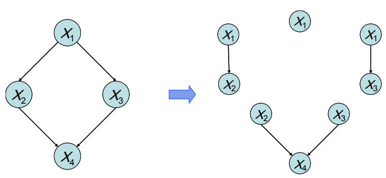
School of Computer Science
Carnegie Mellon

Probabilistic Graphical Models

Variational Inference IV:
Variational Principle II

Junming Yin

Lecture 17, March 21, 2012



Reading:

Recap: Variational Inference

- Variational formulation

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\theta^T \mu - A^*(\mu)\}$$

  - $\mathcal{M}$: the marginal polytope, difficult to characterize
  - A^*: the negative entropy function, no explicit form

- Mean field method: non-convex inner bound and exact form of entropy

- Bethe approximation and loopy belief propagation: polyhedral outer bound and non-convex Bethe approximation

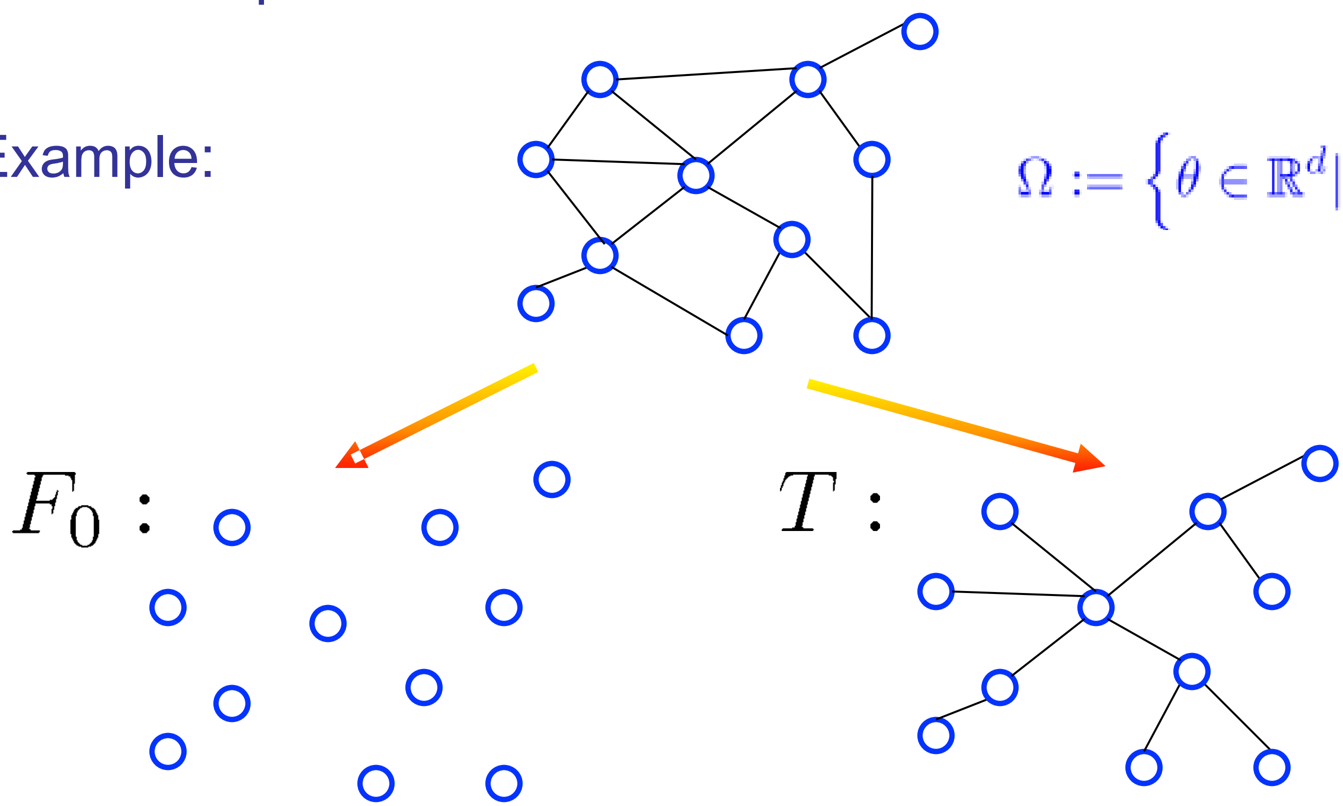
Mean Field Approximation

Tractable Subgraphs

- Definition: A subgraph F of the graph G is *tractable* if it is feasible to perform exact inference

- Example:

$$\Omega := \left\{\theta \in \mathbb{R}^d | A(\theta) < +\infty\right\}$$



$$\Omega(F_0) := \{\theta \in \Omega | \theta_{(s,t)} = 0, \forall (s,t) \in E\} \quad \Omega(T) := \{\theta \in \Omega | \theta_{(s,t)} = 0 \ \forall (s,t) \notin E(T)\}$$

Mean Field Methods

- For an exponential family with sufficient statistics ϕ defined on graph G, the set of realizable mean parameter set

$$\mathcal{M}(G;\phi) := \{\mu \in \mathbb{R}^d \mid \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\}$$

- For a given tractable subgraph F, a subset of mean parameters of interest

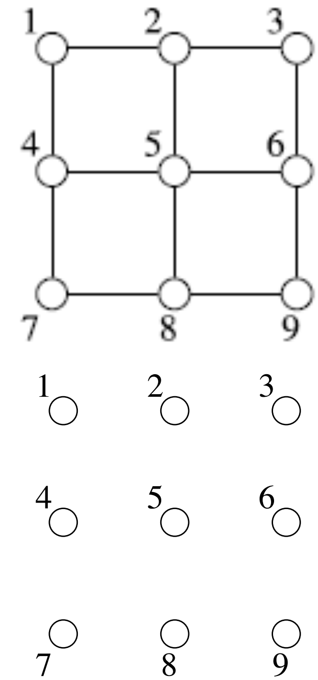
$$\mathcal{M}(F;\phi) := \{\tau \in \mathbb{R}^d \mid \tau = \mathbb{E}_\theta[\phi(X)] \text{ for some } \theta \in \Omega(F)\}$$

- Inner approximation $\mathcal{M}(F;\phi)^o \subseteq \mathcal{M}(G;\phi)^o$
- Mean field solves the relaxed problem

$$\max_{\tau \in \mathcal{M}_F(G)} \{\langle \tau, \theta \rangle - A_F^*(\tau)\}$$

  - $A_F^* = A^*\big|_{\mathcal{M}_F(G)}$ is the exact dual function restricted to $\mathcal{M}_F(G)$

Example: Naïve Mean Field for Ising Model



- Ising model in {0,1} representation

$$p(x) \propto \exp\left\{ \sum_{s \in V} x_s \theta_s + \sum_{(s,t) \in E} x_s x_t \theta_{st} \right\}$$

- Mean parameters

$$\mu_s = \mathbb{E}_p[X_s] = \mathbb{P}[X_s = 1] \quad \text{for all } s \in V, \text{ and}$$
$$\mu_{st} = \mathbb{E}_p[X_s X_t] = \mathbb{P}[(X_s, X_t) = (1,1)] \quad \text{for all } (s,t) \in E.$$

- For fully disconnected graph F,

$$\mathcal{M}_F(G) := \{\tau \in \mathbb{R}^{|V|+|E|} \mid 0 \le \tau_s \le 1, \forall s \in V, \tau_{st} = \tau_s \tau_t, \forall (s,t) \in E\}$$

- The dual decomposes into sum, one for each node

$$A_F^*(\tau) = \sum_{s \in V} [\tau_s \log \tau_s + (1 - \tau_s) \log(1 - \tau_s)]$$

Example: Naïve Mean Field for Ising Model

- Mean field problem

$$A(\theta) \geq \max_{(\tau_1,\ldots,\tau_m)\in[0,1]^m} \left\{ \sum_{s\in V} \theta_s \tau_s + \sum_{(s,t)\in E} \theta_{st} \tau_s \tau_t - A_F^*(\tau) \right\}$$

- The same objective function as in free energy based approach

- The naïve mean field update equations

$$\tau_s \leftarrow \sigma\left(\theta_s + \sum_{t\in N(s)} \theta_s \tau_t \right)$$

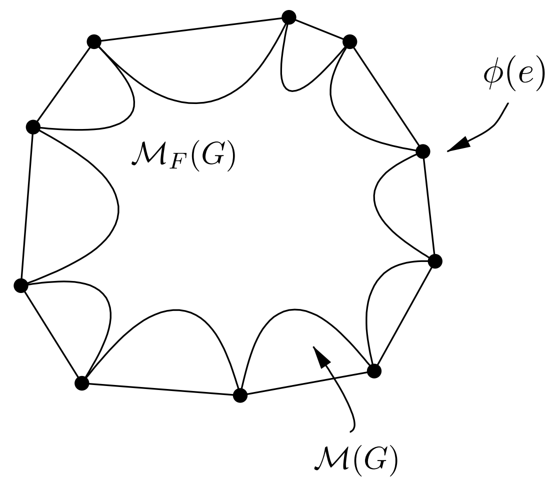
- Also yields lower bound on log partition function

Geometry of Mean Field

- Mean field optimization is always non-convex for any exponential family in which the state space $\mathcal{X}^m$ is finite

- Recall the marginal polytope is a convex hull

$$\mathcal{M}(G) = \mathrm{conv}\{\phi(e); e \in \mathcal{X}^m\}$$



- $\mathcal{M}_F(G)$ contains all the extreme points
  - If it is a strict subset, then it must be non-convex

- Example: two-node Ising model

$$\mathcal{M}_F(G) = \{0 \le \tau_1 \le 1, 0 \le \tau_2 \le 1, \tau_{12} = \tau_1 \tau_2\}$$

  - It has a parabolic cross section along $\tau_1 = \tau_2$, hence non-convex

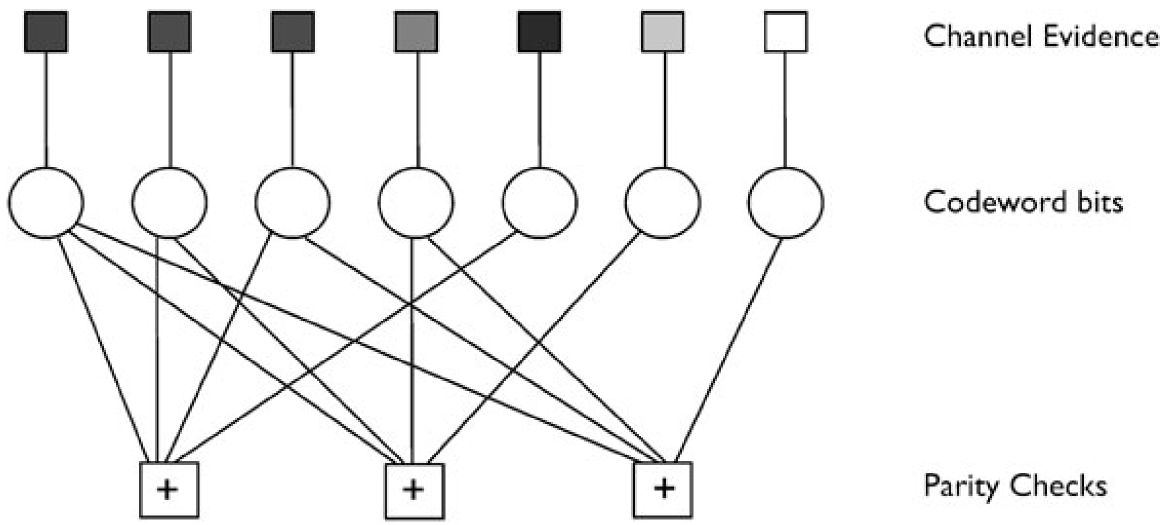
Bethe Approximation and Sum-Product

Historical Information

- Bethe (1935): a physicist who first developed the ideas related to the loopy belief propagation in the Bethe approximation; not fully appreciated outside the physics community until recently
- Gallager (1963): an electrical engineer who explored the loopy belief propagation in his work on LDPC (Low Density Parity Check) codes
- Yedidia (2001): a physicist who made an explicit connection from the loopy belief propagation to the Bethe approximation and further developed generalized belief propagation algorithm

Error Correcting Codes

- Graphical model for (7,4) Hamming code

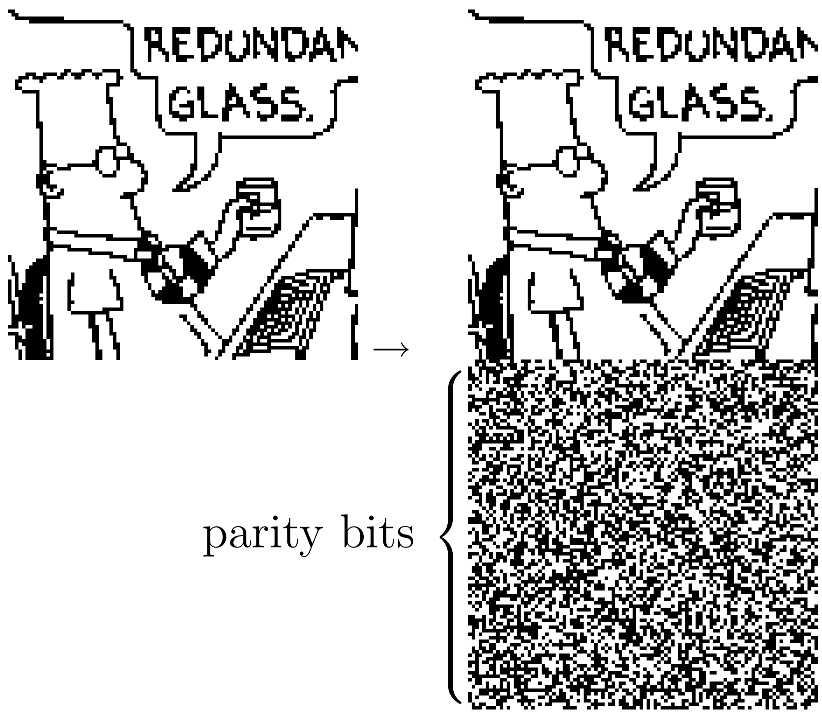


- Potential functions with hard constraint

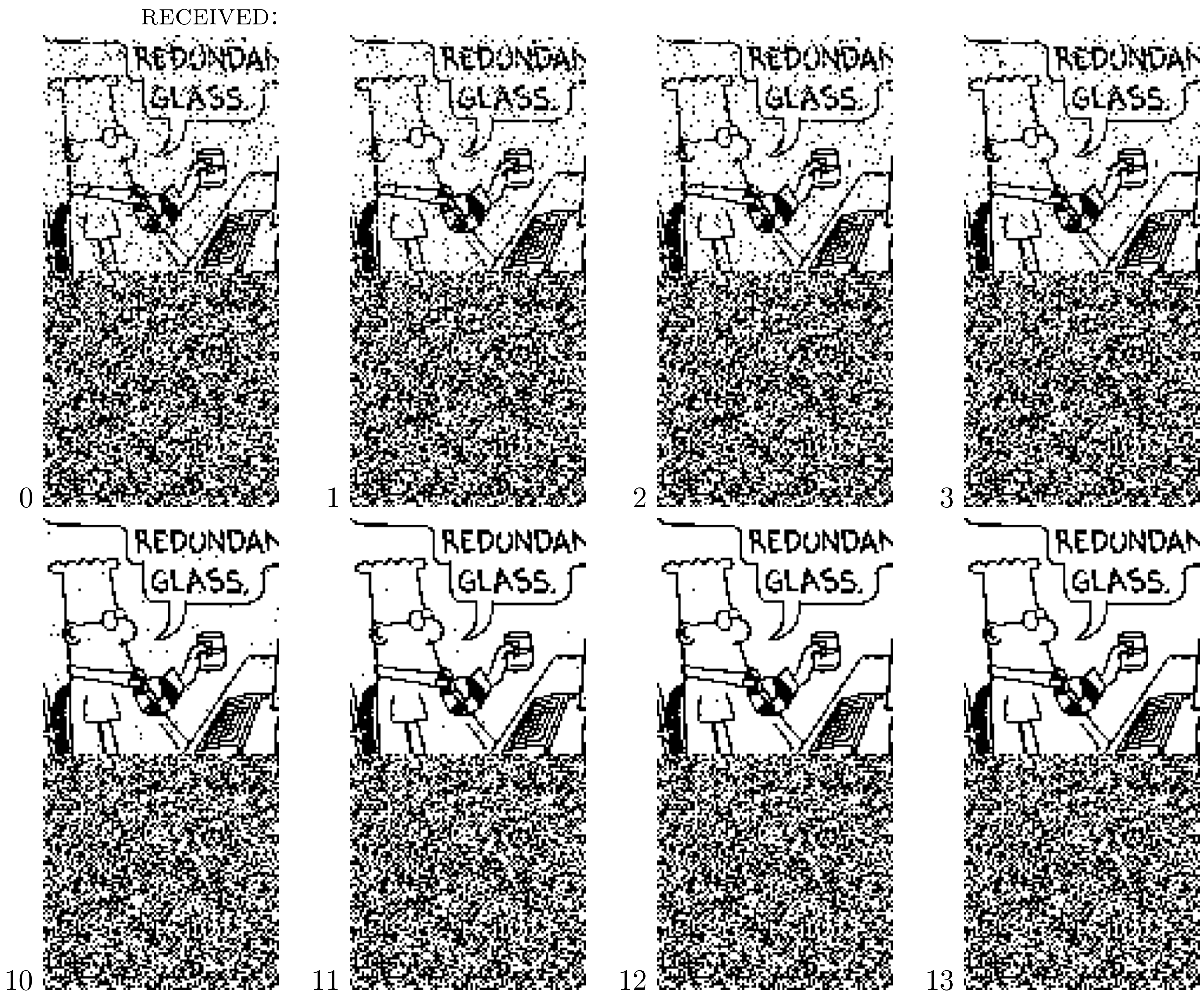
$$\psi_{stu}(x_s, x_t, x_u) := \begin{cases} 1 & \text{if } x_s \oplus x_t \oplus x_u = 1 \\ 0 & \text{otherwise.} \end{cases}$$

- Marginal probabilities = A *posterior* bit probabilities

Example of LDPC Decoding



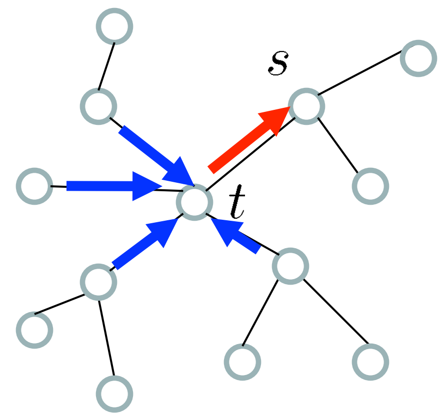
Example of LDPC Decoding



Sum-Product/Belief Propagation Algorithm

- Message passing rule:

$$M_{ts}(x_s) \leftarrow \kappa \sum_{x'_t} \left\{ \psi_{st}(x_s, x'_t)\, \psi_t(x'_t) \prod_{u \in N(t)/s} M_{ut}(x'_t) \right\}$$



- Marginals:

$$\mu_s(x_s) \;=\; \kappa\, \psi_s(x_s) \prod_{t \in N(s)} M^*_{ts}(x_s)$$

- Exact for trees, but approximate for loopy graphs (so called loopy belief propagation)

- Question:
  - How is the algorithm on trees related to variational principle?
  - What is the algorithm doing for graphs with cycles?

Tree Graphical Models

- Discrete variables $X_s \in \{0, 1, \ldots, m_s - 1\}$ on a tree $T = (V, E)$
- Sufficient statistics: $\begin{array}{ll} \mathbb{I}_j(x_s) & \text{for } s = 1, \ldots n, \quad j \in \mathcal{X}_s \\ \mathbb{I}_{jk}(x_s, x_t) & \text{for} (s,t) \in E, \quad (j,k) \in \mathcal{X}_s \times \mathcal{X}_t \end{array}$
- Exponential representation of distribution:

$$p(\mathbf{x}; \theta) \quad \propto \quad \exp\Big\{\sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t)\Big\}$$

where $\theta_s(x_s) := \sum_{j \in \mathcal{X}_s} \theta_{s;j} \mathbb{I}_j(x_s)$ $\qquad$ (and similarly for $\theta_{st}(x_s, x_t)$)

- Mean parameters are marginal probabilities:

$$\mu_{s;j} = \mathbb{E}_p[\mathbb{I}_j(X_s)] = \mathbb{P}[X_s = j] \quad \forall j \in \mathcal{X}_s, \quad \mu_s(x_s) = \sum_{j \in \mathcal{X}_s} \mu_{s;j} \mathbb{I}_j(x_s) = \mathbb{P}(X_s = x_s)$$

$$\mu_{st;jk} = \mathbb{E}_p[\mathbb{I}_{st;jk}(X_s, X_t)] = \mathbb{P}[X_s = j, X_t = k] \quad \forall (j,k) \in \mathcal{X}_s \in \mathcal{X}_t.$$

$$\mu_{st}(x_s, x_t) = \sum_{(j,k) \in \mathcal{X}_s \times \mathcal{X}_t} \mu_{st;jk} \mathbb{I}_{jk}(x_s, x_t) = \mathbb{P}(X_s = x_s, X_t = x_t)$$

Marginal Polytope for Trees

- Recall marginal polytope for general graphs

$$\mathcal{M}(G) = \{\mu \in \mathbb{R}^d \mid \exists p \text{ with marginals } \mu_{s;j}, \mu_{st;jk}\}$$

- By junction tree theorem (see Prop. 2.1 & Prop. 4.1)

$$\mathcal{M}(T) = \left\{\mu \geq 0 \mid \sum_{x_s} \mu_s(x_s) = 1, \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s)\right\}$$

- In particular, if $\mu \in \mathcal{M}(T)$, then

$$p_\mu(x) := \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}.$$

has the corresponding marginals

Decomposition of Entropy for Trees

- For trees, the entropy decomposes as

$$
\begin{aligned}
H(p(x;\mu)) &= -\sum_x p(x;\mu)\log p(x;\mu) \\
&= \sum_{s\in V} \underbrace{\Big(-\sum_{x_s}\mu_s(x_s)\log\mu_s(x_s)\Big)}_{H_s(\mu_s)} - \\
&\quad -\sum_{(s,t)\in E} \underbrace{\Big(\sum_{x_s,x_t}\mu_{st}(x_s,x_t)\log\frac{\mu_{st}(x_s,x_t)}{\mu_s(x_s)\mu_t(x_t)}\Big)}_{I_{st}(\mu_s t),\ \text{KL-Divergence}} \\
&= \sum_{s\in V} H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st})
\end{aligned}
$$

- The dual function has an explicit form $A^*(\mu) = -H(p(x;\mu))$

Exact Variational Principle for Trees

- Variational formulation

$$A(\theta) = \max_{\mu \in \mathcal{M}(T)} \left\{ \langle \theta, \mu \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}) \right\}$$

- Assign Lagrange multiplier λ_{ss} for the normalization constraint $C_{ss}(\mu) := 1 - \sum_{x_s} \mu_s(x_s) = 0$; and $\lambda_{ts}(x_s)$ for each marginalization constraint $C_{ts}(x_s; \mu) := \mu_s(x_s) - \sum_{x_t} \mu_{st}(x_s, x_t) = 0$

- The Lagrangian has the form

$$\mathcal{L}(\mu, \lambda) = \langle \theta, \mu \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}) + \sum_{s \in V} \lambda_{ss} C_{ss}(\mu)$$

$$+ \sum_{(s,t) \in E} \left[\sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t) + \sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s) \right]$$

Lagrangian Derivation

- Taking the derivatives of the Lagrangian w.r.t. μ_s and μ_{st}

$$\frac{\partial \mathcal{L}}{\partial \mu_s(x_s)} = \theta_s(x_s) - \log \mu_s(x_s) + \sum_{t \in \mathcal{N}(s)} \lambda_{ts}(x_s) + C$$

$$\frac{\partial \mathcal{L}}{\partial \mu_{st}(x_s, x_t)} = \theta_{st}(x_s, x_t) - \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} - \lambda_{ts}(x_s) - \lambda_{st}(x_t) + C'$$

- Setting them to zeros yields

$$\mu_s(x_s) \propto \exp\{\theta_s(x_s)\} \prod_{t \in \mathcal{N}(s)} \underbrace{\exp\{\lambda_{ts}(x_s)\}}_{M_{ts}(x_s)}$$

$$\mu_s(x_s, x_t) \propto \exp\left\{\theta_s(x_s) + \theta_t(x_t) + \theta_{st}(x_s, x_t)\right\} \times \prod_{u \in \mathcal{N}(s) \backslash t} \exp\left\{\lambda_{us}(x_s)\right\} \prod_{v \in \mathcal{N}(t) \backslash s} \exp\left\{\lambda_{vt}(x_t)\right\}$$

Lagrangian Derivation (continued)

- Adjusting the Lagrange multipliers or messages to enforce

$$C_{ts}(x_s;\mu) := \mu_s(x_s) - \sum_{x_t} \mu_{st}(x_s, x_t) = 0$$

yields

$$M_{ts}(x_s) \quad \leftarrow \quad \sum_{x_t} \exp\left\{\theta_t(x_t) + \theta_{st}(x_s, x_t)\right\} \quad \prod_{u \in \mathcal{N}(t) \backslash s} M_{ut}(x_t)$$

- Conclusion: the message passing updates are a Lagrange method to solve the stationary condition of the variational formulation

BP on Arbitrary Graphs

- Two main difficulties of the variational formulation

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\theta^T \mu - A^*(\mu)\}$$

  - The marginal polytope $\mathcal{M}$ is hard to characterize, so let's use the tree-based outer bound

$$\mathbb{L}(G) = \left\{ \tau \geq 0 \mid \sum_{x_s} \tau_s(x_s) = 1, \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s) \right\}$$

    These locally consistent vectors τ are called pseudo-marginals.

  - Exact entropy $-A^*(\mu)$ lacks explicit form, so let's approximate it by the exact expression for trees

$$-A^*(\tau) \approx H_{\text{Bethe}}(\tau) := \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}).$$

Bethe Variational Problem (BVP)

- Combining these two ingredient leads to the Bethe variational problem (BVP):

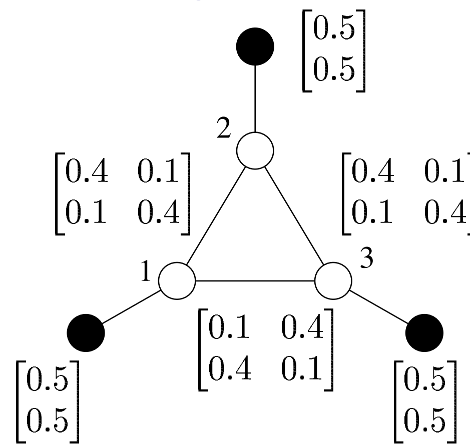
$$\max_{\tau \in \mathbb{L}(G)} \left\{ \langle \theta, \tau \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}) \right\}.$$

  - A simple structured problem (differentiable & constraint set is a simple convex polytope)

  - Loopy BP can be derived as am iterative method for solving a Lagrangian formulation of the BVP (Theorem 4.2); similar proof as for tree graphs

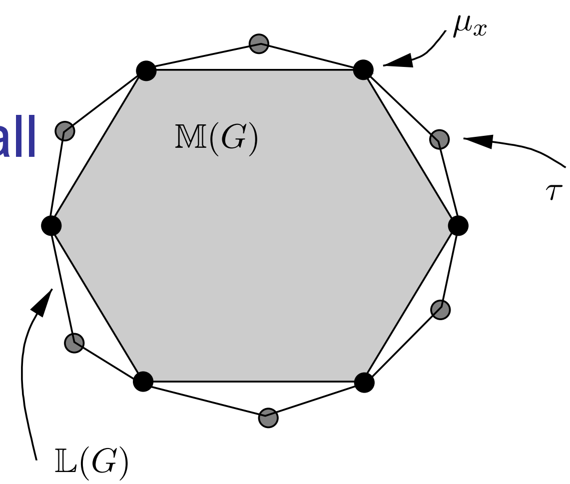
Geometry of BP

- Consider the following assignment of pseudo-marginals
  - Can easily verify $\tau \in \mathbb{L}(G)$
  - However, $\tau \notin \mathcal{M}(G)$ (need a bit more work)



- Tree-based outer bound
  - For any graph, $\mathbb{L}(G) \subseteq \mathcal{M}(G)$
  - Equality holds if and only if the graph is a tree

- Question: does solution to the BVP ever fall into the gap?
  - Yes, for any element of outer bound $\mathbb{L}(G)$, it is possible to construct a distribution with it as a BP fixed point (Wainwright et. al. 2003)

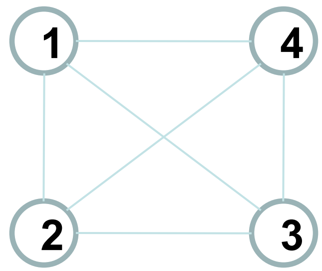


Inexactness of Bethe Entropy Approximation

- Consider a fully connected graph with

$$\mu_s(x_s) = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \quad \text{for } s = 1,2,3,4$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \quad \forall\, (s,t) \in E.$$



- It is globally valid: $\tau \in \mathcal{M}(G)$; realized by the distribution that places mass 1/2 on each of configuration (0,0,0,0) and (1,1,1,1)
- $H_{\text{Bethe}}(\mu) = 4\log 2 - 6\log 2 = -2\log 2 < 0,$
- $-A^*(\mu) = \log 2 > 0.$

Discussions

- This connection provides a principled basis for applying the sum-product algorithm for loopy graphs

- However,
  - Although there is always a fixed point of loopy BP, there is no guarantees on the convergence of the algorithm on loopy graphs
  - The Bethe variational problem is usually non-convex. Therefore, there are no guarantees on the global optimum
  - Generally, no guarantees that $A_{\mathrm{Bethe}}(\theta)$ is a lower bound of $A(\theta)$

- Nevertheless,
  - The connection and understanding suggest a number of avenues for improving upon the ordinary sum-product algorithm, via progressively better approximations to the entropy function and outer bounds on the marginal polytope (Kikuchi clustering)

Summary

- Variational methods in general turn inference into an optimization problem via exponential families and convex duality

- The exact variational principle is intractable to solve; there are two distinct components for approximations:
  - Either inner or outer bound to the marginal polytope
  - Various approximation to the entropy function

- Mean field: non-convex inner bound and exact form of entropy
- BP: polyhedral outer bound and non-convex Bethe approximation
- Kikuchi and variants: tighter polyhedral outer bounds and better entropy approximations (Yedidia et. al. 2002)

Summary

- “Off-the-Shelf” solution to inference problem?
  - Mean field: yields lower bound on the log partition function (likelihood function); widely used as an approximate E-step in EM algorithm
  - Sum-product: works well if the graph is locally tree-like and typically performs better than mean field; successfully used in error-correcting coding and low-level vision community