EE512A – Advanced Inference in Graphical Models

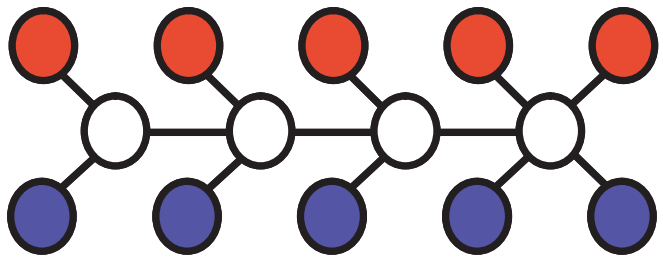
— Fall Quarter, Lecture 13 —

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 12th, 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, tree outer bound
- 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.

Mean Parameters, Convex Cores

- Consider quantities μ_α associated with statistic ϕ_α defined as:

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \nu(dx) \tag{13.10}$$

- this defines a vector of "mean parameters" $(\mu_1, \mu_2, \ldots, \mu_d)$ with $d = |\mathcal{I}|$.
- Define all possible such vectors, with $d = |\mathcal{I}|$,

$$\mathcal{M}(\phi) = \mathcal{M} \stackrel{\Delta}{=} \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \forall \alpha \in \mathcal{I}, \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] \right\} \tag{13.11}$$

- We don't say p was necessarily exponential family
- $\mathcal{M}$ is convex since expected value is a linear operator. So convex combinations of p and p' will lead to convex combinations of μ and μ'
- $\mathcal{M}$ is like a "convex core" of all distributions expressed via ϕ.

Mean Parameters and Marginal Polytopes

- Mean parameters are now true (fully specified) marginals, i.e., $\mu_v(j) = p(x_v = j)$ and $\mu_{st}(j,k) = p(x_s = j, x_t = k)$ since

$$\mu_{v,j} = \mathbb{E}_p[\mathbf{1}(x_v = j)] = p(x_v = j) \quad (13.20)$$

$$\mu_{st,jk} = \mathbb{E}_p[\mathbf{1}(x_s = j, x_t = k)] = p(x_s = j, x_t = k) \quad (13.21)$$

- Such an $\mathcal{M}$ is called the *marginal polytope* for discrete graphical models. Any μ must live in the polytope that corresponds to node and edge true marginals.
- We can also associate such a polytope with a graph G, where we take only $(s,t) \in E(G)$. Denote this as $\mathbb{M}(G)$.
- This polytope can help us to characterize when BP converges (there might be an outer bound of this polytope), or it might characterize the result of a mean-field approximation (an inner bound of this polytope) as we'll see.

Learning is the dual of Inference

- We can view the inference problem as moving from the canonical parameters θ to the point in the marginal polytope, called forward mapping, moving from $\theta \in \Omega$ to $\mu \in \mathcal{M}$.
- We can view the (maximum likelihood) learning problem as moving from a point in the polytope (given by the empirical distribution) to the canonical parameters. Called backwards mapping
- graph structure (e.g., tree-width) makes this easy or hard, and also characterizes the polytope (how complex it is in terms of number of faces).

Maximum entropy estimation

- Goal ("estimation", or "machine learning") is to find

$$p^* \in \operatorname*{argmax}_{p \in \mathcal{U}} H(p) \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \hat{\mu}_\alpha \ \forall \alpha \in \mathcal{I} \tag{13.14}$$

where $H(p) = -\int p(x) \log p(x) \nu(dx)$, and $\forall \alpha \in \mathcal{I}$

$$\mathbb{E}_p[\phi_\alpha(X)] = \int_{\mathsf{D}_X} \phi_\alpha(x) p(x) \nu(dx). \tag{13.15}$$

- $\mathbb{E}_p[\phi_\alpha(X)]$ is mean value as measured by potential function, so above is a form of moment matching.
- Maximum entropy (MaxEnt) distribution is solved by taking distribution in form of $p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$ and then by finding canonical parameters θ that solves

$$E_{p_\theta}[\phi_\alpha(X)] = \hat{\mu}_\alpha \text{ for all } \alpha \in \mathcal{I}. \tag{13.16}$$

Learning is the dual of Inference

- Ex: Estimate θ with $\hat{\theta}$ based on data $\mathbf{D} = \{\bar{x}^{(i)}\}_{i=1}^{M}$ of size M, likelihood function

$$\ell(\theta, \mathbf{D}) = \frac{1}{M} \sum_{i=1}^{M} \log p_\theta(\bar{x}^{(i)}) = \frac{1}{M} \sum_{i=1}^{M} \left(\left\langle \theta, \phi(\bar{x}^{(i)}) \right\rangle - A(\theta) \right) \quad (13.20)$$

$$= \langle \theta, \hat{\mu} \rangle - A(\theta) \quad (13.21)$$

where empirical means are given by:

$$\hat{\mu} = \hat{\mathbb{E}}[\phi(X)] = \frac{1}{M} \sum_{i=1}^{M} \phi(\bar{x}^{(i)}) \quad (13.22)$$

- By taking derivatives of the above, it is easy to see that solution is the point $\hat{\theta} = \theta(\hat{\mu})$ such that empirical matches expected means, or what are called the moment matching conditions:

$$\mathbb{E}_{\hat{\theta}}[\phi(X)] = \hat{\mu} \quad (13.23)$$

this is the the backward mapping problem, going from μ to θ.

- Here, maximum likelihood is identical to maximum entropy problem.

Likelihood and negative entropy

- Entropy definition again: $H(p) = -\int p(x) \log p(x) \nu(dx)$
- Given data, $\mathbf{D} = \{\bar{x}^{(i)}\}_{i=1}^M$, defines an empirical distribution

$$\hat{p}(x) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}(x = \bar{x}^{(i)}) \tag{13.20}$$

so that $\mathbb{E}_{\hat{p}}[\phi(X)] = \int \hat{p}(x)\phi(x)\nu(dx) = \frac{1}{M}\sum_{i=1}^M \phi(\bar{x}^{(i)}) = \hat{\mu}$

- Starting from maximum likelihood solution $\theta(\hat{u})$, meaning we are at moment matching conditions $\mathbb{E}_{p_{\theta(\hat{u})}}[\phi(X)] = \hat{\mu} = \mathbb{E}_{\hat{p}}[\phi(X)]$, we have

$$\ell(\theta(\hat{u}), \mathbf{D}) = \langle \theta(\hat{u}), \hat{\mu} \rangle - A(\theta(\hat{u})) = \frac{1}{M} \sum_{i=1}^{M} \log p_{\theta(\hat{u})}(\bar{x}^{(i)}) \tag{13.21}$$

$$= \int \hat{p}(x) \log p_{\theta(\hat{\mu})}(x) \nu(dx) = \mathbb{E}_{\hat{p}}[\log p_{\theta(\hat{\mu})}(x)] \tag{13.22}$$

$$= \mathbb{E}_{p_{\theta(\hat{\mu})}}[\log p_{\theta(\hat{\mu})}(x)] = -H_{p_{\theta(\hat{\mu})}}[p_{\theta(\hat{\mu})}(x)] \tag{13.23}$$

- Thus, maximum likelihood value and negative entropy are identical, at least for empirical $\hat{\mu}$ (which is $\in \mathcal{M}$).

Dual Mappings: Summary

Summarizing these relationships

- Forward mapping: moving from $\theta \in \Omega$ to $\mu \in \mathcal{M}$, this is the inference problem, getting the marginals.
- Backwards mapping: moving from $\mu \in \mathcal{M}$ to $\theta \in \Omega$, this is the learning problem, getting the parameters for a given set of empirical facts (means).
- In exponential family case, this is maximum entropy and is equivalent to maximum likelihood learning on an exponential family model.
- Turns out log partition function A, and its dual A^* can give us these mappings, and the mappings have interesting forms . . .

Log partition (or cumulant) function: derivative offerings

$$A(\theta) = \log \int_{\mathsf{D}_X} \exp \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{13.20}$$

- If we know the log partition function, we know a lot for an exponential family model. In particular, we know
- $A(\theta)$ is convex in θ (strictly so if minimal representation).
- It yields cumulants of the random vector $\phi(X)$

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] = \int \phi_\alpha(X) p_\theta(x) \nu(dx) = \mu_\alpha \tag{13.21}$$

  in general, derivative of log part. function is expected value of feature
- Also, we get

$$\frac{\partial^2 A}{\partial \theta_{\alpha_1} \partial \theta_{\alpha_2}}(\theta) = \mathbb{E}_\theta[\phi_{\alpha_1}(X)\phi_{\alpha_2}(X)] - \mathbb{E}_\theta[\phi_{\alpha_1}(X)]\mathbb{E}_\theta[\phi_{\alpha_2}(X)] \tag{13.22}$$

- Proof given in book (Proposition 3.1, page 62).

Log partition function: properties

- So derivative of log partition function w.r.t. θ is equal to our mean parameter μ in the discrete case.
- Given $A(\theta)$, we can recover the marginals for each potential function $\phi_\alpha, \alpha \in \mathcal{I}$ (when mean parameters lie in the marginal polytope).
- If we can approximate $A(\theta)$ with $\tilde{A}(\theta)$ then we can get approximate marginals. Perhaps we can bound it without inordinate compute resources. Why do we want bounds? We shall see in future lectures.
- The Bethe approximation (as we'll also see) is such an approximation and corresponds to fixed points of loopy belief propagation.
- In some rarer cases, we can bound the approximation (current research trend).

Exponential Family: Recap

- Exponential Family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{13.1}$$

with

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{13.2}$$

Exponential Family: Recap

- Exponential Family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{13.1}$$

with

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{13.2}$$

- $A(\theta)$ is key.

Exponential Family: Recap

- Exponential Family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{13.1}$$

with

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{13.2}$$

- $A(\theta)$ is key.
- Forward mapping, inference: from $\theta \in \Omega$ to $\mu \in \mathcal{M}$, get marginals.

Exponential Family: Recap

- Exponential Family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{13.1}$$

with

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{13.2}$$

- $A(\theta)$ is key.
- Forward mapping, inference: from $\theta \in \Omega$ to $\mu \in \mathcal{M}$, get marginals.
- Backwards mapping, learning: from $\mu \in \mathcal{M}$ to $\theta \in \Omega$, getting best parameters associated with empirical facts (means).

Exponential Family: Recap

- Exponential Family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{13.1}$$

with

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{13.2}$$

- $A(\theta)$ is key.
- Forward mapping, inference: from $\theta \in \Omega$ to $\mu \in \mathcal{M}$, get marginals.
- Backwards mapping, learning: from $\mu \in \mathcal{M}$ to $\theta \in \Omega$, getting best parameters associated with empirical facts (means).
- So learning is dual of inference.

Log partition function: Properties

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \left\{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\right\}$.

Log partition function: Properties

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \left\{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\right\}$.
- Proofs of the below are in our text:

Log partition function: Properties

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \left\{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\right\}$.
- Proofs of the below are in our text:
- For minimal exponential family models, this mapping is one-to-one, that is there is a unique pairing between μ and θ.

Log partition function: Properties

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \left\{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\right\}$.
- Proofs of the below are in our text:
- For minimal exponential family models, this mapping is one-to-one, that is there is a unique pairing between μ and θ.
- For non-minimal exponential families, more than one θ for a given μ (not surprising since multiple θ's can yield the same distribution).

Log partition function: Properties

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \left\{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\right\}$.
- Proofs of the below are in our text:
- For minimal exponential family models, this mapping is one-to-one, that is there is a unique pairing between μ and θ.
- For non-minimal exponential families, more than one θ for a given μ (not surprising since multiple θ's can yield the same distribution).
- For non-exponential families, other distributions can yield μ, but the exponential family one is the one that has maximum entropy.

Log partition function: Properties

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \left\{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\right\}$.
- Proofs of the below are in our text:
- For minimal exponential family models, this mapping is one-to-one, that is there is a unique pairing between μ and θ.
- For non-minimal exponential families, more than one θ for a given μ (not surprising since multiple θ's can yield the same distribution).
- For non-exponential families, other distributions can yield μ, but the exponential family one is the one that has maximum entropy. ex1: Gaussian, a distribution with maximum entropy amongst all other distributions with same mean and covariance.

Log partition function: Properties

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \left\{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\right\}$.
- Proofs of the below are in our text:
- For minimal exponential family models, this mapping is one-to-one, that is there is a unique pairing between μ and θ.
- For non-minimal exponential families, more than one θ for a given μ (not surprising since multiple θ's can yield the same distribution).
- For non-exponential families, other distributions can yield μ, but the exponential family one is the one that has maximum entropy. ex1: Gaussian, a distribution with maximum entropy amongst all other distributions with same mean and covariance. ex2: Consider the maximum entropy optimization problem, yields a distribution with exactly this property.

Log partition function: Properties

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \left\{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\right\}$.
- Proofs of the below are in our text:
- For minimal exponential family models, this mapping is one-to-one, that is there is a unique pairing between μ and θ.
- For non-minimal exponential families, more than one θ for a given μ (not surprising since multiple θ's can yield the same distribution).
- For non-exponential families, other distributions can yield μ, but the exponential family one is the one that has maximum entropy. ex1: Gaussian, a distribution with maximum entropy amongst all other distributions with same mean and covariance. ex2: Consider the maximum entropy optimization problem, yields a distribution with exactly this property.
- Key point: all mean parameters that are realizable by some dist. are also realizable by member of exp. family.

Mappings - one-to-one

Expanding on one of the previous properties, . . .

Theorem 13.3.1

The gradient map ∇A is one-to-one iff the exponential representation is minimal.

Mappings - one-to-one

Expanding on one of the previous properties, . . .

Theorem 13.3.1

The gradient map ∇A is one-to-one iff the exponential representation is minimal.

- Proof basically uses property that if representation is non-minimal, and $\langle a, \phi(x) \rangle = c$ for all x, then we can form an affine set of equivalent parameters $\theta + \gamma a$.

Mappings - one-to-one

Expanding on one of the previous properties, . . .

Theorem 13.3.1

The gradient map ∇A is one-to-one iff the exponential representation is minimal.

- Proof basically uses property that if representation is non-minimal, and $\langle a, \phi(x) \rangle = c$ for all x, then we can form an affine set of equivalent parameters $\theta + \gamma a$.
- Other direction, uses strict convexity of $A(\theta)$

Mappings - onto

Theorem 13.3.2

In a minimal exponential family, the gradient map ∇A is onto the interior of $\mathcal{M}$ (denoted $\mathcal{M}^\circ$). Consequently, for each $\mu \in \mathcal{M}^\circ$, there exists some $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.

Mappings - onto

Theorem 13.3.2

In a minimal exponential family, the gradient map ∇A is onto the interior of $\mathcal{M}$ (denoted $\mathcal{M}^\circ$). Consequently, for each $\mu \in \mathcal{M}^\circ$, there exists some $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.

- Ex: Gaussian. Any mean parameter (set of means $\mathbb{E}[X]$ and correlations $\mathbb{E}[XX^T]$) can be realized by a Gaussian having those same mean parameters (moments).

Mappings - onto

Theorem 13.3.2

In a minimal exponential family, the gradient map ∇A is onto the interior of $\mathcal{M}$ (denoted $\mathcal{M}^\circ$). Consequently, for each $\mu \in \mathcal{M}^\circ$, there exists some $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.

- Ex: Gaussian. Any mean parameter (set of means $\mathbb{E}[X]$ and correlations $\mathbb{E}[XX^T]$) can be realized by a Gaussian having those same mean parameters (moments).
- The Gaussian won't nec. be the "true" distribtuion (in such case, the "true" distribution would not be a Gaussian, and might be an exponential family distribution with additional moments (e.g., 1D Gaussians have zero skew and kurtosis) or might not be exponential family at all).

Mappings - onto

Theorem 13.3.2

In a minimal exponential family, the gradient map ∇A is onto the interior of $\mathcal{M}$ (denoted $\mathcal{M}^\circ$). Consequently, for each $\mu \in \mathcal{M}^\circ$, there exists some $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.

- Ex: Gaussian. Any mean parameter (set of means $\mathbb{E}[X]$ and correlations $\mathbb{E}[XX^T]$) can be realized by a Gaussian having those same mean parameters (moments).
- The Gaussian won't nec. be the "true" distribtuion (in such case, the "true" distribution would not be a Gaussian, and might be an exponential family distribution with additional moments (e.g., 1D Gaussians have zero skew and kurtosis) or might not be exponential family at all).
- The theorem here is more general and applies for any set of sufficient statistics.

Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{13.3}$$

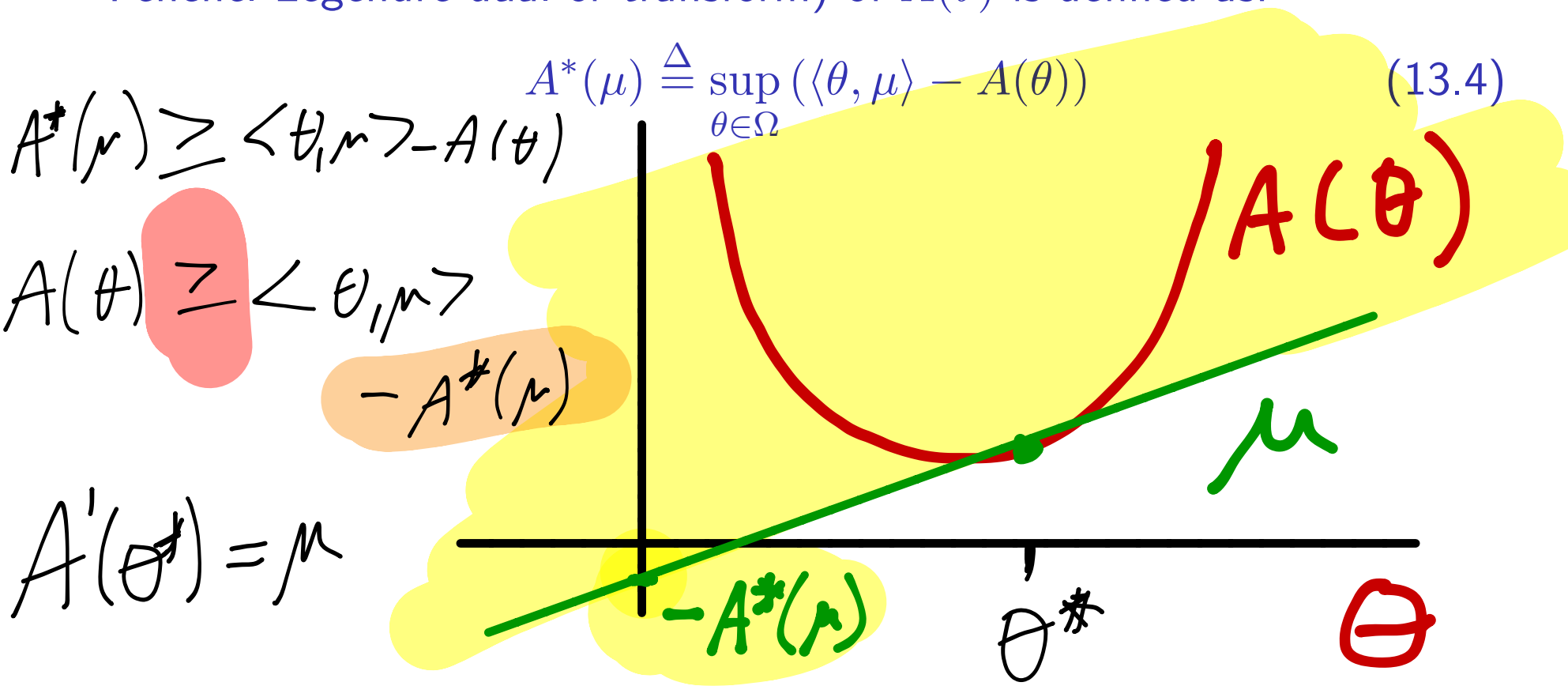
Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{13.3}$$

- Compare this to convex conjugate dual (also sometimes Fenchel-Legendre dual or transform) of $A(\theta)$ is defined as:

$$A^*(\mu) \triangleq \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \tag{13.4}$$



Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{13.3}$$

- Compare this to convex conjugate dual (also sometimes Fenchel-Legendre dual or transform) of $A(\theta)$ is defined as:

$$A^*(\mu) \triangleq \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \tag{13.4}$$

- 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.

Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{13.3}$$

- Compare this to convex conjugate dual (also sometimes Fenchel-Legendre dual or transform) of $A(\theta)$ is defined as:

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \tag{13.4}$$

- 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*

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu \tag{13.5}$$

Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{13.3}$$

- Compare this to convex conjugate dual (also sometimes Fenchel-Legendre dual or transform) of $A(\theta)$ is defined as:

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \tag{13.4}$$

- 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*

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu \tag{13.5}$$

- 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 13.3.3 (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:*

$$A^*(\mu) = \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \tag{13.6}$$

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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \tag{13.7}$$

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

$$\mu = \int_{\mathsf{D}_X} \phi(x) p_\theta(x) \nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta) \tag{13.8}$$

Conjugate Duality

- Note that $A*$ isn't exactly entropy, only entropy sometimes, and depends on matching parameters to μ via the matching mapping $\theta(\mu)$ which achieves

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \mu \tag{13.9}$$

Conjugate Duality

- Note that $A*$ isn't exactly entropy, only entropy sometimes, and depends on matching parameters to μ via the matching mapping $\theta(\mu)$ which achieves

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \mu \tag{13.9}$$

- $A(\theta)$ in Equation 13.7 is the "inference" problem (dual of the dual) for a given θ, since computing it involves computing the desired node/edge marginals.

Conjugate Duality

- Note that $A*$ isn't exactly entropy, only entropy sometimes, and depends on matching parameters to μ via the matching mapping $\theta(\mu)$ which achieves

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \mu \tag{13.9}$$

- $A(\theta)$ in Equation 13.7 is the "inference" problem (dual of the dual) for a given θ, since computing it involves computing the desired node/edge marginals.
- Whenever $\mu \notin \mathcal{M}$, then $A^*(\mu)$ returns ∞ which can't be the resulting sup in Equation 13.7, so Equation 13.7 need only consider $\mathcal{M}$.

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.7}$$

- computing $A(\theta)$ in this way corresponds to the inference problem (finding mean parameters, or node and edge marginals).

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.7}$$

- 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.**

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.7}$$

- 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. ☺

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.7}$$

- 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. ☹

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.7}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.7}$$

- Some good news: The above form gives us new avenues to do approximation. ☺

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.7}$$

- 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. ☺

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.7}$$

- 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.

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.7}$$

- 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). ☺☺

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.7}$$

- 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).

Overcomplete, simple notation

- We'll see: LBP (sum-product alg.) has much to do with an approximation to the aforementioned variational problems.

Overcomplete, simple notation

- We'll see: LBP (sum-product alg.) has much to do with an approximation to the aforementioned variational problems.
- Recall: dealing only with pairwise interactions (natural for image processing) – If not pairwise, we can convert from factor graph to factor graph with factor-width 2 factors.

Overcomplete, simple notation

- We'll see: LBP (sum-product alg.) has much to do with an approximation to the aforementioned variational problems.
- Recall: dealing only with pairwise interactions (natural for image processing) – If not pairwise, we can convert from factor graph to factor graph with factor-width 2 factors.
- Exponential overcomplete family model of form

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{v \in V(G)} \theta_v(x_v) + \sum_{(s,t) \in E(G)} \theta_{st}(x_s, x_t) \right\}$$

with simple new shorthand notation functions θ_v and θ_{st}.

$$\theta_v(x_v) \triangleq \sum_i \theta_{v,i} \mathbf{1}(x_v = i) \text{ and} \tag{13.10}$$

$$\theta_{s,t}(x_s, x_t) \triangleq \sum_{i,j} \theta_{st,ij} \mathbf{1}(x_s = i, x_t = j) \tag{13.11}$$

Marginal notation, and graph

Marginal polytope

- We also have mean parameters that constitute the marginal polytope.

$$\mu_v(x_v) \triangleq \sum_{i \in \mathsf{D}_{X_v}} \mu_{v,i} \mathbf{1}(x_v = i), \text{ for } u \in V(G) \tag{13.12}$$

$$\mu_{st}(x_s, x_t) \triangleq \sum_{(j,k) \in \mathsf{D}_{X_{\{s,t\}}}} \mu_{st,jk} \mathbf{1}(x_s = j, x_t = k), \text{ for } (s,t) \in E(G) \tag{13.13}$$

Marginal notation, and graph

Marginal polytope

- We also have mean parameters that constitute the marginal polytope.

$$\mu_v(x_v) \stackrel{\Delta}{=} \sum_{i \in \mathsf{D}_{X_v}} \mu_{v,i} \mathbf{1}(x_v = i), \text{ for } u \in V(G) \tag{13.12}$$

$$\mu_{st}(x_s, x_t) \stackrel{\Delta}{=} \sum_{(j,k) \in \mathsf{D}_{X_{\{s,t\}}}} \mu_{st,jk} \mathbf{1}(x_s = j, x_t = k), \text{ for } (s,t) \in E(G) \tag{13.13}$$

- And $\mathbb{M}(G)$ corresponds to the set of all singleton and pairwise marginals that can be jointly realized by some underlying probability distribution $p \in \mathcal{F}(G, \mathcal{M}^{(\mathsf{f})})$ that contains only pairwise interactions.

Marginal notation, and graph

Marginal polytope

- We also have mean parameters that constitute the marginal polytope.

$$\mu_v(x_v) \stackrel{\Delta}{=} \sum_{i \in \mathsf{D}_{X_v}} \mu_{v,i} \mathbf{1}(x_v = i), \text{ for } u \in V(G) \tag{13.12}$$

$$\mu_{st}(x_s, x_t) \stackrel{\Delta}{=} \sum_{(j,k) \in \mathsf{D}_{X_{\{s,t\}}}} \mu_{st,jk} \mathbf{1}(x_s = j, x_t = k), \text{ for } (s,t) \in E(G) \tag{13.13}$$

- And $\mathbb{M}(G)$ corresponds to the set of all singleton and pairwise marginals that can be jointly realized by some underlying probability distribution $p \in \mathcal{F}(G, \mathcal{M}^{(\mathsf{f})})$ that contains only pairwise interactions.
- Note, $\mathbb{M}(G)$ is respect to a graph G.

Marginal notation, and graph

Marginal polytope

- We also have mean parameters that constitute the marginal polytope.

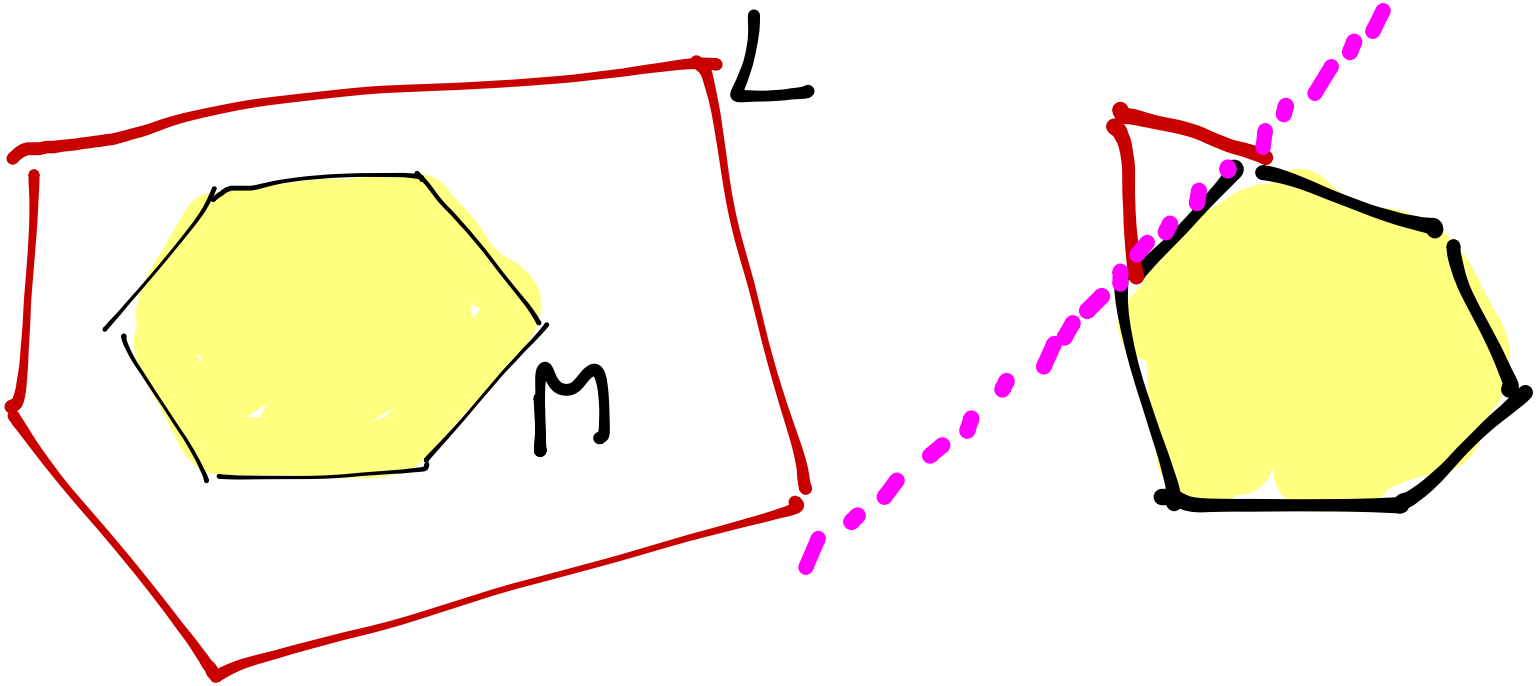
$$\mu_v(x_v) \triangleq \sum_{i \in \mathsf{D}_{X_v}} \mu_{v,i} \mathbf{1}(x_v = i), \text{ for } u \in V(G) \tag{13.12}$$

$$\mu_{st}(x_s, x_t) \triangleq \sum_{(j,k) \in \mathsf{D}_{X_{\{s,t\}}}} \mu_{st,jk} \mathbf{1}(x_s = j, x_t = k), \text{ for } (s,t) \in E(G) \tag{13.13}$$

- And $\mathbb{M}(G)$ corresponds to the set of all singleton and pairwise marginals that can be jointly realized by some underlying probability distribution $p \in \mathcal{F}(G, \mathcal{M}^{(\mathsf{f})})$ that contains only pairwise interactions.
- Note, $\mathbb{M}(G)$ is respect to a graph G.
- Recall, $\mathbb{M}$ can be represented as a convex hull of a set of points, or by a set of linear inequality constraints.

Local consistency polytope

- An "outer bound" of $\mathbb{M}$ consists of a set $\mathbb{L} \supseteq \mathbb{M}$ 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.

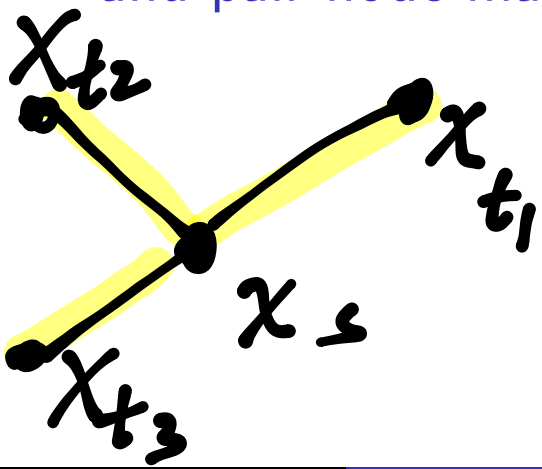


Local consistency polytope

- An "outer bound" of $\mathbb{M}$ consists of a set $\mathbb{L} \supseteq \mathbb{M}$ 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.
- Another way to form outer bound: require only 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

$$\sum_{x_v} \tau_v(x_v) = 1 \tag{13.14}$$

and pair-node marginal consistency constraints



$$\sum_{x'_t} \tau_{s,t}(x_s, x'_t) = \tau_s(x_s) \tag{13.15a}$$

$$\sum_{x'_s} \tau_{s,t}(x'_s, x_t) = \tau_t(x_t) \tag{13.15b}$$

Local consistency polytope

- Define $\mathbb{L}(G)$ to be the (locally consistent) polytope that obeys the constraints in Equations 13.14 and 13.15.
- 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.

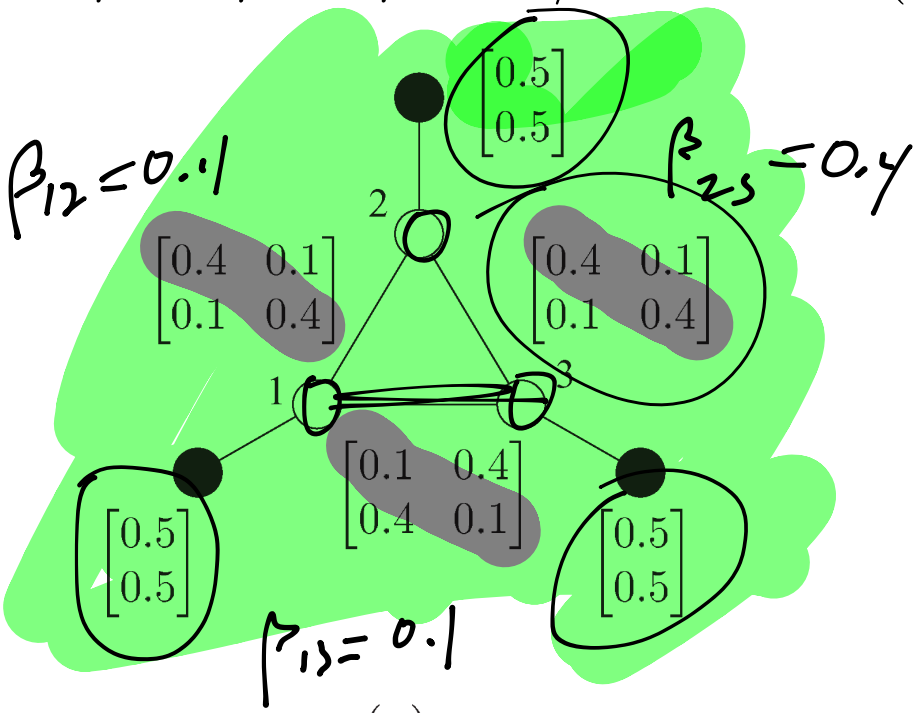
$\mathbb{L} \setminus \mathbb{M}$

Pseudo-marginals

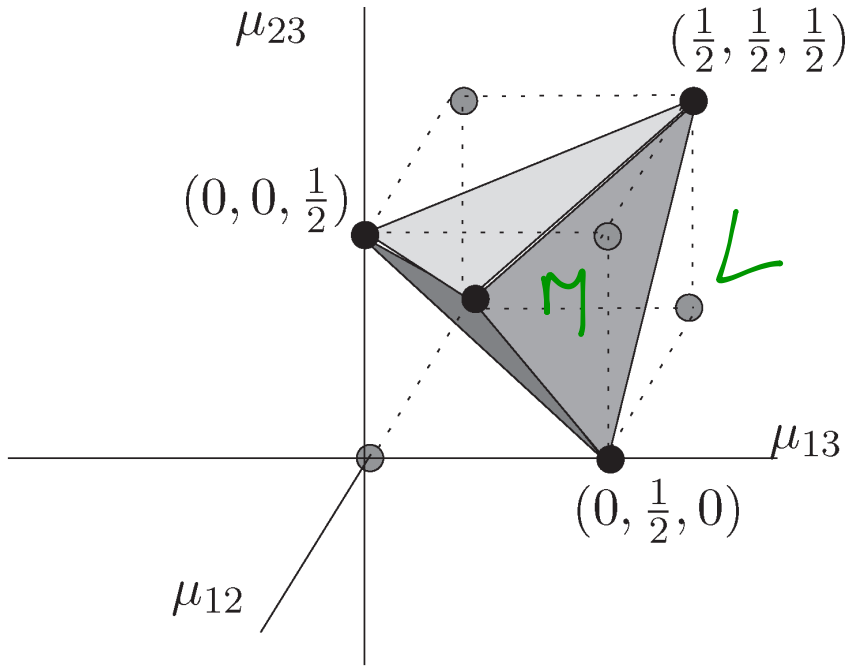
$$\tau_v(x_v) = [0.5, 0.5], \text{ and } \tau_{s,t}(x_s, x_t) = \begin{bmatrix} \beta_{st} & .5 - \beta_{st} \\ .5 - \beta_{st} & \beta_{st} \end{bmatrix} \quad (13.16)$$

$0 \le \beta_{st} \le 0.5$

- 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)



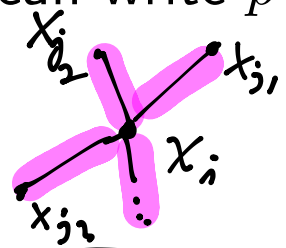
(b)

Bethe Entropy Approximation

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.7}$$

- So inference corresponds to Equation 13.7, and we have two difficulties $\mathcal{M}$ and $A^*(\mu)$.
- Maybe it is hard to compute $A^*(\mu)$ but perhaps we can reasonably approximate it.
- In case when $-A^*(\mu)$ is the entropy, lets use an approximate entropy based on $\mathbb{L}$ being those distributions that factor w.r.t. a tree.
- When $p \in \mathcal{F}(G, \mathcal{M}^{(\mathsf{f})})$ and G is a tree T, then we can write p as:

$$p(x_1, \ldots, x_N) = \frac{\prod_{(i,j) \in E(T)} p_{ij}(x_i, x_j)}{\prod_{v \in V(T)} p_v(x_v)^{d(v)-1}} \tag{13.17}$$



$$= \prod_{v \in V(T)} p_v(x_v) \prod_{(i,j) \in E(T)} \frac{p_{ij}(x_i, x_j)}{p_i(x_i) p_j(x_j)} \tag{13.18}$$

where $d(v)$ is the degree of v (shattering coefficient of v as separator)

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

$$p_\mu(x) = \prod_{s \in V(T)} \mu_s(x_s) \prod_{(s,t) \in E(T)} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{13.19}$$

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

$$p_\mu(x) = \prod_{s \in V(T)} \mu_s(x_s) \prod_{(s,t) \in E(T)} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{13.19}$$

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

$$-A^*(\mu) = H(p_\mu) = \sum_{v \in V(T)} H(X_v) - \sum_{(s,t) \in E(T)} I(X_s; X_t) \tag{13.20}$$

$$= \sum_{v \in V(T)} H_v(\mu_v) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \tag{13.21}$$

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

$$p_\mu(x) = \prod_{s \in V(T)} \mu_s(x_s) \prod_{(s,t) \in E(T)} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{13.19}$$

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

$$-A^*(\mu) = H(p_\mu) = \sum_{v \in V(T)} H(X_v) - \sum_{(s,t) \in E(T)} I(X_s; X_t) \tag{13.20}$$

$$= \sum_{v \in V(T)} H_v(\mu_v) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \tag{13.21}$$

- 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 Entropy Approximation

- We can perhaps just use this as an approximation, i.e., say that for **any** graph $G = (V, E)$ not nec. a tree.

Bethe Entropy Approximation

- We can perhaps just use this as an approximation, i.e., say that for **any** graph $G = (V, E)$ not nec. a tree.
- That is, assuming that the distribution is structured over pairwise potential functions w.r.t. a graph G, we can make an approximation to $-A^*(\tau)$ based on equation that has same form, i.e.,

$$-A^*(\tau) \approx H_{\text{Bethe}}(\tau) \stackrel{\Delta}{=} \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \quad (13.22)$$

$$= \sum_{v \in V(G)} (d(v) - 1) H_v(\tau_v) + \sum_{(i,j) \in E(G)} H_{st}(\tau_s, \tau_t) \quad (13.23)$$

Bethe Entropy Approximation

- We can perhaps just use this as an approximation, i.e., say that for **any** graph $G = (V, E)$ not nec. a tree.
- That is, assuming that the distribution is structured over pairwise potential functions w.r.t. a graph G, we can make an approximation to $-A^*(\tau)$ based on equation that has same form, i.e.,

$$-A^*(\tau) \approx H_{\text{Bethe}}(\tau) \triangleq \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \quad (13.22)$$

$$= \sum_{v \in V(T)} (d(v) - 1) H_v(\tau_v) + \sum_{(i,j) \in E(T)} H_{st}(\tau_s, \tau_t) \quad (13.23)$$

- Key: $H_{\text{Bethe}}(\tau)$ is not necessarily concave as it is not a real entropy.

Bethe Entropy Approximation

- We can perhaps just use this as an approximation, i.e., say that for **any** graph $G = (V, E)$ not nec. a tree.
- That is, assuming that the distribution is structured over pairwise potential functions w.r.t. a graph G, we can make an approximation to $-A^*(\tau)$ based on equation that has same form, i.e.,

$$-A^*(\tau) \approx H_{\text{Bethe}}(\tau) \triangleq \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \quad (13.22)$$

$$= \sum_{v \in V(T)} (d(v) - 1) H_v(\tau_v) + \sum_{(i,j) \in E(T)} H_{st}(\tau_s, \tau_t) \quad (13.23)$$

- Key: $H_{\text{Bethe}}(\tau)$ is not necessarily concave as it is not a real entropy.
- MI equation is not hard to compute $O(r^2)$.

$$I_{st}(\tau_{st}) = I_{st}(\tau_{st}(x_s, x_t)) \quad (13.24)$$

$$= \sum_{x_s, x_t} \tau_{st}(x_s, x_t) \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} \quad (13.25)$$

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.26}$$

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.26}$$

Approximate variational representation of log partition function

$$A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{L}} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\} \tag{13.27}$$

$$= \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \tag{13.28}$$

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.26}$$

Approximate variational representation of log partition function

$$A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{L}} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\} \tag{13.27}$$

$$= \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \tag{13.28}$$

- Exact when $G = T$ but we do this for any G, still commutable

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.26}$$

Approximate variational representation of log partition function

$$A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{L}} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\} \tag{13.27}$$

$$= \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \tag{13.28}$$

- 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.

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.26}$$

Approximate variational representation of log partition function

$$A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{L}} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\} \tag{13.27}$$

$$= \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \tag{13.28}$$

- 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.

Bethe Variational Problem and LBP

- Lagrangian constraints for summing to unity at nodes

$$C_{vv}(\tau) = 1 - \sum_{x_v} \tau_v(x_v) \tag{13.29}$$

- Lagrangian constraints for local consistency

$$C_{ts}(x_s;\tau) = \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t) \tag{13.30}$$

- Yields following Lagrangian

$$\mathcal{L}(\tau, \lambda; \theta) = \langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau) + \sum_{v \in V} \lambda_{vv} C_{vv}(\tau) \tag{13.31}$$

$$+ \sum_{(s,t) \in E(G)} \left[\sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s;\tau) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t;\tau) \right] \tag{13.32}$$

Fixed points: Variational Problem and LBP

Theorem 13.5.1

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.*

$$\nabla_\tau \mathcal{L}(\tau^*, \lambda^*; \theta) = 0 \text{ and } \nabla_\lambda \mathcal{L}(\tau^*, \lambda^*; \theta) = 0 \quad (13.33)$$

(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.

Fixed points: Variational Problem and LBP

- The resulting Lagrange multipliers λ_{st} end up being exactly the messages that we have defined. I.e., we get

$$\lambda_{st}(x_t) = \mu_{s \to t}(x_t) = \sum_{x_s} \psi_{s,t}(x_s, x_t) \prod_{k \in \delta(s) \setminus \{t\}} \mu_{k \to s}(x_s) \quad (13.34)$$

Fixed points: Variational Problem and LBP

- The resulting Lagrange multipliers λ_{st} end up being exactly the messages that we have defined. I.e., we get

$$\lambda_{st}(x_t) = \mu_{s \to t}(x_t) = \sum_{x_s} \psi_{s,t}(x_s, x_t) \prod_{k \in \delta(s) \setminus \{t\}} \mu_{k \to s}(x_s) \qquad (13.34)$$

- Proof: take derivatives of Lagrangian, set equal to zero, use Lagrangian constraints, do a bit of algebra, and amazingly, the BP messages suddenly pop out!!! (see page 86 in book).

Fixed points: Variational Problem and LBP

- The resulting Lagrange multipliers λ_{st} end up being exactly the messages that we have defined. I.e., we get

$$\lambda_{st}(x_t) = \mu_{s \to t}(x_t) = \sum_{x_s} \psi_{s,t}(x_s, x_t) \prod_{k \in \delta(s) \setminus \{t\}} \mu_{k \to s}(x_s) \qquad (13.34)$$

- Proof: take derivatives of Lagrangian, set equal to zero, use Lagrangian constraints, do a bit of algebra, and amazingly, the BP messages suddenly pop out!!! (see page 86 in book).
- So we can now (at least) characterize any stable point of LBP.

Fixed points: Variational Problem and LBP

- The resulting Lagrange multipliers λ_{st} end up being exactly the messages that we have defined. I.e., we get

$$\lambda_{st}(x_t) = \mu_{s \to t}(x_t) = \sum_{x_s} \psi_{s,t}(x_s, x_t) \prod_{k \in \delta(s) \setminus \{t\}} \mu_{k \to s}(x_s) \qquad (13.34)$$

- Proof: take derivatives of Lagrangian, set equal to zero, use Lagrangian constraints, do a bit of algebra, and amazingly, the BP messages suddenly pop out!!! (see page 86 in book).
- So we can now (at least) characterize any stable point of LBP.
- This does not mean that it will converge.

Fixed points: Variational Problem and LBP

- The resulting Lagrange multipliers λ_{st} end up being exactly the messages that we have defined. I.e., we get

$$\lambda_{st}(x_t) = \mu_{s\to t}(x_t) = \sum_{x_s} \psi_{s,t}(x_s, x_t) \prod_{k\in\delta(s)\setminus\{t\}} \mu_{k\to s}(x_s) \qquad (13.34)$$

- Proof: take derivatives of Lagrangian, set equal to zero, use Lagrangian constraints, do a bit of algebra, and amazingly, the BP messages suddenly pop out!!! (see page 86 in book).
- So we can now (at least) characterize any stable point of LBP.
- This does not mean that it will converge.
- For trees, we'll get $A_{\text{Bethe}}(\theta) = A(\theta)$, results of previous lectures (parallel or MPP-based message passing).

Bounds on A: why would we want them?

- Does **not** mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why want bounds?

Bounds on A: why would we want them?

- Does **not** mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why want bounds? Recall Max. Likelihood

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right)$$

(??)

Bounds on A: why would we want them?

- Does **not** mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why want bounds? Recall Max. Likelihood

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

and convex conjugate dual of $A(\theta)$

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

Bounds on A: why would we want them?

- Does **not** mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why want bounds? Recall Max. Likelihood

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

and convex conjugate dual of $A(\theta)$

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

- Recall again the expression for the partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \qquad (13.7)$$

Bounds on A: why would we want them?

- Does **not** mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why want bounds? Recall Max. Likelihood

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

and convex conjugate dual of $A(\theta)$

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

- Recall again the expression for the partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \qquad (13.7)$$

and some approximation to $A(\theta)$, say $A_{\text{approx}}(\theta)$.

Bounds on A: why would we want them?

- Does **not** mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why want bounds? Recall Max. Likelihood

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

and convex conjugate dual of $A(\theta)$

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

- Recall again the expression for the partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \qquad (13.7)$$

and some approximation to $A(\theta)$, say $A_{\text{approx}}(\theta)$.

- Due to sup in Eq. (??), might want upper bound $A_{\text{approx}}(\theta) \geq A(\theta)$,

Bounds on A: why would we want them?

- Does **not** mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why want bounds? Recall Max. Likelihood

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

and convex conjugate dual of $A(\theta)$

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

- Recall again the expression for the partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \qquad (13.7)$$

and some approximation to $A(\theta)$, say $A_{\text{approx}}(\theta)$.

- Due to $\sup$ in Eq. **(??)**, might want upper bound $A_{\text{approx}}(\theta) \geq A(\theta)$,
- mean-field methods (ch 5 in book) provides lower bound on $A(\theta)$.

Bounds on A: why would we want them?

- Does **not** mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why want bounds? Recall Max. Likelihood

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

and convex conjugate dual of $A(\theta)$

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \qquad \textbf{(??)}$$

- Recall again the expression for the partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \qquad (13.7)$$

and some approximation to $A(\theta)$, say $A_{\text{approx}}(\theta)$.

- Due to $\sup$ in Eq. (**??**), might want upper bound $A_{\text{approx}}(\theta) \geq A(\theta)$,
- mean-field methods (ch 5 in book) provides lower bound on $A(\theta)$.
- For certain "attractive" potential functions, we get $A_{\text{Bethe}}(\theta) \leq A(\theta)$, these are common in computer vision and are related to graph cuts.

Bounds on A

- In general, ideally we would like methods that give us (as tight as possible) bounds, and we can use both upper and lower bounds.

Bounds on A

- In general, ideally we would like methods that give us (as tight as possible) bounds, and we can use both upper and lower bounds.
- Recall definition of the family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{13.35}$$

Bounds on A

- In general, ideally we would like methods that give us (as tight as possible) bounds, and we can use both upper and lower bounds.
- Recall definition of the family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{13.35}$$

- So bounds on A can give us bounds on p. E.g., lower bounds on A will give us upper bounds on p.

Bounds on A

- In general, ideally we would like methods that give us (as tight as possible) bounds, and we can use both upper and lower bounds.
- Recall definition of the family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{13.35}$$

- So bounds on A can give us bounds on p. E.g., lower bounds on A will give us upper bounds on p.
- To compute conditionals

$$p(x_A|x_B) = \frac{p(x_{A\cup B})}{p(x_B)} = \frac{\sum_{x_{V\setminus(A\cup B)}} p(x)}{\sum_{x_{V\setminus B}} p(x)} \tag{13.36}$$

we would like both upper and lower bounds on A depending on if we want to upper or lower bound probability estimates.

Bounds on A

- In general, ideally we would like methods that give us (as tight as possible) bounds, and we can use both upper and lower bounds.
- Recall definition of the family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{13.35}$$

- So bounds on A can give us bounds on p. E.g., lower bounds on A will give us upper bounds on p.
- To compute conditionals

$$p(x_A|x_B) = \frac{p(x_{A\cup B})}{p(x_B)} = \frac{\sum_{x_{V\setminus(A\cup B)}} p(x)}{\sum_{x_{V\setminus B}} p(x)} \tag{13.36}$$

  we would like both upper and lower bounds on A depending on if we want to upper or lower bound probability estimates.
- Perhaps more importantly, $\exp(A(\theta))$ is a marginal in and of itself (recall it is marginalization over everything). If we can bound $A(\theta)$, we can come up with other forms of bounds over other marginals.

Lack of bounds for Bethe

- Two reasons A might be inaccurate:

Lack of bounds for Bethe

- Two reasons A might be inaccurate: 1) We have replaced $\mathbb{M}$ with outer bound $\mathbb{L}$;

Lack of bounds for Bethe

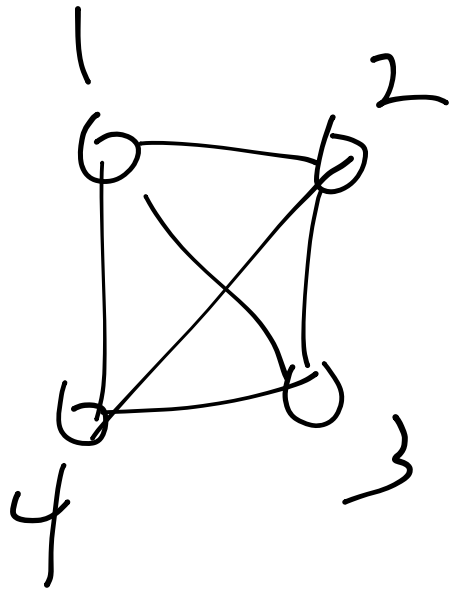
- Two reasons A might be inaccurate: 1) We have replaced $\mathbb{M}$ with outer bound $\mathbb{L}$; and 2) we've used H_{Bethe} in place of the true dual A^*.

Lack of bounds for Bethe

- Two reasons A might be inaccurate: 1) We have replaced $\mathbb{M}$ with outer bound $\mathbb{L}$; and 2) we've used H_{Bethe} in place of the true dual A^*.
- Example of inaccuracy (example 4.2 from book), consider a 4-clique

$$\mu_s(x_s) = [0.5 \;\; 0.5] \qquad \text{for } s = 1, 2, 3, 4 \qquad (13.37a)$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \forall (s,t) \in E(G) \qquad (13.37b)$$



Lack of bounds for Bethe

- Two reasons A might be inaccurate: 1) We have replaced $\mathbb{M}$ with outer bound $\mathbb{L}$; and 2) we've used H_{Bethe} in place of the true dual A^*.
- Example of inaccuracy (example 4.2 from book), consider a 4-clique

$$\mu_s(x_s) = [0.5 \;\; 0.5] \qquad \text{for } s = 1, 2, 3, 4 \tag{13.37a}$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \forall (s,t) \in E(G) \tag{13.37b}$$

- Valid marginals, equal 0.5 probability for $(0,0,0,0)$ and $(1,1,1,1)$.

Lack of bounds for Bethe

- Two reasons A might be inaccurate: 1) We have replaced $\mathbb{M}$ with outer bound $\mathbb{L}$; and 2) we've used H_{Bethe} in place of the true dual A^*.
- Example of inaccuracy (example 4.2 from book), consider a 4-clique

$$\mu_s(x_s) = [0.5 \;\; 0.5] \qquad \text{for } s = 1, 2, 3, 4 \tag{13.37a}$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \forall (s,t) \in E(G) \tag{13.37b}$$

- Valid marginals, equal 0.5 probability for $(0,0,0,0)$ and $(1,1,1,1)$.
- Each $H_s(\mu_s) = \log 2$, and each $I_{st}(\mu_{st}) = \log 2$ giving

$$H_{\text{Bethe}}(\mu) = 4\log 2 - 6\log 2 = -2\log 2 < 0 \tag{13.38}$$

Lack of bounds for Bethe

- Two reasons A might be inaccurate: 1) We have replaced $\mathbb{M}$ with outer bound $\mathbb{L}$; and 2) we've used H_{Bethe} in place of the true dual A^*.
- Example of inaccuracy (example 4.2 from book), consider a 4-clique

$$\mu_s(x_s) = [0.5 \;\; 0.5] \qquad \text{for } s = 1, 2, 3, 4 \tag{13.37a}$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \forall (s,t) \in E(G) \tag{13.37b}$$

- Valid marginals, equal 0.5 probability for $(0,0,0,0)$ and $(1,1,1,1)$.
- Each $H_s(\mu_s) = \log 2$, and each $I_{st}(\mu_{st}) = \log 2$ giving

$$H_{\text{Bethe}}(\mu) = 4 \log 2 - 6 \log 2 = -2 \log 2 < 0 \tag{13.38}$$

which obviously can't be a true entropy since we must have $H > 0$ for discrete distributions.

Lack of bounds for Bethe

- Two reasons A might be inaccurate: 1) We have replaced $\mathbb{M}$ with outer bound $\mathbb{L}$; and 2) we've used H_{Bethe} in place of the true dual A^*.
- Example of inaccuracy (example 4.2 from book), consider a 4-clique

$$\mu_s(x_s) = [0.5 \;\; 0.5] \qquad \text{for } s = 1, 2, 3, 4 \tag{13.37a}$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \forall (s,t) \in E(G) \tag{13.37b}$$

- Valid marginals, equal 0.5 probability for $(0,0,0,0)$ and $(1,1,1,1)$.
- Each $H_s(\mu_s) = \log 2$, and each $I_{st}(\mu_{st}) = \log 2$ giving

$$H_{\text{Bethe}}(\mu) = 4 \log 2 - 6 \log 2 = -2 \log 2 < 0 \tag{13.38}$$

  which obviously can't be a true entropy since we must have $H > 0$ for discrete distributions.
- True $-A^*(\mu) = \log 2 > 0$.

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)?

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.

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. ☹

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., $\phi_i(x_i) = \sum_{x_{V \setminus \{i\}}} p(x)$.

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.

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 identical when global renormalized.

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 identical when global renormalized.
- That is, we have

Reparameterization Properties of Bethe Approximation

Proposition 13.5.2

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

$$p_{\tau^*}(x) \triangleq \frac{1}{Z(\tau^*)} \prod_{s \in V} \tau_s^*(x_s) \prod_{(s,t) \in E(G)} \frac{\tau_{st}^*(x_s, x_t)}{\tau_s^*(x_s)\tau_t^*(x_t)} \tag{13.39}$$

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)

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

- Consider

$$\theta_s(x_s) = \log \tau_s(x_s) = \log[0.5 \;\; 0.5] \qquad \text{for } s = 1, 2, 3, 4 \tag{13.40a}$$

$$\theta_{st}(x_s, x_t) = \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)}$$

$$= \log 4 \begin{bmatrix} \beta_{st} & 0.5 - \beta_{st} \\ 0.5 - \beta_{st} & \beta_{st} \end{bmatrix} \; \forall (s,t) \in E(G) \tag{13.40b}$$

- 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. ☹

If not Bounds, then Better Approximation?

- So we want bounds between $A(\theta)$ and $A_{\text{Bethe}}(\theta)$ in the ideal case.

If not Bounds, then Better Approximation?

- So we want bounds between $A(\theta)$ and $A_{\text{Bethe}}(\theta)$ in the ideal case.
- Perhaps we can come up with an expression for $A(\theta) - A_{\text{Bethe}}(\theta)$

If not Bounds, then Better Approximation?

- So we 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?

If not Bounds, then Better Approximation?

- So we 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.

If not Bounds, then Better Approximation?

- So we 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.
- This is the idea behind Loop Series Expansions

Generalized Loops

- Recall vertex and edge induced subgraphs.
- Notation: Given graph $G = (V, E)$, we have
- Given subset $S \subseteq V$, then $G' = (S, E(S))$ is a vertex induced subgraph.
- Given subset $\tilde{E} \subseteq E$, then $G(\tilde{E}) = (V(\tilde{E}), \tilde{E})$ is edge-induced subgraph.
- Define the degree 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:

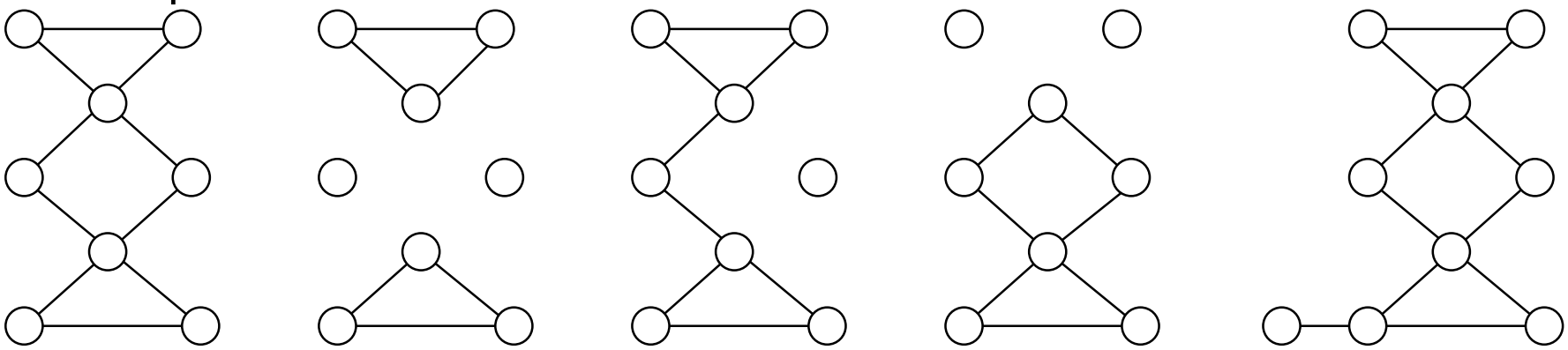


Illustration of generalized loops. (a) An original graph. (b)-(d) Various generalized loops associated with the graph in (a). In this particular case, the original graph is a generalized loop for itself. (e) is not a generalized loop as it has a leaf node.

Edge weights Generalized Loops

- Consider LBP fixed point for binary pairwise MRF (Ising model), and with unary and pairwise pseudomarginals parameterized as:

$$\tau_s(x_s) = \begin{bmatrix} 1-\tau_s \\ \tau_s \end{bmatrix}, \text{ and } \tau_{st}(x_s, x_t) = \begin{bmatrix} 1-\tau_s-\tau_t+\tau_{st} & \tau_t - \tau_{st} \\ \tau_s - \tau_{st} & \tau_{st} \end{bmatrix} \tag{13.41}$$

- Define edge weight as

$$\beta_{st} \triangleq \frac{\tau_{st} - \tau_s\tau_t}{\tau_s(1-\tau_s)\tau_t(1-\tau_t)} \tag{13.42}$$

- and extended to a general set of edges $\tilde{E}$

$$\beta_{st} \triangleq \prod_{(s,t)\in\tilde{E}} \beta_{st} \tag{13.43}$$

Comparison of A and A_{Bethe}

Proposition 13.6.1

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 $A(\theta)$.

$$A(\theta) = A_{Bethe}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \tag{13.44}$$

- For any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, inner term is zero and vanishes.

Comparison of A and A_{Bethe}

Proposition 13.6.1

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 $A(\theta)$.

$$A(\theta) = A_{Bethe}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \tag{13.44}$$

- For any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, inner term is zero and vanishes. why?

Comparison of A and A_{Bethe}

Proposition 13.6.1

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 $A(\theta)$.

$$A(\theta) = A_{Bethe}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \tag{13.44}$$

- For any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, inner term is zero and vanishes. why? Thus, terms in the sum only exists for generalized loops.

Comparison of A and A_{Bethe}

Proposition 13.6.1

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 $A(\theta)$.

$$A(\theta) = A_{Bethe}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \tag{13.44}$$

- For any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, inner term is zero and vanishes. why? Thus, terms in the sum only exists for generalized loops.
- The generalized loops give the correction!

Comparison of A and A_{Bethe}

Proposition 13.6.1

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 $A(\theta)$.

$$A(\theta) = A_{Bethe}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \tag{13.44}$$

- For any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, inner term is zero and vanishes. why? 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 13.6.1

proof sketch.

- Overcomplete, $\exists$ parameters $\hat{\theta}$ s.t. $\left\langle \hat{\theta}, \phi(x) \right\rangle = c$ for all x.

Proof of Proposition 13.6.1

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.

Proof of Proposition 13.6.1

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.
- Choose parameterization

$$\tilde{\theta}_s(x_s) = \log \tau_s(x_s), \text{ and } \tilde{\theta}_{st}(x_s, x_t) = \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} \quad (13.45)$$

Proof of Proposition 13.6.1

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.
- Choose parameterization

$$\tilde{\theta}_s(x_s) = \log \tau_s(x_s), \text{ and } \tilde{\theta}_{st}(x_s, x_t) = \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} \tag{13.45}$$

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

Proof of Proposition 13.6.1

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.
- Choose parameterization

$$\tilde{\theta}_s(x_s) = \log \tau_s(x_s), \text{ and } \tilde{\theta}_{st}(x_s, x_t) = \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} \quad (13.45)$$

- 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

$$A(\tilde{\theta}) = \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \quad (13.46)$$

Proof of Proposition 13.6.1 cont.

proof sketch.

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

$$\frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} = 1 + \beta_{st}(x_s - \tau_s)(x_t - \tau_t) \tag{13.47}$$

Proof of Proposition 13.6.1 cont.

proof sketch.

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

$$\frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} = 1 + \beta_{st}(x_s - \tau_s)(x_t - \tau_t) \tag{13.47}$$

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

$$\exp(A(\tilde{\theta})) = \sum_{x \in \{0,1\}^m} \prod_{s \in V} \tau_s(x_s) \prod_{(s,t) \in E} \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} \tag{13.48}$$

Proof of Proposition 13.6.1 cont.

proof sketch.

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

$$\frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} = 1 + \beta_{st}(x_s - \tau_s)(x_t - \tau_t) \tag{13.47}$$

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

$$\exp(A(\tilde{\theta})) = \sum_{x \in \{0,1\}^m} \prod_{s \in V} \tau_s(x_s) \prod_{(s,t) \in E} \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} \tag{13.48}$$

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

$$\exp(A(\tilde{\theta})) = \mathbb{E}\left[\prod_{(s,t) \in E} (1 + \beta_{st}(X_s - \tau_s)(X_t - \tau_t))\right] \tag{13.49}$$

Proof of Proposition 13.6.1 cont.

proof sketch.

- By polynomial expansion, linearity of expectation, we get

$$\exp(A(\tilde{\theta})) = 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \mathbb{E}\left[\prod_{(s,t)\in\tilde{E}} (\beta_{st}(X_s - \tau_s)(X_t - \tau_t))\right] \quad (13.50)$$

Proof of Proposition 13.6.1 cont.

proof sketch.

- By polynomial expansion, linearity of expectation, we get

$$\exp(A(\tilde{\theta})) = 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \mathbb{E}\left[\prod_{(s,t)\in\tilde{E}} (\beta_{st}(X_s - \tau_s)(X_t - \tau_t))\right] \quad (13.50)$$

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

$$\exp(A(\tilde{\theta})) = 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s\in V} \mathbb{E}_{\tau_s}\left[(X_s - \tau_s)^{d_s(\tilde{E})}\right] \quad (13.51)$$

□

Comparison of A and A_{Bethe}

Proposition 13.6.1

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 $A(\theta)$.

$$A(\theta) = A_{Bethe}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \tag{13.44}$$

- For any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, inner term is zero and vanishes. why? 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).

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.52}$$

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.52}$$

- So far, we have used a replacement for $-A^*(\mu)$ inspired by trees.

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.52}$$

- 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?

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.52}$$

- 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?

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.52}$$

- 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).

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.52}$$

- 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:

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.52}$$

- 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;

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.52}$$

- 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

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.52}$$

- 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?

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{13.52}$$

- 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.

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.

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.

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.

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 maxcliques (which are clusters of graph nodes) in the junction tree are the edges of the hypertree.

Hypergraphs

Definition 13.7.1 (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.

Hypergraphs

Definition 13.7.1 (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 13.7.2 (leaf)

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

Hypergraphs

Definition 13.7.1 (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 13.7.2 (leaf)

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

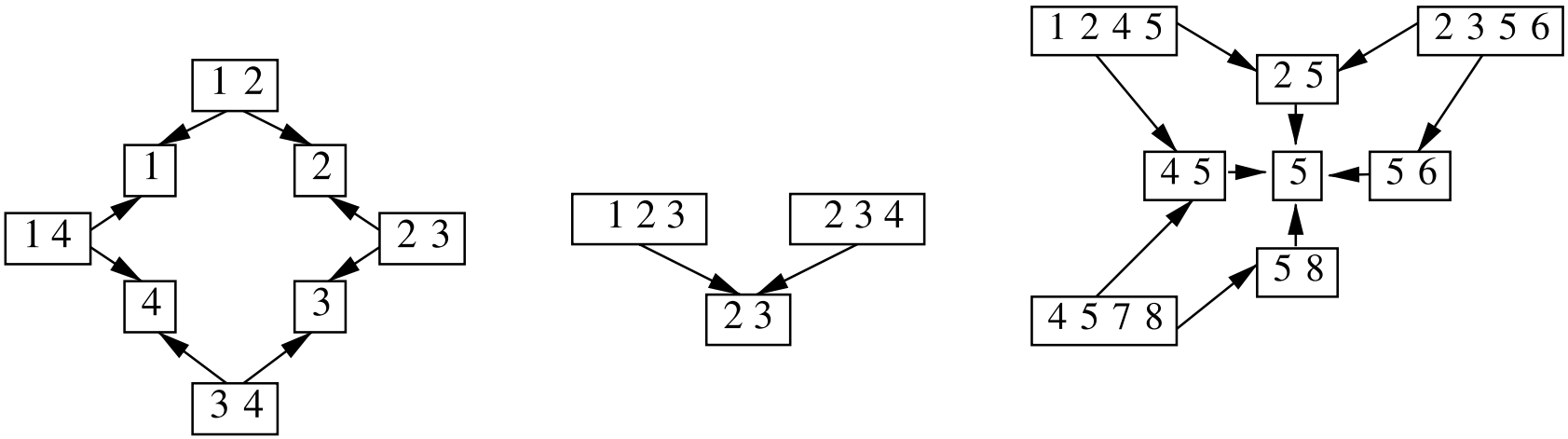
Definition 13.7.3 (acyclic)

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:

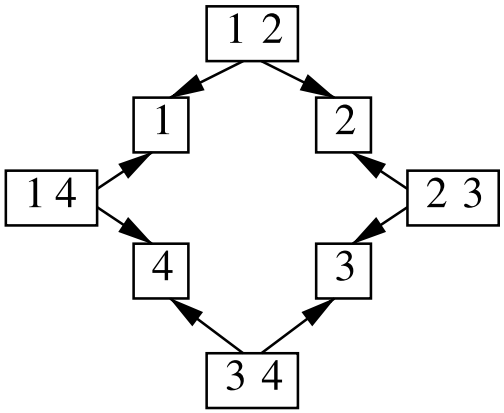
Hypergraphs and posets



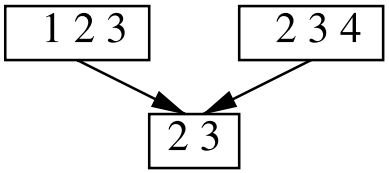
(a) (b) (c)

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 cycle graph represented as a hypergraph. (b) A simple hypertree of width two. (c) A more complex hypertree of width three.

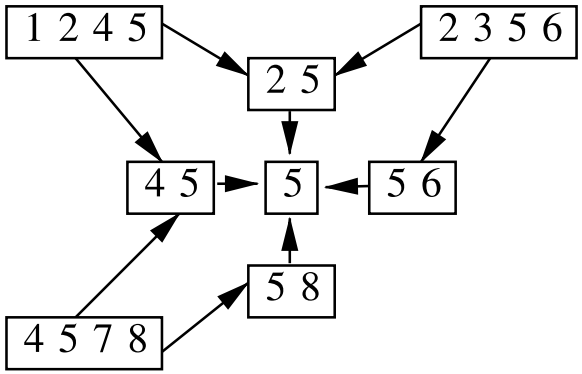
Hypergraphs and posets



(a)



(b)



(c)

As bipartite graphs:

Partially ordered set

- A partially ordered set (poset) is a set $\mathcal{P}$ of objects with an order.

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".

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.

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.)

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

$$\zeta(g,h) = \begin{cases} 1 & \text{if } g \preceq h, \\ 0 & \text{otherwise.} \end{cases} \tag{13.53}$$

- 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 $h : h \not\subseteq g$.
- Given $\omega(g,f)$ defined for f such that $g \subseteq f \subseteq h$, we define

$$\omega(g,h) = -\sum_{\{f | g \subseteq f \subset h\}} \omega(g,f) \tag{13.54}$$

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

$$\sum_{f \in \mathcal{P}} \omega(g,f)\zeta(f,h) = \sum_{\{f | g \subseteq f \subseteq h\}} \omega(g,f) = \delta(g,h) \tag{13.55}$$

General Möbius Inversion Lemma

Lemma 13.7.4

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

$$\Omega(h) = \sum_{g \preceq h} \Upsilon(g) \quad \text{for all } h \in \mathcal{P} \tag{13.56}$$

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

$$\Upsilon(h) = \sum_{g \preceq h} \Omega(g)\omega(g,h) \quad \text{for all } h \in \mathcal{P} \tag{13.57}$$

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 13.7.5 (Möbius Inversion Lemma)

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 group (closure, associativity, identity, and inverse) for which the elements also commute, the real numbers being just one example). The following two equations imply each other.

$$\forall A \subseteq V : \Upsilon(A) = \sum_{B:B\subseteq A} \Omega(B) \tag{13.58}$$

$$\forall A \subseteq V : \Omega(A) = \sum_{B:B\subseteq A} (-1)^{|A\setminus B|}\Upsilon(B) \tag{13.59}$$

Proof of Möbius Inversion Lemma

Proof.

$$\sum_{B:B\subseteq A}\Omega(B) = \sum_{B:B\subseteq A}\sum_{C:C\subseteq B}(-1)^{|B\setminus C|}\Upsilon(C) \tag{13.60}$$

$$= \sum_{C:C\subseteq A}\sum_{B:C\subseteq B\&B\subseteq A}\Upsilon(C)(-1)^{|B\setminus C|} \tag{13.61}$$

$$= \sum_{C:C\subseteq A}\Upsilon(C)\sum_{B:C\subseteq B\&B\subseteq A}(-1)^{|B\setminus C|} \tag{13.62}$$

$$= \sum_{C:C\subseteq A}\Upsilon(C)\sum_{H:H\subseteq A\setminus C}(-1)^{|H|} \tag{13.63}$$

Proof of Möbius Inversion Lemma

Proof Cont.

Also, note that for some set D,

$$\sum_{H:H\subseteq D}(-1)^{|H|} = \sum_{i=0}^{|D|}\binom{|D|}{i}(-1)^i = \sum_{i=0}^{|D|}\binom{|D|}{i}(-1)^i(1)^{|D|-i} \quad (13.64)$$

$$= (1-1)^{|D|} = \begin{cases} 1 & \text{if } |D| = 0 \\ 0 & \text{otherwise} \end{cases} \quad (13.65)$$

which means

$$\sum_{H:H\subseteq A\setminus C}(-1)^{|H|} = \begin{cases} 1 & \text{if } A = C \\ 0 & \text{otherwise} \end{cases} \quad (13.66)$$

Proof of Möbius Inversion Lemma

Proof Cont.

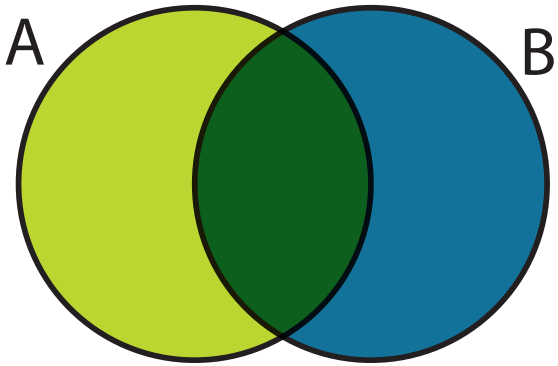
This gives

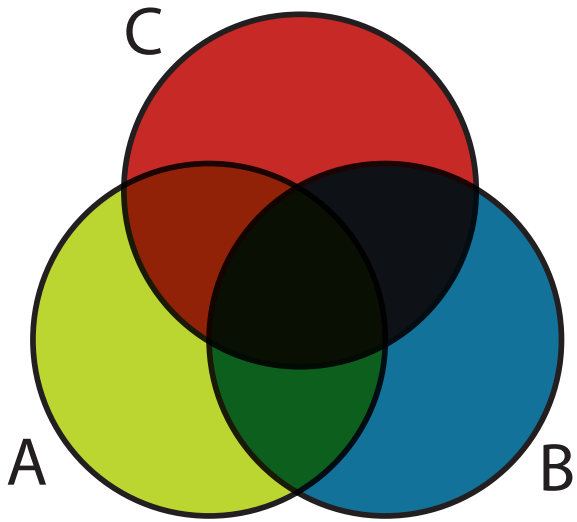
$$\sum_{B:B\subseteq A} \Omega(B) = \sum_{C:C\subseteq A} \Upsilon(C)\mathbf{1}\{A = C\} = \Upsilon(A) \tag{13.67}$$

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

Möbius Inversion Lemma and Inclusion-Exclusion

- This is a general cased 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

- General form of Inclusion-Exclusion: Given $A_1, A_2, \ldots, A_n \subseteq V$,

$$|\cup_{i=1}^{n} A_n| = \sum_{j=1}^{n} (-1)^{j-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} |A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_j}| \quad (13.68)$$

- This is a special case of Möbius Inversion Lemma:

$$\forall A \subseteq V : \Upsilon(A) = \sum_{B:B \subseteq A} \Omega(B) \quad (13.69)$$

$$\forall A \subseteq V : \Omega(A) = \sum_{B:B \subseteq A} (-1)^{|A \setminus B|} \Upsilon(B) \quad (13.70)$$

- 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).

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

$$\log \varphi_h(x_h) \triangleq \sum_{g \preceq h} \omega(g, h) \log \mu_g(x_g) \tag{13.71}$$

(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

$$\log \mu_h(x_h) = \sum_{g \preceq h} \log \varphi_g(x_g) \tag{13.72}$$

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

$$p_\mu(x) = \prod_{h \in E} \varphi_h(x_h) \tag{13.73}$$

$\Rightarrow$ 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

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

$$\varphi_{st}(x_s, x_t) = \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{13.74}$$

giving us the tree factorization we saw early in this course.

expressions of factorization and Möbius

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

$$\varphi_{st}(x_s, x_t) = \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{13.74}$$

  giving us the tree factorization we saw early in this course.

- For more general hypertree, consider edge set $E = \{(12345), (2356), (4578), (25), (45), (56), (58), (5)\}$. Check: is this a junction tree of cliques?

expressions of factorization and Möbius

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

$$\varphi_{st}(x_s, x_t) = \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{13.74}$$

  giving us the tree factorization we saw early in this course.
- For more general hypertree, consider edge set $E = \{(12345), (2356), (4578), (25), (45), (56), (58), (5)\}$. Check: is this a junction tree of cliques?
- Then

$$\varphi_{1245} = \frac{\mu_{1245}}{\varphi_{25}\varphi_{45}\varphi_5} = \frac{\mu_{1245}}{\frac{\mu_{25}}{\mu_5}\frac{\mu_{45}}{\mu_5}\mu_5} = \frac{\mu_{1245}\mu_5}{\mu_{25}\mu_{45}} \tag{13.75}$$

New expressions of entropy

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

$$H_h(\mu_h) = -\sum_{x_h} \mu_h(x_h) \log \mu_h(x_h) \tag{13.76}$$

and the multi-information function

$$I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) \tag{13.77}$$

New expressions of entropy

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

$$H_h(\mu_h) = -\sum_{x_h} \mu_h(x_h) \log \mu_h(x_h) \tag{13.76}$$

and the multi-information function

$$I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) \tag{13.77}$$

- 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.

New expressions of entropy

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

$$H_h(\mu_h) = -\sum_{x_h} \mu_h(x_h) \log \mu_h(x_h) \tag{13.76}$$

and the multi-information function

$$I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) \tag{13.77}$$

- 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

$$H_{\text{hyper}}(\mu) = -\sum_{h \in E} I_h(\mu_h) \tag{13.78}$$

multi-information decomposition

- Using Möbius, we can write

$$I_h(\mu_h) = \sum_{g \preceq h} \omega(g,h) \left\{ \sum_{x_h} \mu_h(x_h) \log \mu_g(x_g) \right\} \tag{13.79}$$

(13.80)

(13.81)

multi-information decomposition

- Using Möbius, we can write

$$I_h(\mu_h) = \sum_{g \preceq h} \omega(g,h) \left\{ \sum_{x_h} \mu_h(x_h) \log \mu_g(x_g) \right\} \tag{13.79}$$

$$= \sum_{f \preceq h} \sum_{e \succeq f} \omega(e,f) \left\{ \sum_{x_f} \mu_f(x_f) \log \mu_f(x_f) \right\} \tag{13.80}$$

$$\tag{13.81}$$

multi-information decomposition

- Using Möbius, we can write

$$I_h(\mu_h) = \sum_{g \preceq h} \omega(g,h) \left\{ \sum_{x_h} \mu_h(x_h) \log \mu_g(x_g) \right\} \tag{13.79}$$

$$= \sum_{f \preceq h} \sum_{e \succeq f} \omega(e,f) \left\{ \sum_{x_f} \mu_f(x_f) \log \mu_f(x_f) \right\} \tag{13.80}$$

$$= -\sum_{f \preceq h} c(f) H_f(\mu_f) \tag{13.81}$$

multi-information decomposition

- Using Möbius, we can write

$$I_h(\mu_h) = \sum_{g \preceq h} \omega(g,h) \left\{ \sum_{x_h} \mu_h(x_h) \log \mu_g(x_g) \right\} \tag{13.79}$$

$$= \sum_{f \preceq h} \sum_{e \succeq f} \omega(e,f) \left\{ \sum_{x_f} \mu_f(x_f) \log \mu_f(x_f) \right\} \tag{13.80}$$

$$= -\sum_{f \preceq h} c(f) H_f(\mu_f) \tag{13.81}$$

where

$$c(f) \triangleq \sum_{e \succeq f} \omega(f,e) \tag{13.82}$$

multi-information decomposition

- Using Möbius, we can write

$$I_h(\mu_h) = \sum_{g \preceq h} \omega(g,h) \left\{ \sum_{x_h} \mu_h(x_h) \log \mu_g(x_g) \right\} \tag{13.79}$$

$$= \sum_{f \preceq h} \sum_{e \succeq f} \omega(e,f) \left\{ \sum_{x_f} \mu_f(x_f) \log \mu_f(x_f) \right\} \tag{13.80}$$

$$= -\sum_{f \preceq h} c(f) H_f(\mu_f) \tag{13.81}$$

where

$$c(f) \triangleq \sum_{e \succeq f} \omega(f,e) \tag{13.82}$$

- This gives us a new expression for the hypertree entropy

$$H_{\text{hyper}}(\mu) = \sum_{h \in E} c(h) H_h(\mu_h) \tag{13.83}$$

Usable to get Kikuchi variational approximation

- Given arbitrary hypergraph now, we'll generalize the hypertree expressions above this arbitrary hypergraph, which will give us a variational expression that approximates cumulant.

Usable to get Kikuchi variational approximation

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

$$p_\theta(x) \propto \exp\left\{\sum_{h \in E} \sigma_h(x_h)\right\} \tag{13.84}$$

using same form of parameterization.

Usable to get Kikuchi variational approximation

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

$$p_\theta(x) \propto \exp\left\{\sum_{h \in E} \sigma_h(x_h)\right\} \tag{13.84}$$

  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

$$\sum_{x_h} \tau_h(x_h) = 1 \tag{13.85}$$

Usable to get Kikuchi variational approximation

- Sum to one constraint:

$$\sum_{x_h} \tau_h(x_h) = 1 \tag{13.86}$$

Usable to get Kikuchi variational approximation

- Sum to one constraint:

$$\sum_{x_h} \tau_h(x_h) = 1 \tag{13.86}$$

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

$$\sum_{x_{h \setminus g}} \tau_h(x_h) = \tau_g(x_g) \tag{13.87}$$

Usable to get Kikuchi variational approximation

- Sum to one constraint:

$$\sum_{x_h} \tau_h(x_h) = 1 \tag{13.86}$$

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

$$\sum_{x_{h \setminus g}} \tau_h(x_h) = \tau_g(x_g) \tag{13.87}$$

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

$$\mathbb{L}_t(G) = \{\tau \geq 0 | \text{ Equations (13.86) } \forall h, \text{ and (13.87) } \forall g \preceq h \text{ hold}\} \tag{13.88}$$

Kikuchi variational approximation

- Generalized entropy for the hypergraph:

$$H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \tag{13.89}$$

where H_g is hyperedge entropy and overcounting number defined by:

$$c(g) = \sum_{f \succeq g} \omega(g, f) \tag{13.90}$$

Kikuchi variational approximation

- Generalized entropy for the hypergraph:

$$H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \tag{13.89}$$

where H_g is hyperedge entropy and overcounting number defined by:

$$c(g) = \sum_{f \succeq g} \omega(g, f) \tag{13.90}$$

- This at last gets the Kikuchi variational approximation

$$A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{\langle \theta, \tau \rangle + H_{\text{app}}(\tau)\} \tag{13.91}$$

Kikuchi variational approximation

- Generalized entropy for the hypergraph:

$$H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \tag{13.89}$$

where H_g is hyperedge entropy and overcounting number defined by:

$$c(g) = \sum_{f \succeq g} \omega(g, f) \tag{13.90}$$

- This at last gets the Kikuchi variational approximation

$$A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{\langle \theta, \tau \rangle + H_{\text{app}}(\tau)\} \tag{13.91}$$

- For a graph, this is exactly $A_{\text{Bethe}}(\theta)$. If, on the other hand, the graph is a junction tree, then this is exact (although it might be expensive, exponential in the tree-width to compute H_{app}).

Kikuchi variational approximation

- Generalized entropy for the hypergraph:

$$H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \tag{13.89}$$

where H_g is hyperedge entropy and overcounting number defined by:

$$c(g) = \sum_{f \succeq g} \omega(g, f) \tag{13.90}$$

- This at last gets the Kikuchi variational approximation

$$A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{\langle \theta, \tau \rangle + H_{\text{app}}(\tau)\} \tag{13.91}$$

- For a graph, this is exactly $A_{\text{Bethe}}(\theta)$. If, on the other hand, the graph is a junction tree, then this is exact (although it might be expensive, exponential in the tree-width to compute H_{app}).
- Can define message passing algorithms on the hypertree, and show that if it converges, it is a fixed point of the Lagrangian associated with this

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`