Outstanding Reading

- Read chapters 3 and 4 in the Wainwright and Jordan book.
- Project status reports are due next Tuesday evening at 11:00pm via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/17463). No more than one page!
- HW2 is due next Wednesday morning at 10:00am via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/17463). Any questions please post directly to our discussion board (https://catalyst.uw.edu/gopost/board/bilmes/23863/) so that all can see the answers to your questions.
We need to find one makeup lecture this term.

- L1 (9/28): Introduction, Families, Semantics
- LX (9/30): No class
- L2 (10/5): Trees, exact inference
- L3 (10/7): More on trees and inference.
- L4 (10/12): To tree or not to tree.
- L5 (10/14): All models lead to trees
- L6 (10/19): Decomposable, JT
- L7 (10/21): Inference on JTs
- L8 (10/26): JT Inference, semi-rings,
- L9 (10/28): time-space tradeoff, conditioning, LBP
- L10 (11/2): LBP, exp. f. models
- L11 (11/4): exp. f. models, marg poly
- L12 (11/9):
- LXX (11/11): Veterans Day, no class
- L13 (11/16):
- L14 (11/18):
- L15 (11/23):
- LXX (11/25): Thanksgiving, no class
- L16 (11/30):
- L17 (12/2):
- L18 (12/7):
- L19: (12/9):

**Review**

- Family of models $p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$
- Results from the problem of maximum entropy: find the entropy maximizing distribution that satisfies a set of constraints.
- Minimal/Overcomplete representations of exponential families.
- Many distribution families are in the exponential family (Gaussian, Ising, General discrete non-negative distribution that factorizes w.r.t. a graph, etc.).
- Mean parameterization, alternative to $\theta$. Set of such means make a polytope.
Based on underlying set of parameters $\theta$, we have family of models

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{\alpha \in I} \theta_\alpha \phi_\alpha(x) \right\} = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$$  \hspace{1cm} (1)

To ensure normalized, we use log partition (cumulant) function

$$A(\theta) = \log \int_{\mathcal{D}_X} \exp(\langle \theta, \phi(x) \rangle) \nu(dx)$$  \hspace{1cm} (2)

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

$A(\theta)$ is convex function of $\theta$, so $\Omega$ is convex.

Exponential family for which $\Omega$ is open is called **regular**

**Exponential family models result from maximum entropy**

Goal is to find

$$p^* \in \arg\max_{p \in \mathcal{U}} H(p) \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \hat{\mu}_\alpha \forall \alpha \in I$$  \hspace{1cm} (3)

where $\forall \alpha \in I$

$$\mathbb{E}_p[\phi_\alpha(X)] = \int_{\mathcal{D}_X} \phi_\alpha(x)p(x)\nu(dx)$$  \hspace{1cm} (4)

This is solved by taking a distribution in the form of Eq. 1, by finding $\theta$ that solves

$$E_{p_\theta}[\phi_\alpha(X)] = \hat{\mu}_\alpha \text{ for all } \alpha \in I$$  \hspace{1cm} (5)
Mean Parameters, Convex Cores

- Consider quantities $\mu_\alpha$ associated with statistic $\phi_\alpha$ defined as:

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

(6)

- this defines a vector of “mean parameters” $(\mu_1, \mu_2, \ldots, \mu_d)$ with $d = |\mathcal{I}|$.

- Define all the possible such vectors

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

(7)

- 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 $\mu$ and $\mu'$

- $\mathcal{M}$ is like a “convex core” of all distributions expressed via $\phi$.

Mean Parameters and Polytopes

- When $X$ is discrete, we get a polytope since

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^b : \mu = \sum_x \phi(x)p(x) \text{ for some } p \in \mathcal{U} \right\}$$

(8)

$$= \text{conv} \{\phi(x), x \in \mathcal{D}_X\}$$

(9)

where $\text{conv} \{\cdot\}$ is the convex hull of the items in argument set.

- So we have a convex polytope
Mean Parameters and Polytopes

- Polytopes can be represented as a set of linear inequalities, i.e., there is a $|J| \times d$ matrix $A$ and $|J|$-element column vector $b$ with

$$M = \{ \mu \in \mathbb{R}^d : A \mu \geq b \} = \{ \mu \in \mathbb{R}^d : \langle a_j, \mu \rangle \geq b_j, \forall j \in J \} \quad (10)$$

with $A$ having rows $a_j$.

Example: Ising mean parameters. Given sufficient statistics

$$\phi(x) = \{ x_s, s \in V; x_s x_t, (s, t) \in E(G) \} \in \mathbb{R}^{\vert V \vert + \vert E \vert} \quad (11)$$

we get

$$\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \quad \forall v \in V \quad (12)$$
$$\mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \quad \forall (s, t) \in E(G) \quad (13)$$

- Mean parameters lie in a polytope that represent the probabilities of a node being 1 or a pair of adjacent nodes being 1,1 for each node and edge in the graph = conv $\{ \phi(x), x \in \{0, 1\}^m \}$.

- Gives complete marginal since $p_s(1) = 1 - p_s(0)$, $p_{s,t}(1, 0) = p_s(1) - p_{s,t}(1, 1)$, $p_{s,t}(0, 1) = p_t(1) - p_{s,t}(1, 1)$, etc.

- Recall: marginals are often the goal of inference. Coincidence?
Example: 2-variable Ising

"Ising model with two variables \((X_1, X_2) \in \{0, 1\}^2\). Three mean parameters \(\mu_1 = \mathbb{E}[X_1], \mu_2 = \mathbb{E}[X_2], \mu_{12} = \mathbb{E}[X_1X_2]\), must satisfy constraints \(0 \leq \mu_{12} \leq \mu_i\) for \(i = 1, 2\), and \(1 + \mu_{12} - \mu_1 - \mu_2 \geq 0\). These constraints carve out a polytope with four facets, contained within the unit hypercube \([0, 1]^3\)."

Mean Parameters and Overcomplete Representation

- We can use overcomplete representation and get a “marginal polytope”, a polytope that represents the marginal distributions at each potential function.
- Example: Ising overcomplete potential functions (generalization of Bernoulli example we saw before)

\[\forall v \in V(G), j \in \{0 \ldots r - 1\}, \text{ define } \phi_{v,j}(x_v) \triangleq 1(x_v = j)\]  \hspace{1cm} (14)

\[\forall (s, t) \in E(G), j, k \in \{0 \ldots r - 1\}, \text{ we define: } \phi_{st,jk}(x_s, x_t) \triangleq 1(x_s = j, x_t = k) = 1(x_s = j)1(x_t = k)\]  \hspace{1cm} (15)

So we now have \(|V|r + 2|E|r^2\) functions each with a corresponding parameter.
Mean Parameters and Marginal Polytopes

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

  \[
  \mu_{v,j} = \mathbb{E}_p[1(x_v = j)] = p(x_v = j) \quad (17)
  \mu_{st,jk} = \mathbb{E}_p[1(x_s = j, x_t = k)] = p(x_s = j, x_t = k) \quad (18)
  \]

- Such an $\mathcal{M}$ is called the *marginal polytope*. Any $\mu$ must live in the polytope that corresponds to node and edge true marginals.

- Any such $\mu$ corresponding to a real distribution and generated by Equations 18 are set to be *realizable*.

- 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.

Marginal Polytopes and Facet complexity

- Number of facets (faces) of a polytope is often (but not always) a good indication of its complexity.

- Corresponds to number of linear constraints in set of linear inequalities describing the polytope.

- “Facet complexity” of $\mathcal{M}$ depends on the graph structure.

- For 1-trees, marginal polytope characterized by local constraints only (pairs of variables on edges of the tree) and has linear growth with graph size.

- For $k$-trees, complexity grows exponentially.

- Key idea: use polyhedral approximations to produce model and inference approximations.
Learning is the dual of Inference

- We can view the inference problem as moving from the canonical parameters $\theta$ 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).

Ex: Estimate $\theta$ with $\hat{\theta}$ based on data $D = \{x^{(i)}_E\}^M_{i=1}$ of size $M$, likelihood function

$$\ell(\theta, D) = \frac{1}{M} \sum_{i=1}^M \log p_\theta(x^{(i)}) = \langle \theta, \hat{\mu} \rangle - A(\theta)$$ (19)

where empirical means given by

$$\hat{\mu} = \hat{\mathbb{E}}[\phi(X)] = \frac{1}{M} \sum_{i=1}^M \phi(x^{(i)})$$ (20)

By taking derivatives of the above, it is easy to see that solution is the point $\hat{\theta}$ such that (empirical matches expected means)

$$\mathbb{E}_{\hat{\theta}}[\phi(X)] = \hat{\mu}$$ (21)

this is the the backward mapping problem, going from $\mu$ to $\theta$.

- This is identical to the maximum entropy problem.
Learning is the dual of Inference

- I.e., solution to the maximum likelihood problem is one that satisfies the moment constraints and has the exponential model form.
- The exponential model form arises when we find the maximum entropy distribution over distributions satisfying the moment constraints.
- Thus, maximum entropy learning under a set of constraints (given by \( \mathbb{E}_\theta[\phi(X)] = \hat{\mu} \)) is the same as maximum likelihood learning of an exponential model form.
- If we do maximum entropy learning, where does the \( \exp(\cdot) \) function come from? From the entropy function. I.e., the exponential form is the distribution that has maximum entropy having those constraints.

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

\[ A(\theta) = \log \int_{\mathcal{D}_X} \langle \theta, \phi(x) \rangle \nu(dx) \quad (22) \]

- 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 \( \theta \) (strictly so if minimal representation).
  - It yields cumulants of the random vector \( \phi(X) \)
    \[ \frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] = \int \phi_\alpha(x) p_\theta(x) \nu(dx) = \mu_\alpha \quad (23) \]
    in general, derivative of log partition function is expected value of feature
  - Also, we get
    \[ \frac{\partial^2 A}{\partial \theta_{\alpha_1} \partial \theta_{\alpha_2}}(\theta) = \mathbb{E}_\theta[\phi_{\alpha_1}(X)\phi_{\alpha_2}(X)] - \mathbb{E}_\theta[\phi_{\alpha_1}(X)]\mathbb{E}_\theta[\phi_{\alpha_2}(X)] \quad (24) \]
  - Proof given in book.

So derivative of log partition function w.r.t. \( \theta \) is equal to our mean parameter \( \mu \) 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.

The Bethe approximation (as we'll 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).
Log partition function

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

Mappings - one-to-one

In fact, we have

**Theorem 1**

*The gradient map $\nabla A$ 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.
Moreover,

**Theorem 2**

*In a minimal exponential family, the gradient map $\nabla A$ 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 $E_{\theta}[\phi(X)] = \mu$.***

- Example: consider, for example, a Gaussian.
- Any mean parameter (set of means $E[X]$ and correlations $E[XX^T]$) can be realized by a Gaussian having those same mean parameters (moments).
- The Gaussian won’t nec. be the “true” distribution (in such case, the “true” distribution would not be an exponential family model with those moments).
- The theorem here is more general and applies for any set of sufficient statistics.

**Conjugate Duality**

- Consider maximum likelihood problem for exp. family
  
  $\theta^* \in \arg\max_{\theta} (\langle \theta, \hat{\mu} \rangle - A(\theta))$  \hspace{1cm} (25)

- Convex conjugate dual of $A(\theta)$ is defined as:
  
  $A^*(\mu) \triangleq \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta))$  \hspace{1cm} (26)

- So dual is optimal value of the ML problem, when $\mu \in \mathcal{M}$
- Key: when $\mu \in \mathcal{M}$, dual is negative entropy of exp. model $p_{\theta(\mu)}$ where $\theta(\mu)$ is the unique set of canonical parameters satisfying this matching condition
  
  $E_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu$  \hspace{1cm} (27)

- When $\mu \notin \mathcal{M}$, then $A^*(\mu) = +\infty$, so dual optimization need consider points only in $\mathcal{M}$. 

Theorem 3 (Relationship between $A$ and $A^*$)

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

$$A^*(\mu) = \sup_{\theta \in \Omega} \left( \langle \theta, \mu \rangle - A(\theta) \right) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \in \bar{\mathcal{M}} \end{cases} \quad (28)$$

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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \quad (29)$$

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

$$\mu = \int_{D_X} \phi(x)p_{\theta}(x)\nu(dx) = \mathbb{E}_{\theta}[\phi(X)] = \nabla A(\theta) \quad (30)$$

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

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \mu \quad (31)$$

$A(\theta)$ in Equation 29 is the “inference” problem (dual of the dual) for a given $\theta$.

Whenever $\mu \notin \mathcal{M}$, then $A^*(\mu)$ returns $\infty$ which can’t be the resulting sup, so Equation 29 need only consider $\mathcal{M}$. 
Conjugate Duality

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \] (29)

- 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. 😞

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. 😊

Surprisingly, this is strongly related to belief propagation (i.e., the sum-product commutative semiring). 😊😊
We’ll see that loopy belief-propagation (sum-product alg.) has much to

do with an approximation to the aforementioned variational problems.

Recall: we’re 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

\[
    p_{0}(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{v \in V(G)} \theta_{v}(x_{v}) + \sum_{(s,t) \in E(G)} \theta_{st}(x_{s}, x_{t}) \right\}
\]

with simple new shorthand notation functions \( \theta_{v} \) and \( \theta_{st} \).

\[
    \theta_{v}(x_{v}) \triangleq \sum_{i} \theta_{v,i} 1(x_{v} = i) \quad \text{and} \quad \theta_{st}(x_{s}, x_{t}) \triangleq \sum_{ij} \theta_{st,ij} 1(x_{s} = i, x_{t} = j)
\]

And \( M(G) \) corresponds to the set of all singleton and pairwise

marginals that can be jointly realized by some underlying probability

distribution \( p \in \mathcal{F}(G, R^{(f)}) \) that contains only pairwise interactions.

Note, \( M(G) \) is respect to a graph \( G \).

\( M \) 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 $\mathcal{M}$ consists of a set that contains $\mathcal{M}$, and if it is formed from a subset of the linear inequalities (subset of the rows of matrix module $(A, b)$), then it is a polyhedral outer bound. Let's call this $\mathbb{L}$.
- 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 non-negative and satisfies normalization
  $$\sum_{x_v} \tau_v(x_v) = 1 \tag{36}$$
  and pair-node marginal consistency constraints
  $$\sum_{x'_t} \tau_{s,t}(x'_s, x'_t) = \tau_s(x_s) \tag{37a}$$
  $$\sum_{x'_s} \tau_{s,t}(x'_s, x_t) = \tau_t(x_t) \tag{37b}$$

Define $\mathbb{L}(G)$ to be the (locally consistent) polytope that obeys the constraints in Equations 36 and 37.
- Recall: local consistency was the necessary conditions for potentials being marginals that, it turned out, for junction tree that also guaranteed global consistency.
- Clearly $\mathcal{M} \subseteq \mathbb{L}(G)$ since any member of $\mathcal{M}$ (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 $\mathcal{M}(T) = \mathbb{L}(T)$
- When $G$ has cycles, however, $\mathcal{M}(G) \subset \mathbb{L}(G)$ 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.
Pseudo-marginals

\[ \tau_v(x_v) = [0.5, 0.5], \quad \text{and} \quad \tau_{s,t}(x_s, x_t) = \begin{bmatrix} \beta_{st} & \cdot 5 - \beta_{st} \\ \beta_{st} & \cdot 5 - \beta_{st} \end{bmatrix} \] (38)

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

Bethe Entropy Approximation

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \] (29)

- So inference corresponds to Equation 29, 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 \( \mathcal{L} \) being those distributions that factor w.r.t. a tree.
- When \( p \in \mathcal{F}(G, R^f) \) and \( G \) is a tree \( T \), then we can write \( p \) as:

\[ p(x_1, \ldots, x_N) = \prod_{v \in \mathcal{V}(T)} p_v(x_v) \prod_{(i,j) \in \mathcal{E}(T)} \frac{p_{ij}(x_i, x_j)}{p_i(x_i)p_j(x_j)} \] (39)
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, any $\mu \in \mathbb{L}(T)$ is such that $\mu \in \mathbb{M}(T)$, thus

$$p_u(x) = \prod_{s \in V(T)} \mu_s(x_s) \prod_{(s,t) \in E(T)} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s) \mu_t(x_t)} \quad (40)$$

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

$$-A^*(\mu) = H(p_\mu) = \sum_{v \in V(T)} H(X_v) - \sum_{(s,t) \in E(T)} I(X_s; X_t) \quad (41)$$

$$= \sum_{v \in V(T)} H_v(\mu_v) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \quad (42)$$

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

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.,

$$-A^*(\tau) \approx H_{\text{Bethe}}(\tau) \triangleq \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \quad (43)$$

- Recall, MI equation is not hard to compute $O(r^2)$.

$$I_{st}(\tau_{st}) = I_{st}(\tau_{st}(x_s, x_t)) \quad (44)$$

$$= \sum_{x_s,x_t} \tau_{st}(x_s, x_t) \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s) \tau_t(x_t)} \quad (45)$$
Original variational representation of log partition function

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \quad (46) \]

Approximate variational representation of log partition function

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

\[ = \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \quad (48) \]

- Exact when \( G = T \) but we do this for any \( G \), still computable
- 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.

Lagrangian constraints for summing to unity at nodes

\[ C_{vv}(\tau) = 1 - \sum_{x_v} \tau_v(x_v) \quad (49) \]

Lagrangian constraints for local consistency

\[ C_{ts}(x_s; \tau) = \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t) \quad (50) \]

Yields following Lagrangian

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

\[ + \sum_{(s,t) \in E(G)} \left[ \sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s; \tau) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t; \tau) \right] \quad (52) \]
Fixed points: Variational Problem and LBP

Theorem 4

LBP updates are Lagrangian method for attempting to solve Bethe variational problem:

(a) For any $G$, any LBP fixed point specifies a pair $(\tau^*, \lambda^*)$ s.t.

$$\nabla_\tau \mathcal{L} (\tau^*, \lambda^*; \theta) = 0 \text{ and } \nabla_\lambda \mathcal{L} (\tau^*, \lambda^*; \theta) = 0 \quad (53)$$

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

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.

The resulting Lagrange multipliers $\lambda^*_{st}$ end up being exactly the messages that we have defined. I.e., we get

$$\lambda^*_{st}(x_t) = \mu_{s \rightarrow t}(x_t) \quad (54)$$

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).

This does not mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it (mean-field methods can provide a lower bound on $A(\theta)$).

For certain potential functions, we’ll get $A_{\text{Bethe}}(\theta) \leq A(\theta)$ as we’ll see later.
Prof. Jeff Bilmes
EE12A/Fall 2011/Graphical Models – Lecture 12 - Nov 9th, 2011

Sources for Today’s Lecture

- Most of this material comes from the Wainwright and Jordan book.