Lecture 5

Belief Propagation and More

Prof. Dahua Lin
dhlin@ie.cuhk.edu.hk

Tree-structured Models

- A *tree-structured* graphical model over X:

$$p(x) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \phi_{st}(x_s, x_t)$$

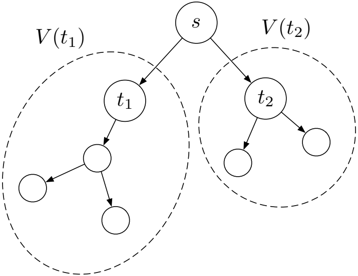
- There is a tractable algorithm, **belief propagation**, to perform exact inference on tree-structured models.

Tree-structured Models (cont'd)

- Given an *undirected tree*, one can designate any vertex r as the *root*, which results in a *directed tree*.
- Then the model can be rewritten as:

$$
\begin{aligned}
p(x) &= \frac{1}{Z}\prod_{s\in V}\psi_s(x_s)\prod_{s\in V(T)\backslash r}\phi_s(x_{\pi(s)}, x_s) \\
&= \frac{1}{Z}\psi_r(x_r)\prod_{s\in V(T)\backslash r}\psi_s(x_s)\phi_s(x_{\pi(s)}, x_s)
\end{aligned}
$$

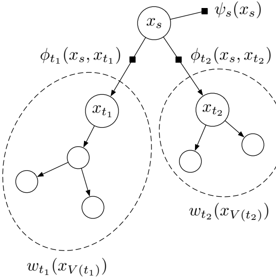
Tree Factorization



- Let $T(s)$ be the *sub-tree* with root s and $V(s)$ be all the vertices contained in $T(s)$. Then, $V(r) = V$.

- For a *non-leaf node* s, let $Ch(s) = \{t_1, \ldots, t_m\}$, then the descendants of s can be partitioned into the vertices of m sub-trees: $V(t_1), \ldots, V(t_m)$.

Factorization of Joint Distribution



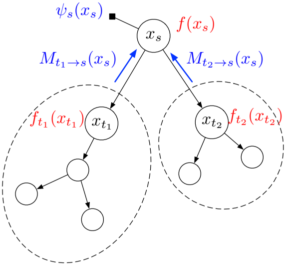
- Let $D(s) = V(s) \backslash s$. Define:

$$\omega_s(x_{V(s)}) = \psi_s(x_s) \prod_{t \in D(s)} \psi_t(x_t) \phi_t(x_{\pi(t)}, x_t)$$

- Then $p(x) \propto \omega_r(x_V)$
- For a leaf t: $\omega_t(x_t) = \psi_t(x_t)$.
- For a non-leaf s:

$$\omega_s(x_{V(s)}) = \psi(x_s) \prod_{t \in Ch(s)} \phi_t(x_s, x_t) \omega_t(x_{V(t)})$$

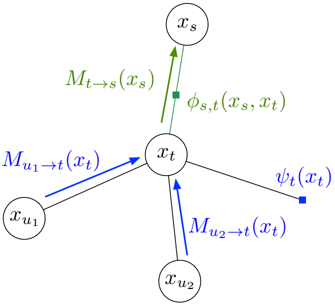
Factorization of Joint Distribution (Cont'd)



- Let $f_s(x_s) = \sum_{x_{D(s)}} \omega_s(x_s, x_{D(s)})$
- Root r: $P(X_r = x_r) \propto f_r(x_r)$
- Leaf t: $f_t(x_t) = \psi_t(x_t)$
- Non-leaf s:

$$
\begin{aligned}
& f_s(x_s) \\
&= \psi_s(x_s) \prod_{t\in Ch(s)} \sum_{x_{V(t)}} \phi_t(x_s, x_t)\omega_t(x_{V(t)}) \\
&= \psi_s(x_s) \prod_{t\in Ch(s)} \sum_{x_t} \phi_t(x_s, x_t) \sum_{x_{D(t)}} \omega_t(x_t, x_{D(t)}) \\
&= \psi_s(x_s) \prod_{t\in Ch(s)} \sum_{x_t} \phi_t(x_s, x_t) f_t(x_t)
\end{aligned}
$$

Message Form



For $t \in Ch(s)$, define:

$$M_{t\to s}(x_s) \triangleq \sum_{x_t} \phi_t(x_s, x_t) f_t(x_t)$$

$$f_s(x_s) = \psi_s(x_s) \prod_{t\in Ch(s)} M_{t\to s}(x_s)$$

Recursive definition:

$$M_{t\to s}(x_s) =$$

$$\sum_{x_t} \phi_{s,t}(x_s, x_t)\psi_t(x_t) \prod_{u\in\mathcal{N}(t)\setminus s} M_{u\to t}(x_t)$$

Belief Propagation (On Tree)

$$M_{t\to s}(x_s) \propto \sum_{x_t} \phi_{s,t}(x_s, x_t)\psi_t(x_t) \prod_{u\in\mathcal{N}(t)\backslash s} M_{u\to t}(x_t)$$

- After one inward/outward pass (for arbitrary choice of r), the marginals have:

$$\mu_s(x_s) \triangleq P(X_s = x_s) \propto \psi_s(x_s) \prod_{t\in\mathcal{N}(u)} M_{t\to s}(x_s)$$

- This is a fixed point of the message update. For a tree-structured graph, this is the *unique* fixed point.

Joint Probabilities

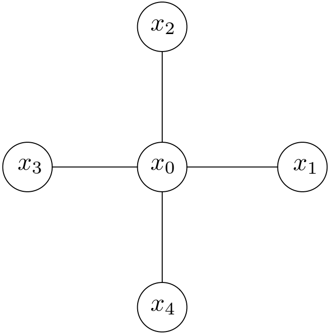
- With the messages, one can also compute the joint probabilities of two linked variables u and v.
- This can be done by merging them as a *compound variable*, which results in another tree-structured model.

$$
\begin{aligned}
\mu_{s,t}(x_s, x_t) \propto & \psi_s(x_s)\psi_t(x_t)\phi_{s,t}(x_s, x_t) \\
& \prod_{u \in \mathcal{N}(s) \setminus t} M_{u \to s}(x_s) \prod_{v \in \mathcal{N}(t) \setminus s} M_{t \to v}(x_t)
\end{aligned}
$$

Complexity Analysis

- For each edge $(s,t) \in E(T)$, there are two messages in opposite directions: $M_{s \to t}(x_t)$ and $M_{t \to s}(x_s)$, respectively of size $|\mathcal{X}_t|$ and $|\mathcal{X}_s|$.
- Total message size: $\sum_{s \in V} \deg(s) \cdot |\mathcal{X}_s|$.
- If $\mathcal{X}_s = \mathcal{X}$ for every $s \in V$, then the total size is $2(|V|-1) \cdot |\mathcal{X}|$.
- The complexity of computing $M_{t \to s}(x_s)$ is $O(m_t m_s)$ with $m_s = |\mathcal{X}_s|$.
- If $\mathcal{X}_s = \mathcal{X}$ for every $s \in V$, then the total time complexity for one-pass is $O(|V| \cdot |\mathcal{X}|^2)$.

Example (Star)



$$p(x) = \psi_0(x_0) \prod_{i=1}^{n} \phi_i(x_0, x_i)\psi_i(x_i)$$

Messages and Beliefs:

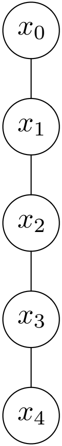
$$M_{i\to 0}(x_0) \propto \sum_{x_i} \phi_i(x_0, x_i)\psi_i(x_i)$$

$$\mu_0(x_0) \propto \psi_0(x_0) \prod_{i=1}^{n} M_{i\to 0}(x_0)$$

$$M_{0\to i}(x_i) \propto \sum_{x_0} \frac{\mu_0(x_0)\phi_i(x_0, x_i)}{M_{i\to 0}(x_0)}$$

$$\mu_i(x_i) \propto \psi_i(x_i) M_{0\to i}(x_i)$$

Example (Chain)



$$p(x) = \prod_{i=0}^{n} \psi_i(x_i) \prod_{i=1}^{n} \phi(x_{i-1}, x_i)$$

Messages and Beliefs:

$$M_{i_1 \to i_2}(x_{i_2}) \propto \sum_x M_{i_0 \to i_1}(x) \psi_{i_1}(x) \phi(x, x_{i_2})$$

$$M_{i_1 \to i_0}(x_{i_0}) \propto \sum_x M_{i_2 \to i_1}(x) \psi_{i_1}(x) \phi(x_{i_0}, x)$$

$$\mu_{i_1}(x) \propto \psi_{i_1}(x) M_{i_0 \to i_1}(x) M_{i_2 \to i_1}(x)$$

Bethe Interpretation

- There are different interpretations of *belief propagation*.
- An representative view is that BP is a fixed-point optimization procedure for the Bethe problem.
- Based on this interpretation, we can extend the analysis to non-tree models.

Marginals of Markov Networks

- Consider a Markov network over a tree $G = (V, E)$, which can generally be written as

$$p_\theta(x) = \frac{1}{Z(\theta)} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \phi_{s,t}(x_s, x_t)$$

$$= \frac{1}{Z(\theta)} \exp\left(\sum_{s \in V} \sum_{i \in \mathcal{X}_s} \theta_s^i 1_i(x_s) + \sum_{(s,t) \in E} \sum_{(x_s, x_t) \in \mathcal{X}_s \times \mathcal{X}_t} \theta_{s,t}^{i,j} 1_i(x_s) 1_j(x_t) \right)$$

- Marginals: Let $\mu_s(i) = P(X_s = i) = E_{p_\theta}[1_i(x_s)]$, and $\mu_{s,t}(i,j) = P(X_s = i, X_t = j) = E_{p_\theta}[1_i(x_s)1_j(x_t)]$

- We will discuss the properties of μ_s and $\mu_{s,t}$.

Global Consistency

- Over a graph $G = (V, E)$, a set of functions: $\{\mu_s\}_{s \in V}$ and $\{\mu_{s,t}\}_{(s,t) \in E}$ are called *globally consistent* if there exist θ such that

$$\mu_s(i) = P(X_s = i) = E_{p_\theta}[1_i(x_s)]$$
$$\mu_{s,t}(i,j) = P(X_s = i, X_t = j) = E_{p_\theta}[1_i(x_s)1_j(x_t)]$$

- We use $\mathbb{M}(G)$ to denote all *globally consistent* function sets as defined above. Such functions constitute the *mean parameters* of p_θ.

Mean Parameters of Trees

Tree-structured Markov models can be parameterized in terms of mean parameters. Arbitrarily choose any $r \in V$ as the root:

$$
\begin{aligned}
p(x_v) &= p_r(x_r) \prod_{s \in V \setminus r} p_{s|\pi(s)}(x_s | x_{\pi(s)}) \\
&= \mu_r(x_r) \prod_{s \in V \setminus r} \frac{\mu_{\pi(s),s}(x_{\pi(s)}, x_s)}{\mu_{\pi(s)}(x_{\pi(s)})} \\
&= \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{s,t}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}
\end{aligned}
$$

Local Consistency

- Globally consistent functions satisfy:

$$\sum_{x_s \in \mathcal{X}_s} \mu_s(x_s) = 1, \quad \forall s \in V$$

$$\sum_{x_s \in \mathcal{X}_s} \mu_{s,t}(x_s, x_t) = \mu_t(x_t), \quad \forall (s,t) \in E,\ x_t \in \mathcal{X}_t$$

$$\sum_{x_t \in \mathcal{X}_t} \mu_{s,t}(x_s, x_t) = \mu_s(x_s), \quad \forall (s,t) \in E,\ x_s \in \mathcal{X}_s$$

- Functions over $G = (V, E)$ which satisfy the above equalities are called *locally consistent*. We use $\mathbb{L}(G)$ to denote the collection of all such function sets.

Global and Local Consistencies

- $\mathbb{M}(G) \subset \mathbb{L}(G)$ holds for any graph G.
- If G is a tree, $\mathbb{M}(G) = \mathbb{L}(G)$.
- One can construct a valid model given $\mu \in \mathbb{L}(G)$ through mean parameterization.

Entropy of Tree Models

Consider a tree-structured Markov network with mean parameters μ, we have

$$\begin{aligned}
H(\mu) &= -A^*(\mu) = E_\mu[-\log p_\mu(X)] \\
&= \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t)} I_{s,t}(\mu_{s,t}) \\
H_s(\mu_s) &= -\sum_{x_s \in \mathcal{X}_s} \mu_s(x_s) \log \mu_s(x_s) \\
I_{s,t}(\mu_{s,t}) &= \sum_{(x_s, x_t) \in \mathcal{X}_s \times \mathcal{X}_t} \mu_{s,t}(x_s, x_t) \log \frac{\mu_{s,t}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}
\end{aligned}$$

Bethe Approximation

- For *loopy graphs*, *i.e.* graphs with cycles, computing $A^*(\mu)$ is generally intractable.

- *Bethe approximation* is to use *Bethe entropy* of *pseudo-marginals*, *i.e.* functions that are locally consistent *w.r.t.* G to approximate the *true entropy*:

$$H_{Be}(\tau) = \sum_{s \in V} H_v(\tau_s) - \sum_{(s,t) \in E} I_{s,t}(\tau_{s,t})$$

where $\tau \in \mathbb{L}(G)$.

Bethe Variational Problem

- Recall: mean parameter can be computed as

$$\mu = \underset{\mu \in \mathbb{M}(G)}{\operatorname{argmax}} \; \theta^T \mu - A^*(\mu)$$

- With *Bethe approximation*:

$$\tau = \underset{\tau \in \mathbb{L}(G)}{\operatorname{argmax}} \; \theta^T \tau + H_{Be}(\tau)$$

- This is called the *Bethe variational problem (BVP)*. The solutions are *pseudo-marginals*.

Bethe Variational Problem (cont'd)

- The *Bethe approximation* is exact when G is a tree.
- It relaxes the solution domain from $\mathbb{M}(G)$ to a convex outer bound $\mathbb{L}(G)$.
- Generally, this is not necessarily a *convex optimization problem* when G is loopy.
- (Loopy) belief propagation is a fixed-point process to find the solution *(Homework Exercise)*.

Discussions

- For tree-structured graph:
  - the Bethe variational problem has a unique solution (τ^*, λ^*), where τ^* corresponds to the single and pairwise marginals.
  - For tree-structured graph, the sum-product belief propagation converges to a unique fixed point, which is equal to this solution.
- For loopy graphs:
  - There is no guarantee that the BP update would converge.
  - The convergence depends on both the topological structure of the graph and the factor values.

Bethe Approximation of $A(\theta)$

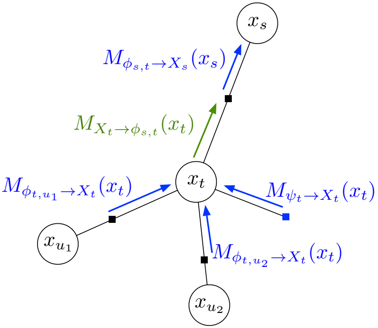
- Define $A_{Be}(\theta)$ as:

$$A_{Be}(\theta) = \sup_{\tau} \left\{ \theta^T \tau + H_{Be}(\tau) \right\}$$

- For a tree-structured model, $A_{Be}(\theta) = A(\theta)$, because H_{Be} is exact.
- In general, $A_{Be}(\theta)$ is an approximation of $A(\theta)$, and there's no guarantee that it is an upper bound or lower bound in general.

BP on Factor Graphs

$$M_{t\to s}(x_s) \propto \sum_{x_t} \phi_{s,t}(x_s, x_t)\psi_t(x_t) \prod_{u\in\mathcal{N}(t)\setminus s} M_{u\to t}(x_t)$$



This message can be decomposed into a series of messages between *variables* and *factors*:

$$M_{\phi_{t,u}\to X_t}(x_t) := M_{u\to t}(x_t)$$

$$M_{\psi_t\to X_t}(x_t) \propto \psi_t(x_t)$$

$$M_{X_t\to\phi_{s,t}}(x_t) \propto M_{\psi_t\to t}(x_t) \prod_{u\in\mathcal{N}(t)\setminus s} M_{\phi_{u,t}\to t}(x_t)$$

$$M_{\phi_{s,t}\to X_s}(x_s) \propto \sum_{x_t} \phi_{s,t}(x_s, x_t) M_{X_t\to\phi_{s,t}}(x_t) = M_{t\to s}(x_s)$$

BP on Factor Graphs (Cont'd)

Belief propagation on factor graphs can be expressed:

- Variable $\to$ factor messages:

$$M_{v\to\phi}(x_v) \propto \prod_{f\in\mathcal{F}(v)\setminus\phi} M_{f\to v}(x_v)$$

- Factor $\to$ variable messages:

$$M_{\phi\to v}(x_v) \propto \sum_{x'_{C(\phi)}:x'_v=x_v} \phi(x'_C) \prod_{u\in C(\phi)\setminus v} M_{u\to\phi}(x'_u)$$

Beliefs on Factor Graphs

- Singleton beliefs:

$$\mu_v(x_v) \propto \prod_{f \in \mathcal{F}(v)} M_{f \to v}(x_v)$$

- Clique beliefs: Let $C := C(\phi)$,

$$\mu_C(x_C) \propto \phi(x_C) \prod_{v \in C} \prod_{f \in \mathcal{F}(v) \setminus \phi} M_{f \to v}(x_v)$$

Discussions

- For a unary factor $\psi_v(x_v)$, one have to only compute messages from the factor to the associated variable as $M_{\psi \to v}(x_v) \propto \psi_v(x_v)$. The message $M_{v \to \psi}$ is never needed.
- The belief propagation over a factor graph is a fixed point algorithm for a generalized Bethe variational problem defined thereon (*Yedidia et al, 2004*)

Tree-Reweighted Message Passing

- A message passing procedure in a way similar to BP. A variant of the algorithm *guarantees* convergence.
- It is based on a *variational approximation* that provides an *upper bound* of $A(\theta)$.
- Work on $\mathbb{L}(G)$ instead of $\mathbb{M}(G)$, like BP.
- Very effective in practice.

Setup

Consider a *Markov network* over $G = (V, E)$ as:

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

- Let $\mu_s^{(i)} = E_p[\delta_i(x_s)]$ and $\mu_{st}^{(i,j)} = E_p[\delta_i(x_s)\delta_j(x_t)]$.
- $\mathbb{M}(G)$: *(globally) realizable marginals* and $\mathbb{L}(G)$: *locally consistent pseudo-marginals*.

Motivating Idea

- Computing the log-partition $A(\theta)$ is intractable in general.
- Computing $A(\theta)$ for trees is tractable.
- **Idea:** approximate $A(\theta)$ of a loopy graphical model with a *convex combination* of *tree-based* log-partitions.

Parameters on a Spanning Tree

Let $T = (V, E(T))$ be a *spanning tree* of $G = (V, E)$:

- Define $\mathcal{I}(T)$ to be the set of *the indices of the parameters* on the tree T:

$$\mathcal{I}(T) := \{s \mid s \in V\} \cup \{st \mid (s,t) \in E(T)\}.$$

- Define $\mathcal{E}(T)$ to be the set of *parameters* whose coefficients are non-zeros only on the tree T:

$$\mathcal{E}(T) := \{\theta \mid \theta_\alpha = 0 \ \forall \alpha \in \mathcal{I}(G) \backslash \mathcal{I}(T)\}.$$

- When $\theta \in \mathcal{E}(T)$, computing $A(\theta)$ is tractable.

Distribution over Spanning Trees

- Let $\mathfrak{T}$ be the set of all *spanning trees*.
- Let ρ be a *distribution* over $\mathfrak{T}$, s.t. $\rho(T) \geq 0$ for every T and $\sum_{T \in \mathfrak{T}} \rho(T) = 1$.
- Given $T \sim \rho$, $\rho_e \triangleq \Pr(e \in T)$ for each $e \in E$ is called the *edge appearance probability* of e.
- $\boldsymbol{\rho} : e \mapsto \rho_e$ is a vector of $|E|$-dimension.

Convex Combination of Trees

- Let $\boldsymbol{\theta} := (\theta(T))_{T \in \mathfrak{T}}$ be a collection of parameters, each associated with a spanning tree T.
- Let $\mathcal{E} \triangleq \{\boldsymbol{\theta} \mid \theta(T) \in \mathcal{E}(T) \ \forall T \in \mathcal{E}(T)\}$.
- With $\boldsymbol{\theta} \in \mathcal{E}$ and ρ, we can form a *convex combination* of exponential family parameters:

$$E_\rho[\theta(T)] = \sum_T \rho(T)\theta(T)$$

.

Convex Combination of Trees (cont'd)

- Given a target parameter $\bar{\theta}$ and a distribution ρ over $\mathfrak{T}$, we define:

$$\mathcal{Q}_\rho(\bar{\theta}) = \{\boldsymbol{\theta} \in \mathcal{E} \mid E_\rho[\theta(T)] = \bar{\theta}\}$$

- Any member $\boldsymbol{\theta} \in \mathcal{A}_\rho(\bar{\theta})$ is called a *ρ-reparameterization* of $p_{\bar{\theta}}$.
- $\mathcal{Q}_\rho(\bar{\theta})$ is never empty as long as $\rho \succ 0$. Why?

Convex Upper Bounds

- Let $F(\theta)$ be a *convex function*. Given $\bar{\theta}$ and ρ, for any $\boldsymbol{\theta} \in \mathcal{Q}_\rho(\bar{\theta})$, we have

$$F(\bar{\theta}) \leq \underbrace{\sum_T \rho(T) F(\theta(T))}_{\text{convex upper bound}}$$

- $\sum_T \rho(T) A(\theta(T))$ is an upper bound of $A(\bar{\theta})$.

Optimal Upper Bound

To find the optimal (*i.e.* smallest) upper bound, we formulate the following problem. Given a fixed ρ,

$$\underset{\boldsymbol{\theta}\in\mathcal{E}}{\text{minimize}} \sum_{T\in\mathfrak{T}} \rho(T)A(\theta(T))$$

$$\text{s.t.} \sum_{T\in\mathfrak{T}} \rho(T)\theta(T) = \bar{\theta}$$

This is a convex problem, but $\mathfrak{T}$ is excessively large.

Tree-consistent Pseudo-marginals

To construct a *dual problem*, we introduce *dual variables* μ:

- μ_s for the constraint $E_\rho[\theta_s(T)] = \bar{\theta}_s$
- μ_{st} for the constraint $E_\rho[\theta_{st}(T)] = \bar{\theta}_{st}$.
- When optimality is attained, we have $\mu \in \mathbb{L}(G)$.
  We are going to show this.

Tree-consistent Pseudo-marginals (Cont'd)

We form the *Lagrangian* L:

$$\begin{aligned} L(\boldsymbol{\theta}, \mu) &= E_\rho[A(\theta(T))] + \langle \mu, \bar{\theta} - E_\rho[\theta(T)] \rangle \\ &= \mu^T \bar{\theta} + E_\rho[A(\theta(T)) - \mu^T \theta(T)] \end{aligned}$$

Then $\nabla_{\theta_\alpha(T)} L = 0 \implies$

$$E_{\hat{\theta}(T)}[\phi_\alpha] = \hat{\mu}_\alpha, \ \forall T \in \mathfrak{T}, \alpha \in \mathcal{I}(\alpha)$$

For every $T \in \mathfrak{T}$:

- $E_{\hat{\theta}(T)}[\delta_i(X_s)] = \hat{\mu}_s^{(i)}$
- $E_{\hat{\theta}(T)}[\delta_i(X_s)\delta_j(X_t)] = \hat{\mu}_{st}^{(i,j)}, \ \forall (s,t) \in E(T)$

Dual Problem

- By the duality of exponential family:

$$A^*(\Pi_T(\hat{\mu})) = \langle \hat{\theta}(T), \hat{\mu} \rangle - A(\hat{\theta}(T))$$

- Substituting this into the Lagrangian yields:

$$L(\hat{\boldsymbol{\theta}}, \hat{\mu}) = \bar{\theta}^T \hat{\mu} - E_\rho[A^*(\Pi_T(\hat{\mu}))]$$

- The dual problem is to maximize $\hat{\theta}^T \mu - E_\rho[A^*(\Pi_T(\mu))]$, s.t. $\mu \in \mathbb{L}(G)$.

Approximating $E_\rho[A^*(\Pi_T(\hat{\mu}))]$

For a tree-based model:

$$A^*(\Pi_T(\mu)) = -\sum_{s\in V} H_s(\mu_s) + \sum_{(s,t)\in E(T)} I_{st}(\mu_{st})$$

The expectation is then:

$$\begin{aligned} E_\rho[A^*(\Pi_T(\mu))] &= \sum_T \rho(T)\left[-\sum_{s\in V} H_s(\mu_s) + \sum_{(s,t)\in E(T)} I_{st}(\mu_{st})\right] \\ &= -\sum_{s\in V} H_s(\mu_s) + \sum_{(s,t)\in E} \rho_{st} I_{st}(\mu_{st}) \\ &=: -H_{Trw}(\mu; \boldsymbol{\rho}_e) \end{aligned}$$

Dual Problem

We finally obtain a dual problem:

$$\underset{\mu \in \mathbb{L}(G)}{\text{maximize}} \;\; \bar{\theta}^T \mu + H_{Trw}(\mu; \boldsymbol{\rho}_e)$$

$$\text{s.t. } \mu \in \mathbb{L}(G).$$

- This is also a *convex optimization* problem, and it is *tractable*.
- Solving this problem is actually to perform *approximate inference* of the marginals, called *tree-reweighted inference*.

A Closer Look

- Recall: Bethe variational problem:

$$\underset{\mu \in \mathbb{L}(G)}{\text{maximize}} \;\; \theta^T \mu + H_{Be}(\mu)$$

- Tree-reweighted variational problem:

$$\underset{\mu \in \mathbb{L}(G)}{\text{maximize}} \;\; \theta^T \mu + H_{Trw}(\mu)$$

- Key difference: H_{Be} vs. H_{Trw}.

H_{Be} vs. H_{Trw}

Compare:

$$H_{Be}(\mu) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st})$$

$$H_{Trw}(\mu; \boldsymbol{\rho}_e) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\mu_{st})$$

When $G = (V, E)$ is a tree, they reduce to the same problem, *i.e.* exact inference on the tree.

Tree-reweighted Message Passing

The fixed point update based on H_{Trw}:

$$M_{t,s}(x_s) \propto \sum_{x_t} \exp\left(\rho_{st}^{-1}\theta_{st}(x_s, x_t) + \theta_t(x_t)\right) \cdot \frac{\prod_{u \in \mathcal{N}(t) \setminus s} [M_{u,t}(x_t)]^{\rho_{ut}}}{[M_{s,t}(x_t)]^{(1-\rho_{ts})}}$$

- When $\rho_{st} = 1$ for every $(s, t) \in E$, this reduces to the sum-product belief propagation.
- When G is a tree, this performs exact inference.
- No guarantee of convergence in general.

Summary

- Variable elimination
- Belief propagation
- Bethe approximation
- Tree-reweighted message passing