EE512A – Advanced Inference in Graphical Models

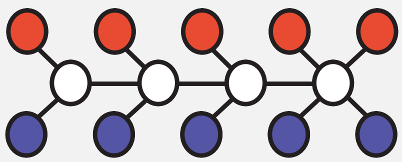
— Fall Quarter, Lecture 14 —

http://j.ee.washington.edu/~bilmes/classes/ee512a_fall_2014/

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
http://melodi.ee.washington.edu/~bilmes

Nov 17th, 2014



Announcements

- Wainwright and Jordan *Graphical Models, Exponential Families, and Variational Inference* http://www.nowpublishers.com/product.aspx?product=MAL&doi=2200000001
- Read chapters 1,2, and 3 in this book. Start reading chapter 4.
- Assignment due Wednesday (Nov 12th) night, 11:45pm. Non-binding final project proposals (one page max).

Class Road Map - EE512a

- L1 (9/29): Introduction, Families, Semantics
- L2 (10/1): MRFs, elimination, Inference on Trees
- L3 (10/6): Tree inference, message passing, more general queries, non-tree)
- L4 (10/8): Non-trees, perfect elimination, triangulated graphs
- L5 (10/13): triangulated graphs, k-trees, the triangulation process/heuristics
- L6 (10/15): multiple queries, decomposable models, junction trees
- L7 (10/20): junction trees, begin intersection graphs
- L8 (10/22): intersection graphs, inference on junction trees
- L9 (10/27): inference on junction trees, semirings,
- L10 (11/3): conditioning, hardness, LBP
- L11 (11/5): LBP, exponential models,
- L12 (11/10): exponential models, mean params and polytopes,
- L13 (11/12): polytopes, tree outer bound, Bethe entropy approx.
- L14 (11/17):
- L15 (11/19):
- L16 (11/24):
- L17 (11/26):
- L18 (12/1):
- L19 (12/3):
- Final Presentations: (12/10):

Finals Week: Dec 8th-12th, 2014.

Conjugate Duality

- Consider maximum likelihood problem for exp. family
$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{14.3}$$
- Compare this to convex conjugate dual (also sometimes Fenchel-Legendre dual or transform) of $A(\theta)$ is defined as:
$$A^*(\mu) \triangleq \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \tag{14.4}$$
- So dual is optimal value of the ML problem, when $\mu \in \mathcal{M}$, and we saw the relationship between ML and negative entropy before.
- Key: when $\mu \in \mathcal{M}$, dual is negative entropy of exponential model $p_{\theta(\mu)}$ where $\theta(\mu)$ is the unique set of canonical parameters satisfying this *matching condition*
$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu \tag{14.5}$$
- When $\mu \notin \mathcal{M}$, then $A^*(\mu) = +\infty$, optimization with dual need consider points only in $\mathcal{M}$.

Conjugate Duality, Maximum Likelihood, Negative Entropy

Theorem 14.2.3 (Relationship between A and A^*)

(a) *For any $\mu \in \mathcal{M}^\circ$, $\theta(\mu)$ unique canonical parameter sat. matching condition, then conj. dual takes form:*

$$A^*(\mu) = \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \in \overline{\mathcal{M}} \end{cases} \tag{14.3}$$

(b) *Partition function has variational representation (dual of dual)*

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\} \tag{14.4}$$

(c) *For $\theta \in \Omega$, sup occurs at $\mu \in \mathcal{M}^\circ$ of moment matching conditions*

$$\mu = \int_{\mathsf{D}_X} \phi(x) p_\theta(x) \nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta) \tag{14.5}$$

Conjugate Duality, Good and Bad News

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\} \tag{14.4}$$

- Computing $A(\theta)$ in this way corresponds to the inference problem (finding mean parameters, or node and edge marginals).
- Key: **we compute the log partition function simultaneously with solving inference, given the dual.**
- Good news: problem is concave objective over a convex set. Should be easy. In simple examples, indeed, it is easy. ☺
- Bad news: $\mathcal{M}$ is quite complicated to characterize, depends on the complexity of the graphical model. ☹
- More bad news: A^* not given explicitly in general and hard to compute. ☹

Conjugate Duality, Avenues to Approximation

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{14.4}$$

- Some good news: The above form gives us new avenues to do approximation. ☺
- For example, we might either relax $\mathcal{M}$ (making it less complex), relax $A^*(\mu)$ (making it easier to compute over), or both. ☺
- $A^*(\mu)$'s relationship to entropy gives avenues for relaxation.
- Surprisingly, this is strongly related to belief propagation (i.e., the sum-product commutative semiring). ☺☺
- Much of the rest of the class will be above approaches to the above — giving not only to junction tree algorithm (that we've seen) but also to well-known approximation methods (LBP, mean-field, Bethe, expectation-propagation (EP), Kikuchi methods, linear programming relaxations, and semidefnite relaxations, some of which we will cover).

Local consistency (tree outer bound) polytope

- An "outer bound" of $\mathbb{M}$ consists of a set that contains $\mathbb{M}$. If formed from a **subset** of the linear inequalities (subset of the rows of matrix module (A, b)), then it is a polyhedral outer bound.
- A simple way to form outer bound: require only local consistency, i.e., consider set $\{\tau_v, v \in V(G)\} \cup \{\tau_{s,t}, (s,t) \in E(G)\}$ that is, always non-negative , and that satisfies normalization

$$\sum_{x_v} \tau_v(x_v) = 1 \tag{14.8}$$

and pair-node marginal consistency constraints

$$\sum_{x_t'} \tau_{s,t}(x_s, x_t') = \tau_s(x_s) \tag{14.9a}$$

$$\sum_{x_s'} \tau_{s,t}(x_s', x_t) = \tau_t(x_t) \tag{14.9b}$$

Local consistency (tree outer bound) polytope: properties

- Define $\mathbb{L}(G)$ to be the (locally consistent) polytope that obeys the constraints in Equations 14.8 and 14.9.
- Recall: local consistency was the necessary conditions for potentials being marginals that, it turned out, for junction tree that also guaranteed global consistency.
- Clearly $\mathbb{M} \subseteq \mathbb{L}(G)$ since any member of $\mathbb{M}$ (true marginals) will be locally consistent.
- When G is a tree, we say that local consistency implies global consistency, so for any tree T, we have $\mathbb{M}(T) = \mathbb{L}(T)$
- When G has cycles, however, $\mathbb{M}(G) \subset \mathbb{L}(G)$ strictly. We refer to members of $\mathbb{L}(G)$ as **pseudo-marginals**
- Key problem is that members of $\mathbb{L}$ might not be true possible marginals for any distribution.

Bethe Entropy Approximation

- In terms of current notation, we can let $\mu \in \mathbb{L}(T)$, the pseudo marginals associated with T. Since local consistency requires global consistency, for a tree, any $\mu \in \mathbb{L}(T)$ is such that $\mu \in \mathbb{M}(T)$, thus

$$p_\mu(x) = \prod_{s \in V(T)} \mu_s(x_s) \prod_{(s,t) \in E(T)} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{14.10}$$

- When $G = T$ is a tree, and $\mu \in \mathbb{L}(T) = \mathbb{M}(T)$ we have

$$-A^*(\mu) = H(p_\mu) = \sum_{v \in V(T)} H(X_v) - \sum_{(s,t) \in E(T)} I(X_s; X_t) \tag{14.11}$$

$$= \sum_{v \in V(T)} H_v(\mu_v) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \tag{14.12}$$

- That is, for $G = T$, $-A^*(\mu)$ is very easy to compute (only need to compute entropy and mutual information over at most pairs).

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{14.14}$$

Approximate variational representation of log partition function

$$A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{L}} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\} \tag{14.15}$$

$$= \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \tag{14.16}$$

- Exact when $G = T$ but we do this for any G, still commutable
- we get an approximate log partition function, and approximate (pseudo) marginals (in $\mathbb{L}$), but this is perhaps much easier to compute.
- We can optimize this directly using a Lagrangian formulation.

Fixed points: Variational Problem and LBP

Theorem 14.2.3

LBP updates are Lagrangian method for attempting to solve Bethe variational problem:
(a) *For any G, any LBP fixed point specifies a pair (τ^*, λ^*) s.t.*

$$\nabla_\tau \mathcal{L}(\tau^*, \lambda^*; \theta) = 0 \text{ and } \nabla_\lambda \mathcal{L}(\tau^*, \lambda^*; \theta) = 0 \tag{14.18}$$

(b) *For tree MRFs, Lagrangian equations have unique solution (τ^*, λ^*) where τ^* are exact node and edge marginals for the tree and the optimal value obtained is the true log partition function.*

- Not guaranteed convex optimization, but is if graph is tree.
- Remarkably, this means if we run loopy belief propagation, and we reach a point where we have converged, then we will have achieved a fixed-point of the above Lagrangian, and thus a (perhaps reasonable) local optimum of the underlying variational problem.

What about $\mathbb{L} \setminus \mathbb{M}$?

- Do solutions to Bethe variational problem (equivalently fixed points of LBP) ever fall into $\mathbb{L}(G) \setminus \mathbb{M}(G)$ (which we know to be non-empty for non-tree graphs)?
- Unfortunately, for **all** $\tau \in \mathbb{L}(G)$, then it can be a fixed point for LBP for some p_θ. true for Lagrangian optimization as well. ☹
- Recall notion of reparameterization: for tree graph, such that we can reparameterize so that the edges and nodes are true marginals. e.g., $\phi_i(x_i) = \sum_{x_{V \setminus \{i\}}} p(x)$. A goal of inference is to change factors to become true marginals, can't be done for graphs with cycles in general.
- Fixed points of LBP do not get marginal reparameterization but it does get something that does still preserve the original when global renormalized.
- That is, we have

Reparameterization Properties of Bethe Approximation

Proposition 14.3.1

Let $\tau^* = (\tau^*_s, s \in V; \tau^*_{st}, (s,t) \in E(G))$ *denote any optimum of the Bethe variational principle defined by the distribution* p_θ. *Then the distribution defined by the fixed point as*

$$p_{\tau^*}(x) \triangleq \frac{1}{Z(\tau^*)} \prod_{s \in V} \tau^*_s(x_s) \prod_{(s,t) \in E(G)} \frac{\tau^*_{st}(x_s, x_t)}{\tau^*_s(x_s)\tau^*_t(x_t)} \tag{14.1}$$

is a reparameterization of the original. That is, we have $p_\theta(x) = p_{\tau^*}(x)$ *for all* x.

- For trees, we have $Z(\tau^*) = 1$.
- Form gives strategies for seeing how bad we are doing for any given instance (by, say, comparing marginals) - approximation error (possibly a bound)

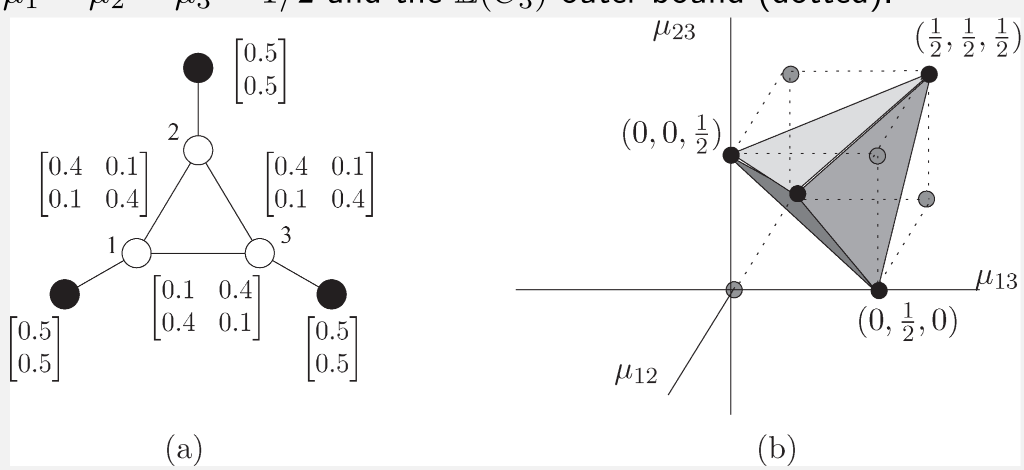
Review

the next slide is a repeat from lecture 13.

Pseudo-marginals

$$\tau_v(x_v) = [0.5, 0.5], \text{ and } \tau_{s,t}(x_s, x_t) = \begin{bmatrix} \beta_{st} & .5 - \beta_{st} \\ .5 - \beta_{st} & \beta_{st} \end{bmatrix} \quad (14.8)$$

- Consider on 3-cycle C_3, satisfies local consistency.
- But for this won't give us a marginal. Below shows $\mathbb{M}(C_3)$ for $\mu_1 = \mu_2 = \mu_3 = 1/2$ and the $\mathbb{L}(C_3)$ outer bound (dotted).



A fixed point in $\mathbb{L} \setminus \mathbb{M}$ is possible.

- Consider

$$\theta_s(x_s) = \log \tau_s(x_s) = \log[0.5 \;\; 0.5] \qquad \text{for } s = 1,2,3,4 \tag{14.2a}$$

$$\begin{aligned}\theta_{st}(x_s, x_t) &= \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} \\ &= \log 4 \begin{bmatrix} \beta_{st} & 0.5 - \beta_{st} \\ 0.5 - \beta_{st} & \beta_{st} \end{bmatrix} \; \forall (s,t) \in E(G)\end{aligned} \tag{14.2b}$$

- We saw in the pseudo marginals slide that, for a 3-cycle, a choice of parameters that gave us $\tau \in \mathbb{L} \setminus \mathbb{M}$. Is this achievable as fixed point of LBP?
- For this choice of parameters, if we start sending messages, starting from the uniform messages, then this will be a fixed point. ☹

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{14.14}$$

Approximate variational representation of log partition function

$$\begin{aligned} A_{\text{Bethe}}(\theta) &= \sup_{\tau \in \mathbb{L}} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\} &(14.15)\\ &= \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} &(14.16)\end{aligned}$$

- Exact when $G = T$ but we do this for any G, still commutable
- we get an approximate log partition function, and approximate (pseudo) marginals (in $\mathbb{L}$), but this is perhaps much easier to compute.
- We can optimize this directly using a Lagrangian formulation.

If not Bounds, then Better Approximation?

- We might want bounds between $A(\theta)$ and $A_{\text{Bethe}}(\theta)$ in the ideal case.
- Perhaps we can come up with an expression for $A(\theta) - A_{\text{Bethe}}(\theta)$
- We don't expect the expression to be easy to compute. Why?
- Expression, however, could help make the difference smaller by approximating the difference in a computationally practical way. I.e.,

$$A(\theta) = A_{\text{Bethe}}(\theta) + \underbrace{A(\theta) - A_{\text{Bethe}}(\theta)}_{\text{expression to approximate}} \tag{14.3}$$

- This is the idea behind Loop Series Expansions

Vertex and Edge Induced Subgraphs, Degree, and Generalized Loops

- Given a graph $G = (V, E)$, it is possible to construct either a vertex- or an edge-induced subgraph.
- Given subset $S \subseteq V$, then $G' = (S, E(S))$ is a vertex induced subgraph, $E(S) = E \cap (S \times S)$.
- Given subset $\tilde{E} \subseteq E$, then $G(\tilde{E}) = (V(\tilde{E}), \tilde{E})$ is edge-induced subgraph, $V(\tilde{E}) = V \cap \left\{ u \in V : \exists v \text{ s.t. } (u, v) \in \tilde{E} \right\}$.
- Usually, "induced subgraph" means "vertex induced subgraph" when it is not specified.
- Define the degree of s in the subgraph as $d_s(\tilde{E}) = |\delta_s(\tilde{E})|$ where $\delta_s(\tilde{E}) = \left\{ t \in V | (s, t) \in \tilde{E} \right\}$ is the set of neighbors of s in $G(\tilde{E})$.
- Definition: a generalized loop is a subgraph $G(\tilde{E})$ where no node has degree 1 (i.e., $d_s(\tilde{E}) \neq 1$ for all $s \in V(G(\tilde{E}))$.

Generalized Loops

- Definition: a generalized loop is a subgraph $G(\tilde{E})$ where no node has degree 1 (i.e., $d_s(\tilde{E}) \neq 1$ for all $s \in V(G(\tilde{E}))$.
- Example:

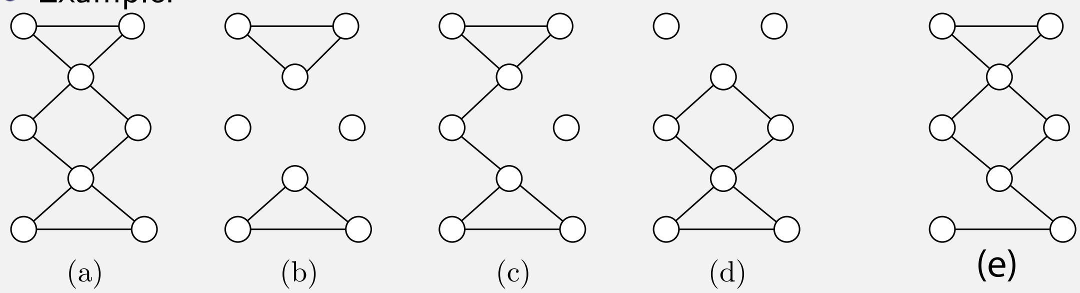


Illustration of generalized loops. (a) An original graph. (b)-(d) Various generalized loops associated with the graph in (a). In this particular case, the original graph is a generalized loop for itself. (e) is not a generalized loop as it has a leaf node.

Edge weights Generalized Loops

- Consider LBP fixed point for binary pairwise MRF (Ising model).
- Unary and pairwise pseudomarginals can be parameterized using $\{\tau_s\}_{s \in V}$ and $\{\tau_{st}\}_{(s,t) \in E}$, where

$$\tau_s(x_s) = \begin{bmatrix} 1 - \tau_s \\ \tau_s \end{bmatrix}, \text{ and } \tau_{st}(x_s, x_t) = \begin{bmatrix} 1 - \tau_s - \tau_t + \tau_{st} & \tau_t - \tau_{st} \\ \tau_s - \tau_{st} & \tau_{st} \end{bmatrix}$$

- Being in local consistency (tree outer bound) polytope $\mathbb{L}(T)$ is the same as: $\tau_s \geq 0$, $\tau_{st} \geq 0$, $1 - \tau_s - \tau_t + \tau_{st} \geq 0$, and $\tau_s - \tau_{st} \geq 0$.
- Define an edge weight β_{st} as follows:

$$\beta_{st} \triangleq \frac{\tau_{st} - \tau_s \tau_t}{\tau_s(1 - \tau_s)\tau_t(1 - \tau_t)} \tag{14.4}$$

- and the weight can be extended to an edge-induced subgraph via $\tilde{E}$

$$\beta_{\tilde{E}} \triangleq \prod_{(s,t) \in \tilde{E}} \beta_{st} \tag{14.5}$$

Comparison of A and A_{Bethe}: loop series expansion

Proposition 14.4.1

Consider a pairwise MRF with binary variables, with $A_{Bethe}(\theta)$ being the optimized free energy evaluated at a LBP fixed point $\tau = (\tau_s, s \in V; \tau_{st}, (s,t) \in E(G))$. Then we have the following relationship with the cumulant function $A(\theta)$.

$$A(\theta) = A_{Bethe}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \quad (14.6)$$

- For any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, inner term is zero and vanishes. why? Since $E_{\tau_s}\left[(X_s - \tau_s)^d\right]$ is the d^{th} central moment. Thus, terms in the sum only exists for generalized loops.
- The generalized loops give the correction!
- For trees, there are no generalized loops, and so if G is a tree then we have an equality between $A(\theta)$ and $A_{\text{Bethe}}(\theta)$ (recall both defs here).

Proof of Proposition 14.4.1

proof sketch.

- Overcomplete, $\exists$ parameters $\hat{\theta}$ s.t. $\left\langle \hat{\theta}, \phi(x) \right\rangle = c$ for all x.
- Thus, we can show this for just one set of parameters θ since $A(\theta)$ and $A_{\text{Bethe}}(\theta)$ both shift by same amount.
- We choose parameterization at a LBP fixed point, so

$$\tilde{\theta}_s(x_s) = \log \tau_s(x_s), \text{ and } \tilde{\theta}_{st}(x_s, x_t) = \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} \quad (14.7)$$

- With this paramterization, $A_{\text{Bethe}}(\tilde{\theta}) = 0$ (since the optimization attempts to maximize a set of negative KL-divergence terms).
- Thus, we need only show

$$A(\tilde{\theta}) = \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \quad (14.8)$$

Proof of Proposition 14.4.1 cont.

proof sketch.

- By checking for each value of $(x_s, x_t) \in \{0,1\}^2$, we have

$$\frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} = 1 + \beta_{st}(x_s - \tau_s)(x_t - \tau_t) \tag{14.9}$$

- Moreover, at current parameterization $\tilde{\theta}$, we have

$$\exp(A(\tilde{\theta})) = \sum_{x \in \{0,1\}^m} \prod_{s \in V} \tau_s(x_s) \prod_{(s,t) \in E} \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} \tag{14.10}$$

- Let $\tau_{\text{fact}} = \prod_s \tau_s(x_s)$ and let $\mathbb{E}$ be w.r.t. τ_{fact}, then

$$\exp(A(\tilde{\theta})) = \mathbb{E}\left[\prod_{(s,t) \in E} (1 + \beta_{st}(X_s - \tau_s)(X_t - \tau_t))\right] \tag{14.11}$$

Bethe Entropy Approx | Bethe & Loop Series | Kikuchi and Hypertree-based Methods | Refs

Proof of Proposition 14.4.1 cont.

proof sketch.

- By polynomial expansion, linearity of expectation, we get

$$\exp(A(\tilde{\theta})) = 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \mathbb{E}\left[\prod_{(s,t) \in \tilde{E}} (\beta_{st}(X_s - \tau_s)(X_t - \tau_t))\right] \tag{14.12}$$

- And by independence of τ_{frac}, we get

$$\exp(A(\tilde{\theta})) = 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s}\left[(X_s - \tau_s)^{d_s(\tilde{E})}\right] \tag{14.13}$$

□

Comparison of A and A_{Bethe}: loop series expansion

Proposition 14.4.1

Consider a pairwise MRF with binary variables, with $A_{Bethe}(\theta)$ being the optimized free energy evaluated at a LBP fixed point $\tau = (\tau_s, s \in V; \tau_{st}, (s,t) \in E(G))$. Then we have the following relationship with the cumulant function $A(\theta)$.

$$A(\theta) = A_{Bethe}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \quad (14.6)$$

- For any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, inner term is zero and vanishes. why? Since $E_{\tau_s}\left[(X_s - \tau_s)^d\right]$ is the d^{th} central moment. Thus, terms in the sum only exists for generalized loops.
- The generalized loops give the correction!
- For trees, there are no generalized loops, and so if G is a tree then we have an equality between $A(\theta)$ and $A_{\text{Bethe}}(\theta)$ (recall both defs here).

Loop Series Approximations

- So, various forms of approximation can be made by taking, rather than a sum over all $2^E \setminus \{\emptyset\}$, some small set of subsets of $\emptyset \notin \mathcal{E} \subseteq 2^E$ for which the summation is tractable.
- For attractive potentials (which we'll define later in the class, and which are related to submodular potentials, and which essentially always encourage neighbors to be the same), it is the case that we have:

$$A_{\text{Bethe}}(\theta) \leq A(\theta) \quad (14.14)$$

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{14.15}$$

- So far, we have used a replacement for $-A^*(\mu)$ inspired by trees.
- But we know a tree is really a 1-tree. Why not k-tree?
- Why not some other junction tree?
- Junction trees are really hypertrees (special case of hypergraphs).
- So can we come up with: 1) replacement for $-A^*(\mu)$ associated with a hypertree/junction tree; 2) a generalization for this replacement for any hypergraph; and 3) a corresponding generalized polytope associated with the hypergraph?
- This is the Kikuchi variational approach (or "clustered variational approximation").

Hypergraphs

- A graph $G = (V, E)$ is a set of nodes V and edges E where every $(s, t) = e \in E$ is only two nodes.
- A hypergraph is a system (V, E) where every $e \in E$ can consist of any number of nodes. I.e., we might have $\{v_1, v_2, \ldots, v_k\} = e \in E(G)$ for a hypergraph.
- A hypertree is a hypergraph that can be reduced to a tree in a particular way, we've already seen them in the forms of junction trees.
- A junction tree (which, recall, satisfies r.i.p.) is a hypertree where the cliques (which are clusters of graph nodes) in the junction tree are the edges of the hypertree.

Hypergraphs

Definition 14.5.1 (hypergraph)

A *hypergraph* $H = (V, E)$ is a set of vertices V and a collection of hyperedges E, where each element $e \in E$ is a subset of V, so $\forall e \in E, e \subseteq V$. In a graph, $|e| = 2$. In a hypergraph, it can be larger.

Definition 14.5.2 (leaf)

A vertex $v \in V(H)$ of H is called a *leaf* if it appears only in a single maximal hyper-edge $h \in H$.

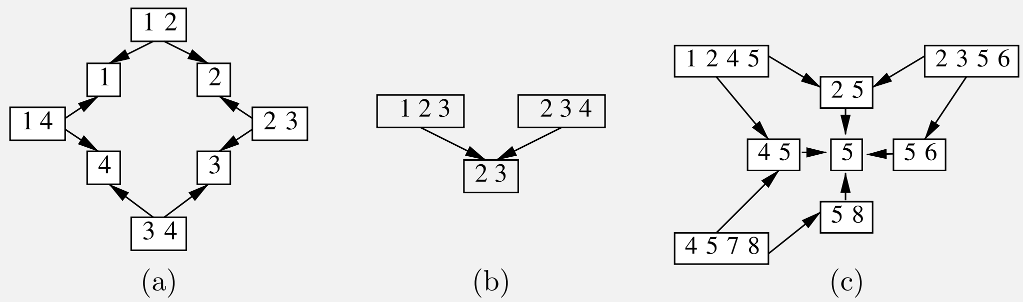
Definition 14.5.3 (acyclic)

A hypergraph H is called *acyclic* if it is empty, or if it contains a leaf v such that induced hypergraph $H(V - \{v\})$ is acyclic (note, generalization of perfect elimination order in a triangulated graph, junction tree).

Hypergraphs and bipartite graphs

Hypergraphs can be represented by a bipartite $G = (V, F, E)$ graphs where V is a set of left-nodes, F is a set of right nodes, and E is a set of size-two edges. Right nodes are hyperedges in the hypergraphs.
Some hand-drawn examples:

Hypergraphs and posets



Graphical representations of hypergraphs. Subsets of nodes corresponding to hyperedges are shown in rectangles, whereas the arrows represent inclusion relations among hyperedges. (a) An ordinary single 4-cycle graph represented as a hypergraph. (b) A simple hypertree of "width" two. (c) A more complex hypertree of "width" three.
Here, $a \to b$ if it is the case that $b \preceq a$ and there does not exist a c such that $b \preceq c \preceq a$, similar to a Hasse lattice diagram. As bipartite graphs:

Partially ordered set

- A partially ordered set (poset) is a set $\mathcal{P}$ of objects with an order.
- Set of objects $\mathcal{P}$ and a binary relation $\preceq$ which can be read as "is contained in" or "is part of" or "is less than or equal to".
- For any $x, y \in \mathcal{P}$, we may ask is $x \preceq y$ which is either true or false.
- In a poset, for any $x, y, z \in \mathcal{P}$ the following conditions hold (by definition):

For all $x, x \preceq x$.	(Reflexive)	(P1.)
If $x \preceq y$ and $y \preceq x$, then $x = y$	(Antisymmetriy)	(P2.)
If $x \preceq y$ and $y \preceq z$, then $x \preceq z$.	(Transitivity)	(P3.)

- We can use the above to get other operators as well such as "less than" via $x \preceq y$ and $x \neq y$ implies $x \prec y$. Also, we get $x \succ y$ if not $x \preceq y$. And $x \succeq y$ is read "x contains y". And so on.

Möbius Inversion Lemma

- A zeta function of a poset is a mapping $\zeta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by

$$\zeta(g,h) = \begin{cases} 1 & \text{if } g \preceq h, \\ 0 & \text{otherwise.} \end{cases} \tag{14.16}$$

- The Möbius function $\omega : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ is the multiplicative inverse of this function. It is defined recursively:
- $\omega(g,g) = 1$ for all $g \in \mathcal{P}$
- $\omega(g,h) = 0$ for all $h : h \not\preceq g$.
- Given $\omega(g,f)$ defined for f such that $g \preceq f \preceq h$, we define

$$\omega(g,h) = - \sum_{\{f | g \preceq f \prec h\}} \omega(g,f) \tag{14.17}$$

- Then, ω and ζ are multiplicative inverses, in that

$$\sum_{f \in \mathcal{P}} \omega(g,f)\zeta(f,h) = \sum_{\{f | g \subseteq f \subseteq h\}} \omega(g,f) = \delta(g,h) \tag{14.18}$$

General Möbius Inversion Lemma

Lemma 14.5.4 (General Möbius Inversion Lemma)

Given real valued functions Υ and Ω defined on poset $\mathcal{P}$, then $\Omega(h)$ may be expressed via $\Upsilon(\cdot)$ via

$$\Omega(h) = \sum_{g \preceq h} \Upsilon(g) \quad \textit{for all } h \in \mathcal{P} \tag{14.19}$$

iff $\Upsilon(h)$ may be expressed via $\Omega(\cdot)$ via

$$\Upsilon(h) = \sum_{g \preceq h} \Omega(g)\omega(g,h) \quad \textit{for all } h \in \mathcal{P} \tag{14.20}$$

When $\mathcal{P} = 2^V$ for some set V (so this means that the poset consists of sets and all subsets of an underlying set V) this can be simplified, where $\preceq$ becomes $\subseteq$; and $\succeq$ becomes $\supseteq$.

Möbius Inversion Lemma

Lemma 14.5.5 (Möbius Inversion Lemma (for sets))

Let Υ and Ω be functions defined on the set of all subsets of a finite set V, taking values in an Abelian group (i.e., a set and an operator having properties closure, associativity, identity, and inverse, and for which the elements also commute, the real numbers being just one example). The following two equations imply each other.

$$\forall A \subseteq V : \Upsilon(A) = \sum_{B : B \subseteq A} \Omega(B) \tag{14.21}$$

$$\forall A \subseteq V : \Omega(A) = \sum_{B : B \subseteq A} (-1)^{|A \setminus B|} \Upsilon(B) \tag{14.22}$$

Proof of Möbius Inversion Lemma

Proof.

$$\sum_{B : B \subseteq A} \Omega(B) = \sum_{B : B \subseteq A} \sum_{C : C \subseteq B} (-1)^{|B \setminus C|} \Upsilon(C) \tag{14.23}$$

$$= \sum_{C : C \subseteq A} \sum_{B : C \subseteq B \& B \subseteq A} \Upsilon(C)(-1)^{|B \setminus C|} \tag{14.24}$$

$$= \sum_{C : C \subseteq A} \Upsilon(C) \sum_{B : C \subseteq B \& B \subseteq A} (-1)^{|B \setminus C|} \tag{14.25}$$

$$= \sum_{C : C \subseteq A} \Upsilon(C) \sum_{H : H \subseteq A \setminus C} (-1)^{|H|} \tag{14.26}$$

Proof of Möbius Inversion Lemma

Proof Cont.

Also, note that for some set D,

$$\sum_{H:H\subseteq D}(-1)^{|H|} = \sum_{i=0}^{|D|}\binom{|D|}{i}(-1)^i = \sum_{i=0}^{|D|}\binom{|D|}{i}(-1)^i(1)^{|D|-i} \tag{14.27}$$

$$= (1-1)^{|D|} = \begin{cases} 1 & \text{if } |D| = 0 \\ 0 & \text{otherwise} \end{cases} \tag{14.28}$$

which means

$$\sum_{H:H\subseteq A\setminus C}(-1)^{|H|} = \begin{cases} 1 & \text{if } A = C \\ 0 & \text{otherwise} \end{cases} \tag{14.29}$$

Proof of Möbius Inversion Lemma

Proof Cont.

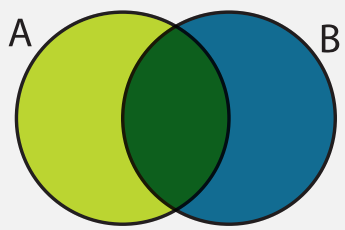
This gives

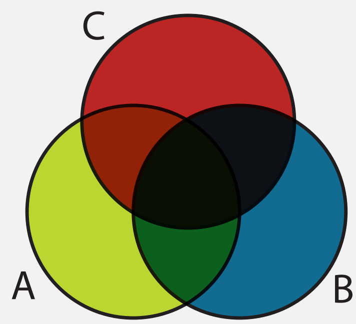
$$\sum_{B:B\subseteq A}\Omega(B) = \sum_{C:C\subseteq A}\Upsilon(C)\mathbf{1}\{A=C\} = \Upsilon(A) \tag{14.30}$$

thus proving one direction. The other direction is very similar. □

Möbius Inversion Lemma and Inclusion-Exclusion

- This is a general case of inclusion-exclusion.
- Given ground set V and $A, B \subseteq V$, to compute the size $|A \cup B| = |A| + |B| - |A \cap B|$.
- $A, B, C \subseteq V$, then $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$. Start by including, then excluding, and then including again.





- Also consider entropy: $H(X, Y) = H(X) + H(Y) - I(X; Y)$. $H(X, Y, Z) = H(X) + H(Y) + H(Z) - I(X; Y) - I(X; Z) - I(Y; Z) + I(X; Y; Z)$.

Möbius Inversion Lemma and Inclusion-Exclusion

- Ex: Set cardinality inclusion-exclusion: Given $A_1, A_2, \ldots, A_n \subseteq V$,

$$|\cup_{i=1}^n A_n| = \sum_{j=1}^{n} (-1)^{j-1} \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} |A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_j}| \quad (14.31)$$

- This is a special case of Möbius Inversion Lemma:

$$\forall A \subseteq V : \Upsilon(A) = \sum_{B : B \subseteq A} \Omega(B) \quad (14.32)$$

$$\forall A \subseteq V : \Omega(A) = \sum_{B : B \subseteq A} (-1)^{|A \setminus B|} \Upsilon(B) \quad (14.33)$$

- Möbius Inversion lemma is also used to prove the Hammersley-Clifford theorem (that factorization and Markov property definitions of families are identical for positive distributions).
- We use it here to come up with alternative expressions for the entropy and for the marginal polytope.

Back to Kikuchi: Möbius and expressions of factorization

- Suppose we are given marginals that factor w.r.t. a hypergraph $G = (V, E)$, so we have $\mu = (\mu_h, h \in E)$, then we can define new functions $\varphi = (\varphi_h, h \in E)$ via Möbius inversion lemma as follows

$$\log \varphi_h(x_h) \triangleq \sum_{g \preceq h} \omega(g, h) \log \mu_g(x_g) \tag{14.34}$$

  (see Stanley, "Enumerative Combinatorics" for more info.)
- From Möbius inversion lemma, this then gives us a new way to write the log marginals, i.e., as

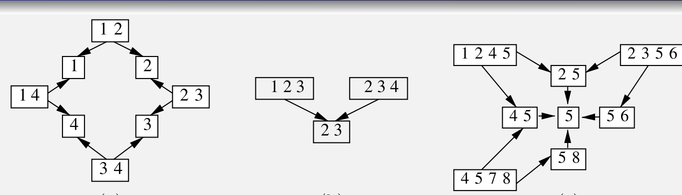
$$\log \mu_h(x_h) = \sum_{g \preceq h} \log \varphi_g(x_g) \tag{14.35}$$

- Key, when φ_h is defined as above, and G is a hypertree we have

$$p_\mu(x) = \prod_{h \in E} \varphi_h(x_h) \tag{14.36}$$

  $\Rightarrow$ general way to factorize a distribution that factors w.r.t. a hypergraph. When a 1-tree, we recover factorization we already know.

expressions of factorization and Möbius



(a) (b) (c)

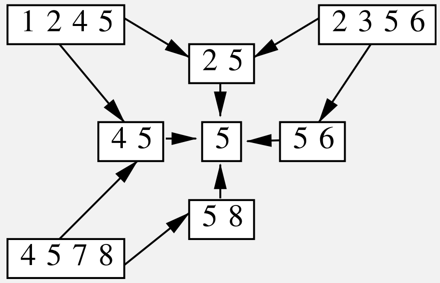
- When the graph is a tree (a 1-tree), we have $\varphi_s(x_s) = \mu_s(x_s)$ and

$$\varphi_{st}(x_s, x_t) = \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{14.37}$$

  giving us the tree factorization we saw early in this course.
- For the more general hypertree, consider edge set $E = \{(12345), (2356), (4578), (25), (45), (56), (58), (5)\}$. Check: is this a junction tree of cliques?
- Then, from Eqn. (14.35), we get unaries $\varphi_s(x_s) = \mu_s(x_s)$ and pairwise (e.g., $\varphi_{25} = \mu_{25}/\mu_5$, etc.) and

$$\varphi_{1245} = \frac{\mu_{1245}}{\varphi_{25}\varphi_{45}\varphi_5} = \frac{\mu_{1245}}{\frac{\mu_{25}}{\mu_5}\frac{\mu_{45}}{\mu_5}\mu_5} = \frac{\mu_{1245}\mu_5}{\mu_{25}\mu_{45}} \tag{14.38}$$

expressions of factorization and Möbius



(c)

- Doing this for all maxcliques of the figure, we get a factorization of the form:

$$p_\mu = \frac{\mu_{1245}\mu_5}{\mu_{25}\mu_{45}} \frac{\mu_{2356}\mu_5}{\mu_{25}\mu_{56}} \frac{\mu_{4578}\mu_5}{\mu_{45}\mu_{58}} \frac{\mu_{25}}{\mu_5} \frac{\mu_{45}}{\mu_5} \frac{\mu_{56}}{\mu_5} \frac{\mu_{58}}{\mu_5} \mu_5 \tag{14.39}$$

$$= \frac{\mu_{1245}\mu_{2356}\mu_{4578}}{\mu_{25}\mu_{45}} \tag{14.40}$$

New expressions of entropy

- We can express entropic quantities as well, such as the hyperedge entropy

$$H_h(\mu_h) = -\sum_{x_h} \mu_h(x_h) \log \mu_h(x_h) \tag{14.41}$$

and the multi-information function

$$I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) \tag{14.42}$$

- In the case of a single tree edge $h = (s,t)$, then $I_h(\mu_h) = I(X_s; X_t)$ the standard mutual information.
- Then the overall entropy of any hypertree distribution becomes

$$H_{\text{hyper}}(\mu) = -\sum_{h \in E} I_h(\mu_h) \tag{14.43}$$

multi-information decomposition

- Using Möbius, we can write

$$I_h(\mu_h) = \sum_{g \preceq h} \omega(g,h) \left\{ \sum_{x_h} \mu_h(x_h) \log \mu_g(x_g) \right\} \tag{14.44}$$

$$= \sum_{f \preceq h} \sum_{e \succeq f} \omega(e,f) \left\{ \sum_{x_f} \mu_f(x_f) \log \mu_f(x_f) \right\} \tag{14.45}$$

$$= -\sum_{f \preceq h} c(f) H_f(\mu_f) \tag{14.46}$$

where

$$c(f) \triangleq \sum_{e \succeq f} \omega(f,e) \tag{14.47}$$

- This gives us a new expression for the hypertree entropy

$$H_{\text{hyper}}(\mu) = \sum_{h \in E} c(h) H_h(\mu_h) \tag{14.48}$$

Usable to get Kikuchi variational approximation

- Given arbitrary hypergraph now, we can generalize the hypertree-specific expressions above to this arbitrary hypergraph. This will give us a variational expression that approximates cumulant.
- Given hypergraph $G = (V, E)$, we have

$$p_\theta(x) \propto \exp \left\{ \sum_{h \in E} \sigma_h(x_h) \right\} \tag{14.49}$$

using same form of parameterization.
- Hypergraph will give us local marginal constraints on hypergraph pseudo marginals, i.e., for each $h \in E$, we form marginal $\tau_h(x_h)$ and define constraints, non-negative, and

$$\sum_{x_h} \tau_h(x_h) = 1 \tag{14.50}$$

Usable to get Kikuchi variational approximation

- Sum to one constraint:

$$\sum_{x_h} \tau_h(x_h) = 1 \tag{14.51}$$

- Local agreement via the hypergraph constraint. For any $g \preceq h$ must have marginalization condition

$$\sum_{x_{h \setminus g}} \tau_h(x_h) = \tau_g(x_g) \tag{14.52}$$

- Define new polyhedral constraint set $\mathbb{L}_t(G)$

$$\mathbb{L}_t(G) = \{\tau \geq 0 | \text{ Equations (14.51) } \forall h, \text{ and (14.52) } \forall g \preceq h \text{ hold}\} \tag{14.53}$$

Kikuchi variational approximation

- Generalized entropy for the hypergraph:

$$H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \tag{14.54}$$

where H_g is hyperedge entropy and overcounting number defined by:

$$c(g) = \sum_{f \succeq g} \omega(g, f) \tag{14.55}$$

Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

$$A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{\langle \theta, \tau \rangle + H_{\text{app}}(\tau)\} \tag{14.56}$$

- For a graph, this is exactly $A_{\text{Bethe}}(\theta)$!
- Also, if hypergraph is a junction tree (r.i.p. holds, and tree-local consistency implies global consistency), then this is also exact (although it might be expensive, exponential in the tree-width to compute H_{app}).
- We can define message passing algorithms on the hypertree, and show that if it converges, it is a fixed point of the associated Lagrangian.

Sources for Today's Lecture

- Wainwright and Jordan *Graphical Models, Exponential Families, and Variational Inference* `http://www.nowpublishers.com/product.aspx?product=MAL&doi=2200000001`