in one-to-one correspondence with stationary points of Bethe free energy.
- Also, fixed point BP messages ν^* are (exponentials of) the dual parameters λ^* at the fixed points

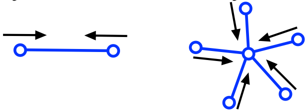
Fixed point condition for BP messages

- BP fixed point messages ν^* satisfy

$$\nu^*_{i\to j}(x_i) \propto \prod_{k\in\partial i\setminus j} \Big\{ \sum_{x_k} \psi_{ik}(x_i, x_k)\nu^*_{k\to i}(x_k) \Big\}$$

- define a set of marginals (which are exact on a tree)

$$\begin{aligned} b^*_i(x_i) &\propto \prod_{k\in\partial i} \Big\{ \sum_{x_k\in\mathcal{X}} \psi_{ik}(x_i, x_k)\,\nu^*_{k\to i}(x_k) \Big\} \\ &\propto \prod_{k\in\partial i} \Big\{ \big(\nu^*_{i\to k}(x_i)\big)^{\frac{1}{\deg(i)-1}} \Big\} \\ b^*_{ij}(x_i, x_j) &\propto \nu^*_{i\to j}(x_i)\,\psi_{ij}(x_i, x_j)\,\nu^*_{j\to i}(x_j) \end{aligned}$$



- **exercise.** show $\{b_i, b_{ij}\}$ is locally consistent
- **claim.** b^* corresponds to a stationary point of Bethe free energy

Stationarity condition for Bethe free energy

Lagrangian with λ_i for condition $\sum_{x_i} b_i(x_i) = 1$, and $\lambda_{i\to j}(x_i)$ for condition $\sum_{x_j} b_{ij}(x_i, x_j) = b_i(x_i)$

$$
\begin{aligned}
\mathcal{L}(b,\lambda) &= \mathbb{F}(b) - \sum_{i\in V} \lambda_i \Big\{ \sum_{x_i} b_i(x_i) - 1 \Big\} \\
&\quad - \sum_{(i,j)\in\vec{E}} \sum_{x_i} \lambda_{i\to j}(x_i) \Big\{ \sum_{x_j} b_{ij}(x_i,x_j) - b_i(x_i) \Big\}
\end{aligned}
$$

taking the derivative

$$
\begin{aligned}
\nabla_{b_{ij}(x_i,x_j)} \mathcal{L}(b,\lambda) &= -1 - \log b_{ij}(x_i,x_j) + \log \psi_{ij}(x_i,x_j) - \lambda_{i\to j}(x_i) - \lambda_{j\to i}(x_j) \\
\nabla_{b_i(x_i)} \mathcal{L}(b,\lambda) &= -(1-\deg(i)) \log[b_i(x_i)\, e] - \lambda_i + \sum_{j\in\partial i} \lambda_{i\to j}(x_i)
\end{aligned}
$$

setting the derivatives to zero

$$
\begin{aligned}
b^*_{ij}(x_i,x_j) &= \psi_{ij}(x_i,x_j) \exp\Big\{ -1 - \lambda_{i\to j}(x_i) - \lambda_{j\to i}(x_j) \Big\}, \\
b_i(x_i)^* &\propto \exp\Big\{ -\frac{1}{\deg(i)-1} \sum_{j\in\partial i} \lambda_{i\to j}(x_i) \Big\}
\end{aligned}
$$

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

- changing variables: $\nu_{i\to j}(x_i) \propto e^{-\lambda_{i\to j}(x_i)}$

$$
\begin{aligned}
b^*_{ij}(x_i, x_j) &\propto \nu_{i\to j}(x_i)\,\psi_{ij}(x_i, x_j)\,\nu_{j\to i}(x_j) \\
b^*_i(x_i) &\propto \prod_{j\in\partial i}\left\{\left(\nu_{i\to j}(x_i)\right)^{\frac{1}{\mathbf{deg}(i)-1}}\right\}
\end{aligned}
$$

- imposing locally consistency constraints $\sum_{x_j} b^*_{ij}(x_i, x_j) = b^*_i(x_i)$, we can show that the $\nu_{i\to j}$'s are at BP fixed point. Start with the identity

$$
\begin{aligned}
\prod_{k\in\partial i\setminus j}\Big\{\underbrace{\sum_{x_k} b^*_{ik}(x_i, x_k)}_{=b^*_i(x_i)}\Big\} &= b^*_i(x_i)^{\mathbf{deg}(i)-1} \text{ , substitute } \nu\text{'s} \\
\prod_{k\in\partial i\setminus j}\Big\{\nu_{i\to k}(x_i)\sum_{x_k}\nu_{k\to i}(x_k)\psi_{ik}(x_i, x_k)\Big\} &\propto \prod_{k\in\partial i}\Big\{\nu_{i\to k}(x_i)\Big\} \text{ , after a division} \\
\prod_{k\in\partial i\setminus j}\Big\{\sum_{x_k}\nu_{k\to i}(x_k)\psi_{ik}(x_i, x_k)\Big\} &\propto \nu_{i\to j}(x_i)
\end{aligned}
$$

- we have established that each of the BP fixed points correspond to a stationary point of the Bethe free energy

- Alternative algorithms to find fixed points (e.g. gradient ascent)
  [e.g. Heskes 2002]
- Include higher order marginals
  [Yedidia, Freeman, Weiss 2003]
- Convexify Bethe free energy
  [Wainwright, Jaakkola, Willsky 2005]
- Asymptotically tight estimates on $\log Z$ for graph sequences
  [e.g. Dembo, Montanari 2010]

- Historically, statistical physics study systems in thermal equilibrium, whose state is given by **Boltzmann's law**

$$\mu(x) = \frac{1}{Z(T)} e^{-E(x)/T}$$

where T is the temperature, $E(x)$ is the energy at a state x, and $Z(T)$ is the partition function given by

$$Z(T) = \sum_{x \in S} e^{-E(x)/T}$$

Helmholtz free energy (log partition function) is an important quantity for understanding how the system and statistical physicists have devoted significant energy to find good approximations to it:

$$F_H = -\ln Z(T)$$

An important technique is based on variational approaches, where the maximum of Gibbs free energy is studied

$$\mathbb{G}(b) = \sum_{x \in S} b(x) E(x) + \sum_{x \in S} b(x) \log b(x)$$

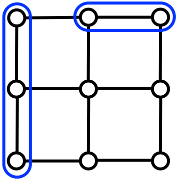
Region-based approximation

Consistency on Vertices	$\rightarrow$	Edges	$\rightarrow$	Regions
Naive Mean Field	$\rightarrow$	Bethe Free Energy	$\rightarrow$	Region-Based Free Energy
MF Equations	$\rightarrow$	Belief Propagation	$\rightarrow$	Generalized BP

[Cluster variational method, Kikuchi 1951]

- Idea: decompose the system into sub-systems (regions) and approximate the free energy by combining the free energies of the sub-systems

Region



- **Definitions.**
  - a **region** $R = (V_R, E_R)$ is a subgraph such that if $(i,j) \in E_R$ then $i,j \in V_R$
  - **region free energy** $\mathbb{F}_R : \mathrm{M}(\mathcal{X}^{V_R}) \to \mathbb{R}$

$$
\begin{aligned}
\mathbb{F}_R(b_R) &= \mathbb{E}_{b_R} \log \psi_{\mathrm{tot},R}(x_R) + H(b_R) \\
&= \underbrace{-\sum_{x_R} \sum_{(i,j)\in E_R} -b_R(x_R) \log \psi_{ij}(x_i, x_j)}_{\text{region energy}} + \underbrace{\sum_{x_R} -b_R(x_R) \log b_R(x_R)}_{\text{region entropy}}
\end{aligned}
$$

  - can be evaluated for small regions (complexity $|\mathcal{X}|^{|R|}$)

Region-based approximation

- **collection of regions**

$$\mathbf{R} \;=\; \{R_1, R_2, \ldots, R_m\}\,.$$

- **coefficients**

$$c_{\mathbf{R}} \;=\; \{c_{R_1}, c_{R_2}, \ldots, c_{R_m}\}\,, \qquad c_{R_i} \in \mathbb{R}\,.$$

- **marginals**

$$b_{\mathbf{R}} \;=\; \{b_{R_1}, b_{R_2}, \ldots, b_{R_m}\}\,, \qquad b_{R_i} \in \mathsf{M}(\mathcal{X}^{V_{R_i}})\,.$$

- **region-based free energy approximation**:

$$\mathbb{F}_{\mathbf{R}}(b_{\mathbf{R}}) \;=\; \sum_{R \in \mathbf{R}} c_R\, \mathbb{F}_R(b_R)$$

Example: Bethe Free Energy

- regions

$$
\begin{aligned}
\mathbf{R} &= \{R_i : i \in V\} \cup \{R_{ij} : (i,j) \in E\} \\
R_i &= (\{i\}, \emptyset) \\
R_{ij} &= (\{i,j\}, \{(i,j)\})
\end{aligned}
$$

- coefficients

$$
c_i = 1 - \deg(i)\,, \qquad c_{ij} = 1\,.
$$

- Bethe free energy as a special case of the region based free energy

$$
\mathbb{F}_{\mathbf{R}}(b) = \sum_{i \in V} \{1 - \deg(i)\}\, H(b_i) + \sum_{(i,j) \in E} \Big\{ H(b_{ij}) + \mathbb{E}_{b_{ij}} \log \psi_{ij}(x_i, x_j) \Big\}
$$

- main questions
  1. What about domain/consistency of $b_R(x_R)$?
  2. How to choose coefficients?
  3. How to choose regions?
- **valid** region-based approximations [Yedidia, Freeman, Weiss, 2003]
  - **condition 1:** local consistency

$$R \in \mathbf{R},\, R' \subseteq R \quad \Rightarrow \quad R' \in \mathbf{R}\,.$$

$$\sum_{x_{R \setminus R'}} b_R(x_R) = b_{R'}(x_{R'}) \qquad \text{for all } R' \subseteq R\,.$$

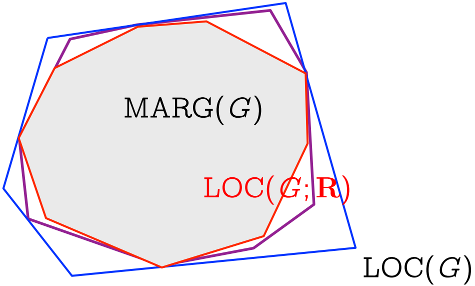
    let $\mathrm{LOC}(G;\mathbf{R})$ be a set of marginals $b = \{b_R : R \in \mathbf{R}\}$ that are locally consistent w.r.t. a collection of regions $\mathbf{R}$
  - **condition 2:** vertex counting

$$\sum_{R \in \mathbf{R}} c_R\, \mathbb{I}(i \in R) = 1 \qquad \text{for all } i \in V\,.$$

  - **condition 3:** edge counting

$$\sum_{R \in \mathbf{R}} c_R\, \mathbb{I}((i,j) \in R) = 1 \qquad \text{for all } (i,j) \in E$$

Geometric picture



Polytopes

Justification of condition #2

$$\sum_{R \in \mathbf{R}} c_R \, \mathbb{I}(i \in R) = 1 \qquad \text{for all } i \in V \,.$$

- consider a special case of uniform distribution: $\mu(x) = 1/|\mathcal{X}|^{|V|}$ and $\psi_{ij}(x_i, x_j) = 1$
- suppose $b_R(x_R)$ are true marginals, i.e. $b_R^*(x_R) = 1/|\mathcal{X}|^{|V_R|}$
- then for any graph, the region based approximation is exact: $\mathbb{F}_{\mathbf{R}}(b^*) = \log Z$

since $\log \psi_{ij}(x_i, x_j) = 0$, energy terms are zeros

$$\begin{aligned}
\sum_{R \in \mathbf{R}} c_R \mathbb{F}_R(b_R^*) &= \sum_{R \in \mathbf{R}} c_R \, H(b_R^*) \\
&= \sum_{R \in \mathbf{R}} c_R \underbrace{|V_R|}_{\sum_{i \in V} \mathbb{I}(i \in R)} \log |\mathcal{X}| \\
&= \sum_{i \in V} \left\{ \sum_{R \in \mathbf{R}} c_R \, \mathbb{I}(i \in R) \right\} \log |\mathcal{X}| \\
&= |V| \log |\mathcal{X}|
\end{aligned}$$

Justification of condition #3

$$\sum_{R \in \mathbf{R}} c_R \, \mathbb{I}((i,j) \in R) = 1 \qquad \text{for all } (i,j) \in E\,.$$

- neglect entropy (e.g. suppose $\psi_{ij}(x_i, x_j) = e^{\beta\, \theta_{ij}(x_i,x_j)}$, $\beta \to \infty$)
- suppose $b^*_R(x_R)$ are true marginals, i.e. $b^*_R(x_R) = \sum_{x_{V \setminus V(R)}} b^*(X)$
- then the region based approximation correctly recovers the energy

$$\begin{aligned}
\sum_{R \in \mathbf{R}} c_R \mathbb{F}_R(b^*_R) &= \beta \sum_{R \in \mathbf{R}} c_R \sum_{x_R} b^*_R(x_R) \sum_{(ij) \in E(R)} \theta_{ij}(x_i, x_j) + O_\beta(1) \\
&= \beta \sum_{R \in \mathbf{R}} c_R \sum_{(ij) \in E(R)} \mathbb{E}_{b^*_{ij}}[\theta_{ij}(X_i, X_j)] + O_\beta(1) \\
&= \beta \sum_{(ij) \in E} \left\{ \sum_{R \in \mathbf{R}} c_R \, \mathbb{I}((i,j) \in R) \right\} \mathbb{E}_{b^*_{ij}}[\theta_{ij}(X_i, X_j)] + O_\beta(1) \\
&= \beta \sum_{(ij) \in E} \mathbb{E}_{b^*_{ij}}[\theta_{ij}(X_i, X_j)] + O_\beta(1)
\end{aligned}$$

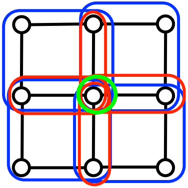
How should the regions be chosen?

- **Cluster variational method (Kikuchi approximations):**
  - First, choose a basic set of clusters (with $c_R = 1$)
  - Then, add all intersections of those basic clusters with

$$c_R = 1 - \sum_{R' \in \text{ancestor of R}} c'_R$$

  - Continue until all intersections are included

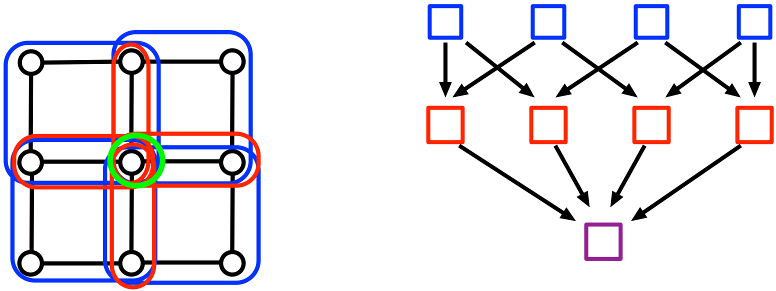
  - the above choice of c_R ensures that the **vertex counting condition** is satisfied, i.e. $\sum_{R \in \mathbf{R}} \mathbb{I}(i \in R) = 1$
  - an example with a choice of a basic set of $\{(x_1, x_2, x_4, x_5), (x_2, x_3, x_5, x_6), (x_4, x_5, x_7, x_8), (x_5, x_6, x_8, x_9)\}$



- larger basic regions give better approximations
- for pairwise MRFs, Bethe free energy has the correct energy term
- Region based methods improve in giving the increasingly accurate entropy term as clusters become larger

The Region Graph

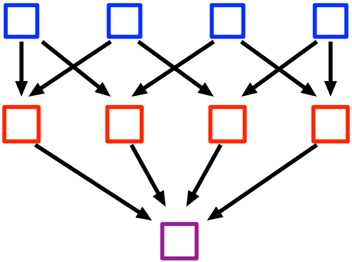
- given a collection of regions $\mathbf{R}$, how do you compute the (consistent) coefficients?
- **region graph** is a directed acyclic graph where an edge from R to R' may exist if $R' \subseteq R$
- child, parent, ancestor, descendant
- region graph is not unique



The Region Graph

- given a region graph, the weights of the regions can be computed according to

$$c_R = 1 - \sum_{R' \in \text{ANCESTORS}(R)} c_R$$



$$c_{\text{loop}} = +1$$

$$c_{\text{edge}} = -1$$

$$c_{\text{vert}} = +1$$

- region based free energy is exact if the corresponding region graph has no (undirected) cycles and the weights c_R are valid
- in general, how to generate a good region graph is still open

Generalized belief propagation

$$\begin{array}{ll} \text{maximize} & \mathbb{F}_{\mathbf{R}}(b_{\mathbf{R}}) = \sum_{R\in\mathbf{R}} c_R\,\mathbb{F}_R(b_R) \\ \text{subject to} & \sum_{x_{R\setminus R'}} b_R(x_R) = b_{R'}(x_{R'}), \quad \forall R \to R' \end{array}$$

- we form the Lagrangian

$$\mathcal{L}(\{b_R\},\{\lambda_{R\to R'}\}) = \mathbb{F}_{\mathbf{R}}(b_{\mathbf{R}}) - \sum_{R\to R'}\sum_{x_{R'}}\left\{\lambda_{R\to R'}(x_{R'})\left(\sum_{x_{R\setminus R'}} b_R(x_R) - b_{R'}(x_{R'})\right)\right\}$$

- setting derivative to zero

$$\nabla_{b_R(x_R)}\mathcal{L}(\{b_R\},\{\lambda_{R\to R'}\}) \quad = \quad 0$$

- setting $\nabla_{b_R(x_R)}\mathcal{L}(\{b_R\},\{\lambda_{R\to R'}\}) = 0$ gives an marginal computation rule for **generalized belief propagation** algorithm

$$b_R(x_R) \propto \prod_{(i,j)\in E_R} \psi_{ij}(x_i,x_j) \prod_{P\in\mathcal{P}(R)} \nu_{P\to R}(x_R) \prod_{D\in\mathcal{D}(R)} \prod_{P'\in\mathcal{P}(D)\setminus R,\mathcal{D}(R)} \nu_{P'\to D}(x_D)$$

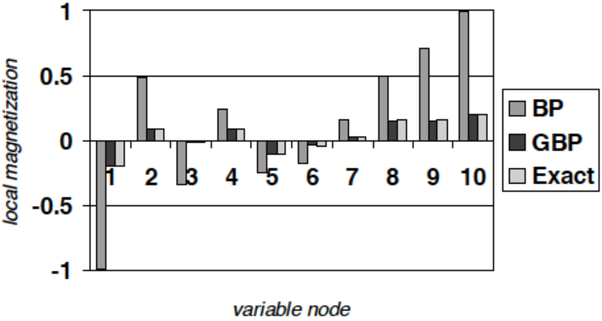
- each consistency constraint $b_R(x_R) = \sum_{x_{P\setminus R}} b_P(x_R, x_{P\setminus R})$ gives message update rule

$$\nu_{P\to R}(x_R) \propto \frac{\sum_{x_{P\setminus R}} \prod_{(i,j)\in E_R} \psi_{ij}(x_i,x_j) \prod_{(I,J)\in\mathcal{N}(P,R)} \nu_{I\to J}(x_J)}{\prod_{(I,J)\in\mathcal{D}(P,R)} \nu_{I\to J}(x_J)}$$

  - $\mathcal{P}(R) = \{$ parent of $R\}$
  - $\mathcal{D}(R) = \{$ all descendants of $R\}$
  - $\mathcal{E}(R) = R \cup \mathcal{D}(R)$
  - $\mathcal{N}(P,R) = \{I \to J \,:\, J \in \mathcal{E}(P)\setminus\mathcal{E}(R), I \notin \mathcal{E}(P)\}$
  - $\mathcal{D}(P,R) = \{I \to J \,:\, J \in \mathcal{E}(R), I \in \mathcal{E}(P)\setminus\mathcal{E}(R)\}$

- GBP fixed points are region-based free energy stationary points

Was it worth it?



10×10 Ising model with random potentials [Yedidia et al. 2003]
2×2 overlapping clusters are used with clustering variational method
GBP improves over BP significantly

Upper bound using tree-reweighted belief propagation

- consider all spanning trees of G
- each spanning tree $\tau_k = (V, E_k)$ has its own compatibility functions $\{\psi_{ij}^{(k)}\}_{(i,j)\in E_k}$ and a weight c_k such that

$$\begin{aligned} \sum_k c_k &= 1 \\ \log \psi_{ij}(x_i, x_j) &= \sum_k c_k \log \psi_{ij}^{(k)}(x_i, x_j) \end{aligned}$$

- decomposing the energy

$$\begin{aligned} \mathbb{E}_b\Big[- \sum_{(i,j)\in E} \log \psi_{ij}(x_i, x_j) \Big] &= \mathbb{E}_b\Big[- \sum_{(i,j)\in E} \sum_k c_k \log \psi_{ij}^{(k)}(x_i, x_j) \Big] \\ &= \sum_k c_k \underbrace{\mathbb{E}_b\Big[- \sum_{(i,j)\in E_k} \log \psi_{ij}^{(k)}(x_i, x_j) \Big]}_{\text{expectation over a tree } E_k} \end{aligned}$$

- from Gibbs variational principle

$$
\begin{aligned}
\log Z &= \sup_{b \in \mathsf{M}(\mathcal{X}^V)} \Big\{ \mathbb{E}_b \Big[\sum_{(i,j)\in E} \log \psi_{ij}(x_i, x_j) \Big] + H(b) \Big\} \\
&= \sup_{b \in \mathsf{M}(\mathcal{X}^V)} \Big\{ \sum_k c_k \Big\{ \mathbb{E}_b \Big[\sum_{(i,j)\in E_k} \log \psi^{(k)}_{ij}(x_i, x_j) \Big] + H(b) \Big\} \Big\} \\
&\leq \sum_k c_k \sup_{b^{(k)} \in \mathsf{M}(\mathcal{X}^V)} \Big\{ \mathbb{E}_{b^{(k)}} \Big[\sum_{(i,j)\in E_k} \log \psi^{(k)}_{ij}(x_i, x_j) \Big] + H(b) \Big\} \\
&= \sum_k c_k \underbrace{\sup_{b^{(k)} \in \mathrm{LOC}(\tau_k)} \Big\{ \mathbb{E}_{b^{(k)}} \Big[\sum_{(i,j)\in E_k} \log \psi^{(k)}_{ij}(x_i, x_j) \Big] + H(b) \Big\}}_{\text{can be solved exactly using BP}}
\end{aligned}
$$

- to get the tightest upper bound, we want to minimize the right-hand side over $\{c_k\}$ and $\{\psi^{(k)}_{ij}\}$
- the number of spanning trees can explode
- all these loose ends are resolved in [Wainwright, Jaakkola, Willsky, 2003]

Exponential families

- given a *finite* space $\mathcal{X}^V$ and a collection of functions

$$\begin{aligned} T : \mathcal{X}^V &\rightarrow \mathbb{R}^m , \\ x &\mapsto T(x) = (T_1(x), \ldots, T_m(x)) \, . \end{aligned}$$

- the corresponding **exponential family** is a family of distributions parametrized by a vector θ such that $\{\mu_\theta \, : \, \theta \in \mathbb{R}^m\}$ where

$$\mu_\theta(x) = \frac{1}{Z(\theta)} \exp\left\{ \langle \theta, T(x) \rangle \right\} , \qquad F(\theta) = \log Z(\theta)$$

where $\langle x, y \rangle = \sum_i x_i y_i$ denotes the inner product

Basic properties of exponential families

$$F(\theta) \quad = \quad \log\left(\sum_x e^{\sum_{i=1}^m \theta_i T_i(x)}\right)$$

(1) $\theta \mapsto F(\theta)$ is convex [log-sum-exps are convex]

(2) $\nabla_\theta F(\theta) = \sum_x \frac{e^{\langle\theta, T(x)\rangle}}{Z(\theta)}[T_1(x), \ldots, T_m(x)]^T = \mathbb{E}_\theta\{T(x)\} \equiv \tau(\theta)$

(3) $\nabla^2_\theta F(\theta) = \mathrm{Cov}_\theta\{T(x); T(x)\}$

(4) define a polytope

$$\begin{aligned} \mathrm{MARG}(T) \quad &\equiv \quad \mathrm{conv}\Big(\{T(x) : \; x \in \mathcal{X}^V\}\Big) \\ &= \quad \Big\{\mathbb{E}_\nu[T(x)] \; : \; \nu \in \mathrm{M}(\mathcal{X}^V)\Big\}, \text{ and} \\ \overline{\mathrm{Image}}(\tau) \quad &= \quad \mathrm{closure}\Big(\{\mathbb{E}_\theta[T(x)] \; : \; \theta \in \mathbb{R}^m\}\Big) \end{aligned}$$

then exponential families allow to realize any point in the interior of $\mathrm{MARG}(T)$

$$\overline{\mathrm{Image}(\tau)} = \mathrm{MARG}(T)$$

Proofs

(1), **(2)**, **(3):** exercises
(4): a bit more difficult

Claim 1: A closed convex set is the closure of its relative interior. [Hint: Assume the set has full dimension. Each point has a cone of full dimension around it.]

Claim 2: Let $\tau_* \in \mathrm{relint}(\mathrm{MARG}(T))$. Then $\tau_* = \mathbb{E}_{\nu_*}\{T(x)\}$ for some ν_* s.t. $\nu_*(x) > 0$ for all $x \in \mathcal{X}^V$. [Hint: Consider the set of signed weigths ν such that $\sum_x \nu(x) T(x) = \tau_*$. If the claim was false, it would be tangent to the simplex.]

Claim 3: There exists $\theta_* \in \mathbb{R}^m$ such that $\mathbb{E}_{\theta_*}\{T(x)\} = \mathbb{E}_{\nu_*}\{T(x)\}$.

Proof of Claim 3

Wlog $\{1, T_1, \dots, T_m\}$ linearly independent. Consider

$$\begin{aligned} F(\theta;\tau_*) &\equiv F(\theta) - \langle \tau_*, \theta\rangle \\ &= \log\Big\{ \sum_{x\in\mathcal{X}^V} \exp\big(\langle\theta, T(x)\rangle\big)\Big\} - \mathbb{E}_{\nu_*}\{\langle\theta, T(x)\rangle\} \end{aligned}$$

- $F(\,\cdot\,;\tau_*) : \mathbb{R}^m \to \mathbb{R}$ is differentiable and convex.
- If θ_* is a stationary point, then $\mathbb{E}_{\theta_*}\{T(x)\} = \mathbb{E}_{\nu_*}\{T(x)\}$.
- As $\theta \to \infty$, $F(\theta;\tau_*) \to \infty$.

Implies the thesis.

As $\theta \to \infty$, $F_{\tau_*}(\theta) \to \infty$

Let $\theta = \beta\, v$, $\beta \in \mathbb{R}_+$

$$\begin{aligned} F(\theta;\tau_*) &= \log\Big\{ \sum_{x\in\mathcal{X}^V} \exp\big(\langle\theta, T(x)\rangle\big)\Big\} - \mathbb{E}_{\nu_*}\{\langle\theta, T(x)\rangle\} \\ &\geq \beta\Big[\max_x \langle v, T(x)\rangle - \mathbb{E}_{\nu_*}\{\langle v, T(x)\rangle\}\Big] \end{aligned}$$

and $[\,\dots\,] > 0$ strictly because $\nu_*(x) > 0$ for all x.

Duality structure

$$
\begin{aligned}
F_*(\tau) &\equiv \inf_{\theta \in \mathbb{R}^m} \{ F(\theta) - \langle \tau, \theta \rangle \}\,, \\
F_* : &\quad \mathrm{MARG}(T) \to \mathbb{R}\,, \qquad \text{concave.}
\end{aligned}
$$

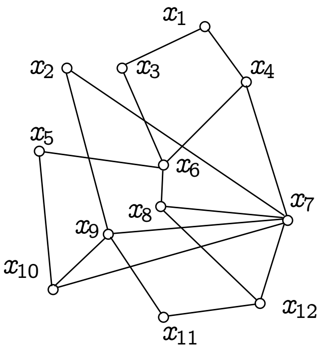
$$
\begin{aligned}
F(\theta) &\equiv \sup_{\tau \in \mathrm{MARG}(T)} \{ F_*(\tau) + \langle \tau, \theta \rangle \}\,, \\
F : &\quad \mathbb{R}^m \to \mathbb{R}\,, \qquad \text{convex.}
\end{aligned}
$$

Duality structure

$$\begin{aligned} F_*(\tau) &\equiv \inf_{\theta \in \mathbb{R}^m} \{ F(\theta) - \langle \tau, \theta \rangle \}, \\ F_* : &\quad \mathrm{MARG}(T) \to \mathbb{R}, \qquad \text{concave.} \end{aligned}$$

$$\begin{aligned} F(\theta) &\equiv \sup_{\tau \in \mathrm{MARG}(T)} \{ F_*(\tau) + \langle \tau, \theta \rangle \}, \\ F : &\quad \mathbb{R}^m \to \mathbb{R}, \qquad \text{convex.} \end{aligned}$$

Let's apply all this



$G = (V, E)$, $V = [n]$, $x = (x_1, \ldots, x_n)$, $x_i \in \mathcal{X}$,

$$
\begin{aligned}
T_{i,\xi}(x) &= \mathbb{I}(x_i = \xi)\,, & i \in V, \xi \in \mathcal{X}, \\
T_{ij,\xi_1,\xi_2}(x) &= \mathbb{I}(x_i = \xi_1)\,\mathbb{I}(x_j = \xi_2)\,, & (i,j) \in E, \xi_1, \xi_2 \in \mathcal{X},
\end{aligned}
$$

overcomplete!

The exponential family

$$\mu_\theta(x) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{(i,j)\in E, \xi_1,\xi_2\in\mathcal{X}} \theta_{ij}(\xi_1,\xi_2)\, T_{ij\xi_1\xi_2}(x) + \sum_{i\in V, \xi\in\mathcal{X}} \theta_i(\xi) T_{i\xi}(x) \right\}$$
$$= \frac{1}{Z(\theta)} \exp\left\{ \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j) + \sum_{i\in V} \theta_i(x_i) \right\}$$

(General pairwise model)

The τ parameters

$$\begin{aligned} b_i(\xi) &= \mathbb{E}_\theta\{T_i(\xi)\} = \mu_\theta(x_i = \xi), & \text{for } i \in V, \\ b_{ij}(\xi_1,\xi_2) &= \mathbb{E}_\theta\{T_{ij}(\xi_1,\xi_2)\} = \mu_\theta(x_i = \xi_1, x_j = \xi_2), & \text{for } (i,j) \in E. \end{aligned}$$

The exponential family

$$
\begin{aligned}
\mu_\theta(x) &= \frac{1}{Z(\theta)} \exp\left\{ \sum_{(i,j)\in E, \xi_1,\xi_2 \in \mathcal{X}} \theta_{ij}(\xi_1,\xi_2)\, T_{ij\xi_1\xi_2}(x) + \sum_{i\in V, \xi\in\mathcal{X}} \theta_i(\xi) T_{i\xi}(x) \right\} \\
&= \frac{1}{Z(\theta)} \exp\left\{ \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j) + \sum_{i\in V} \theta_i(x_i) \right\}
\end{aligned}
$$

(General pairwise model)

The τ parameters

$$
\begin{aligned}
b_i(\xi) &= \mathbb{E}_\theta\{T_i(\xi)\} = \mu_\theta(x_i = \xi), && \text{for } i \in V, \\
b_{ij}(\xi_1,\xi_2) &= \mathbb{E}_\theta\{T_{ij}(\xi_1,\xi_2)\} = \mu_\theta(x_i = \xi_1, x_j = \xi_2), && \text{for } (i,j) \in E.
\end{aligned}
$$

The duality structure

$$
\begin{aligned}
& F(\theta) \leftrightarrow F_*(b)\,, \\
F_* : \quad & \mathrm{MARG}(G) \to \mathbb{R}\,.
\end{aligned}
$$

We want to evaluate at $\Phi = F(\theta_* = \log \psi)$':

$$
\begin{aligned}
\Phi &= \sup_{b \in \mathrm{MARG}(G)} \Big\{ F_*(b) + \langle \theta_*, b \rangle \Big\} \\
&= \mathrm{Entropy} + \mathrm{Energy}
\end{aligned}
$$

New interpretation

Bethe entropy is an approximate expression for $F_*(b)$.

The duality structure

$$
\begin{aligned}
& F(\theta) \leftrightarrow F_*(b)\,, \\
F_* : \quad & \mathrm{MARG}(G) \to \mathbb{R}\,.
\end{aligned}
$$

We want to evaluate at $\Phi = F(\theta_* = \log \psi)$':

$$
\begin{aligned}
\Phi &= \sup_{b \in \mathrm{MARG}(G)} \Big\{ F_*(b) + \langle \theta_*, b \rangle \Big\} \\
&= \mathrm{Entropy} + \mathrm{Energy}
\end{aligned}
$$

New interpretation

Bethe entropy is an approximate expression for $F_*(b)$.

The duality structure

$$\begin{aligned} & F(\theta) \leftrightarrow F_*(b)\,, \\ F_*: \quad & \mathrm{MARG}(G) \to \mathbb{R}\,. \end{aligned}$$

We want to evaluate at $\Phi = F(\theta_* = \log\psi)$':

$$\begin{aligned} \Phi &= \sup_{b \in \mathrm{MARG}(G)} \left\{ F_*(b) + \langle \theta_*, b \rangle \right\} \\ &= \mathrm{Entropy} + \mathrm{Energy} \end{aligned}$$

New interpretation

Bethe entropy is an approximate expression for $F_*(b)$.

Interpretation works fine on trees

Proposition

If G is a tree, then $\mathrm{MARG}(G) = \mathrm{LOC}(G)$ and

$$F_*(b) = \sum_{i \in V} H(b_i) - \sum_{(i,j) \in E} I(b_{ij}) = \mathbb{F}_{\psi=1}(b)$$

As a consequence, $\mathbb{F} : \mathrm{LOC}(G) \to \mathbb{R}$ is concave.

Proof: Exercise.

Interpretation works fine on trees

Proposition

If G is a tree, then $\mathrm{MARG}(G) = \mathrm{LOC}(G)$ and

$$F_*(b) = \sum_{i \in V} H(b_i) - \sum_{(i,j) \in E} I(b_{ij}) = \mathbb{F}_{\psi=1}(b)$$

As a consequence, $\mathbb{F} : \mathrm{LOC}(G) \to \mathbb{R}$ is concave.

Proof: Exercise.

Interpretation works fine on trees

Proposition

If G is a tree, then $\mathrm{MARG}(G) = \mathrm{LOC}(G)$ and

$$F_*(b) = \sum_{i \in V} H(b_i) - \sum_{(i,j) \in E} I(b_{ij}) = \mathbb{F}_{\psi=1}(b)$$

As a consequence, $\mathbb{F} : \mathrm{LOC}(G) \to \mathbb{R}$ is concave.

Proof: Exercise.

What about general graphs?

Write G as a convex combination of trees.

Abuse: I will use T to denote trees, not functions.

Convex combinations

$$\mathcal{T}(G) \quad = \quad \{ \text{ spanning trees in } G \ \},$$

$$\begin{aligned} \rho : \mathcal{T}(G) \quad & \rightarrow [0,1]\,, \\ T \quad & \mapsto \rho_T\,, \qquad \text{weights}\,, \end{aligned}$$

$$\sum_{T \in \mathcal{T}(G)} \rho_T = 1\,,$$

$$\sum_{T \in \mathcal{T}(G)} \rho_T\, \theta^T = \theta\,.$$

Convex combinations

$$\mathcal{T}(G) = \{ \text{ spanning trees in } G \},$$

$$\begin{aligned} \rho : \mathcal{T}(G) &\to [0,1], \\ T &\mapsto \rho_T, \qquad \text{weights}, \end{aligned}$$

$$\sum_{T \in \mathcal{T}(G)} \rho_T = 1,$$

$$\sum_{T \in \mathcal{T}(G)} \rho_T \, \theta^T = \theta.$$

Convex combinations

$$\mathcal{T}(G) = \{ \text{ spanning trees in } G \},$$

$$\begin{aligned} \rho : \mathcal{T}(G) &\to [0,1], \\ T &\mapsto \rho_T, \qquad \text{weights}, \end{aligned}$$

$$\sum_{T \in \mathcal{T}(G)} \rho_T = 1,$$

$$\sum_{T \in \mathcal{T}(G)} \rho_T \, \theta^T = \theta.$$

Convex combinations

$$\begin{aligned} \Phi &= F(\theta) = F\big(\sum_{T \in \mathcal{T}(G)} \rho_T \theta^T\big) \\ &\leq \sum_{T \in \mathcal{T}(G)} \rho_T \, F(\theta^T) \end{aligned}$$

- Fix weigths ρ_T.
- Minimize over θ^T (convex!)

Problem: Exponentially many spanning trees.

Convex combinations

$$\begin{aligned}\Phi &= F(\theta) = F\Big(\sum_{T \in \mathcal{T}(G)} \rho_T \theta^T\Big) \\ &\leq \sum_{T \in \mathcal{T}(G)} \rho_T \, F(\theta^T)\end{aligned}$$

- Fix weigths ρ_T.
- Minimize over θ^T (convex!)

Problem: Exponentially many spanning trees.

Convex combinations

$$
\begin{aligned}
\Phi &= F(\theta) = F\Big(\sum_{T \in \mathcal{T}(G)} \rho_T \theta^T\Big) \\
&\leq \sum_{T \in \mathcal{T}(G)} \rho_T \, F(\theta^T)
\end{aligned}
$$

- Fix weigths ρ_T.
- Minimize over θ^T (convex!)

Problem: Exponentially many spanning trees.

Minimization over $(\theta^T)_{T \in \mathcal{T}(G)}$

$$
\begin{aligned}
\text{minimize} \quad & \sum_{T \in \mathcal{T}(G)} \rho_T \, F(\theta^T)\,, \\
\text{subject to} \quad & \sum_{T \in \mathcal{T}(G)} \rho_T \theta^T_{ij}(x_i, x_j) = \theta_{ij}(x_i, x_j)\,, \\
& \sum_{T \in \mathcal{T}(G)} \rho_T \theta^T_{i}(x_i) = \theta_{i}(x_i)\,.
\end{aligned}
$$

Convex Problem

Minimization over $(\theta^T)_{T \in \mathcal{T}(G)}$

$$
\begin{aligned}
\text{minimize} \quad & \sum_{T \in \mathcal{T}(G)} \rho_T \, F(\theta^T) \,, \\
\text{subject to} \quad & \sum_{T \in \mathcal{T}(G)} \rho_T \theta^T_{ij}(x_i, x_j) = \theta_{ij}(x_i, x_j) \,, \\
& \sum_{T \in \mathcal{T}(G)} \rho_T \theta^T_{i}(x_i) = \theta_{i}(x_i) \,.
\end{aligned}
$$

Convex Problem

Lagrangian

$$
\begin{aligned}
\mathcal{L}((\theta^T), b) = & \sum_T \rho_T \, F(\theta^T) \\
& - \sum_{(ij)\in E} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \Big\{ \sum_T \rho_T \theta_{ij}^T(x_i, x_j) - \theta_{ij}(x_i, x_j) \Big\} \\
& - \sum_{i\in V} \sum_{x_i} b_i(x_i) \Big\{ \sum_T \rho_T \theta_i^T(x_i) - \theta_i(x_i) \Big\}
\end{aligned}
$$

$$
= \sum_T \rho_T \Big\{ F(\theta^T) - \langle b, \theta^T \rangle \Big\} + \langle b, \theta \rangle
$$

Separable in θ^T

Lagrangian

$$
\begin{aligned}
\mathcal{L}((\theta^T), b) = & \sum_T \rho_T \, F(\theta^T) \\
& - \sum_{(ij)\in E} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \Big\{ \sum_T \rho_T \theta_{ij}^T(x_i, x_j) - \theta_{ij}(x_i, x_j) \Big\} \\
& - \sum_{i \in V} \sum_{x_i} b_i(x_i) \Big\{ \sum_T \rho_T \theta_i^T(x_i) - \theta_i(x_i) \Big\}
\end{aligned}
$$

$$
= \sum_T \rho_T \Big\{ F(\theta^T) - \langle b, \theta^T \rangle \Big\} + \langle b, \theta \rangle
$$

Separable in θ^T

Lagrangian

$$
\begin{aligned}
\mathcal{L}((\theta^T), b) = & \sum_T \rho_T \, F(\theta^T) \\
& - \sum_{(ij)\in E} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \Big\{ \sum_T \rho_T \theta_{ij}^T(x_i, x_j) - \theta_{ij}(x_i, x_j) \Big\} \\
& - \sum_{i \in V} \sum_{x_i} b_i(x_i) \Big\{ \sum_T \rho_T \theta_i^T(x_i) - \theta_i(x_i) \Big\}
\end{aligned}
$$

$$
= \sum_T \rho_T \Big\{ F(\theta^T) - \langle b, \theta^T \rangle \Big\} + \langle b, \theta \rangle
$$

Separable in θ^T

Lagrangian

$$\min_{(\theta^T)} \mathcal{L}((\theta^T), b) = \sum_T \rho_T F_*(b;\theta) + \langle b, \theta \rangle$$

$$= \sum_T \rho_T \Big\{ \sum_{i \in V} H(b_i) - \sum_{(ij) \in E(T)} I(b_{ij}) \Big\} + \langle b, \theta \rangle$$

$$= \sum_{i \in V} H(b_i) \Big\{ \sum_{T: i \in V} \rho_T \Big\} - \sum_{(i,j) \in V} I(b_{ij}) \Big\{ \sum_{T: (i,j) \in E(T)} \rho_T \Big\} + \langle b, \theta \rangle$$

$$= \sum_{i \in V} H(b_i) - \sum_{(i,j) \in V} \rho(ij) I(b_{ij}) + \langle b, \theta \rangle$$

Lagrangian

$$
\begin{aligned}
\min_{(\theta^T)} \mathcal{L}((\theta^T), b) &= \sum_T \rho_T F_*(b;\theta) + \langle b, \theta \rangle \\
&= \sum_T \rho_T \Big\{ \sum_{i \in V} H(b_i) - \sum_{(ij) \in E(T)} I(b_{ij}) \Big\} + \langle b, \theta \rangle \\
&= \sum_{i \in V} H(b_i) \Big\{ \sum_{T:\, i \in V} \rho_T \Big\} - \sum_{(i,j) \in V} I(b_{ij}) \Big\{ \sum_{T:\, (i,j) \in E(T)} \rho_T \Big\} + \langle b, \theta \rangle \\
&= \sum_{i \in V} H(b_i) - \sum_{(i,j) \in V} \rho(ij) I(b_{ij}) + \langle b, \theta \rangle
\end{aligned}
$$

Lagrangian

$$
\begin{aligned}
\min_{(\theta^T)} \mathcal{L}((\theta^T), b) &= \sum_T \rho_T F_*(b;\theta) + \langle b, \theta \rangle \\
&= \sum_T \rho_T \Big\{ \sum_{i\in V} H(b_i) - \sum_{(ij)\in E(T)} I(b_{ij}) \Big\} + \langle b, \theta \rangle \\
&= \sum_{i\in V} H(b_i) \Big\{ \sum_{T:\, i\in V} \rho_T \Big\} - \sum_{(i,j)\in V} I(b_{ij}) \Big\{ \sum_{T:\, (i,j)\in E(T)} \rho_T \Big\} + \langle b, \theta \rangle \\
&= \sum_{i\in V} H(b_i) - \sum_{(i,j)\in V} \rho(ij) I(b_{ij}) + \langle b, \theta \rangle
\end{aligned}
$$

Lagrangian

$$\min_{(\theta^T)} \mathcal{L}((\theta^T), b) = \sum_T \rho_T F_*(b;\theta) + \langle b, \theta \rangle$$

$$= \sum_T \rho_T \Big\{ \sum_{i \in V} H(b_i) - \sum_{(ij) \in E(T)} I(b_{ij}) \Big\} + \langle b, \theta \rangle$$

$$= \sum_{i \in V} H(b_i) \Big\{ \sum_{T:\, i \in V} \rho_T \Big\} - \sum_{(i,j) \in V} I(b_{ij}) \Big\{ \sum_{T:\, (i,j) \in E(T)} \rho_T \Big\} + \langle b, \theta \rangle$$

$$= \sum_{i \in V} H(b_i) - \sum_{(i,j) \in V} \rho(ij) I(b_{ij}) + \langle b, \theta \rangle$$

Tree-reweighted free energy

$$\mathbb{F}_{\mathsf{TRW}}(b) = \sum_{i \in V} H(b_i) - \sum_{(i,j) \in V} \rho(ij) I(b_{ij}) + \langle b, \theta \rangle$$

Compare with Bethe free energy

$$\mathbb{F}(b) \sum_{i \in V} H(b_i) - \sum_{(i,j) \in V} I(b_{ij}) + \langle b, \theta \rangle$$

$\rho(i,j) = 0$ Obviously concave upper bound.

$\rho(i,j) = 1$ Bethe free energy.

Tree-reweighted free energy

$$\mathbb{F}_{\mathsf{TRW}}(b) = \sum_{i \in V} H(b_i) - \sum_{(i,j) \in V} \rho(ij) I(b_{ij}) + \langle b, \theta \rangle$$

Compare with Bethe free energy

$$\mathbb{F}(b) \sum_{i \in V} H(b_i) - \sum_{(i,j) \in V} I(b_{ij}) + \langle b, \theta \rangle$$

$\rho(i,j) = 0$ Obviously concave upper bound.

$\rho(i,j) = 1$ Bethe free energy.

Tree-reweighted free energy

$$\mathbb{F}_{\mathsf{TRW}}(b) = \sum_{i \in V} H(b_i) - \sum_{(i,j) \in V} \rho(ij) I(b_{ij}) + \langle b, \theta \rangle$$

Compare with Bethe free energy

$$\mathbb{F}(b) \sum_{i \in V} H(b_i) - \sum_{(i,j) \in V} I(b_{ij}) + \langle b, \theta \rangle$$

$\rho(i,j) = 0$ Obviously concave upper bound.

$\rho(i,j) = 1$ Bethe free energy.

Edge weights

$\rho = (\, \rho(e) \; : \; e \in E)$
Intepretation

$$\rho(e) = \mathbb{P}_\rho\{e \in E(T)\}\,, \qquad \mathbb{P}_\rho(T) = \rho_T\,.$$

Spanning-Tree polytope

$$\begin{aligned} \sum_{(i,j)\in E} \rho(i,j) &= |V| - 1\,, \\ \sum_{(i,j)\in E(U)} \rho(i,j) &\leq |U| - 1\,, \qquad \text{for all } U \subseteq V\,. \end{aligned}$$

Edge weights

$\rho = (\,\rho(e)\;:\;e \in E)$
Intepretation

$$\rho(e) = \mathbb{P}_\rho\{e \in E(T)\}\,, \qquad \mathbb{P}_\rho(T) = \rho_T\,.$$

Spanning-Tree polytope

$$\begin{aligned}\sum_{(i,j)\in E} \rho(i,j) &= |V| - 1\,, \\ \sum_{(i,j)\in E(U)} \rho(i,j) &\leq |U| - 1\,, \qquad \text{for all } U \subseteq V\,.\end{aligned}$$

Example

k-regular graph

$$|V| = n, \qquad |E| = \frac{nk}{2}\,.$$

Take all the weights equal (not necessarily ok, but. . .)

$$\rho(i,j) = \frac{2(n-1)}{nk} \approx \frac{2}{k}$$

For (some) models on locally tree-like graphs, $\rho(i,j) = 1$ is approximately correct $\rightarrow \Theta(n)$ error.

Example

k-regular graph

$$|V| = n, \qquad |E| = \frac{nk}{2}\,.$$

Take all the weights equal (not necessarily ok, but...)

$$\rho(i,j) = \frac{2(n-1)}{nk} \approx \frac{2}{k}$$

For (some) models on locally tree-like graphs, $\rho(i,j) = 1$ is approximately correct $\rightarrow \Theta(n)$ error.