Announcements

- 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).
Mean Parameters, Convex Cores

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

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

(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 $d = |\mathcal{I}|$,

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

(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 $\mu$ and $\mu'$

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

  \[
  \begin{align*}
  \mu_{v,j} &= \mathbb{E}_p[1(x_v = j)] = p(x_v = j) \quad (13.20) \\
  \mu_{st,jk} &= \mathbb{E}_p[1(x_s = j, x_t = k)] = p(x_s = j, x_t = k) \quad (13.21)
  \end{align*}
  \]

- Such an \( \mathcal{M} \) is called the marginal polytope for discrete graphical models. Any \( \mu \) 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 \( \theta \) 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).
Maximum entropy estimation

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

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

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

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

- \( \mathbb{E}_p[\phi_\alpha(X)] \) 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 \( \theta \) that solves

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

Learning is the dual of Inference

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

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

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

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:

\[ E_{\hat{\theta}}[\phi(X)] = \hat{\mu} \quad (13.23) \]

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

- 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, \( D = \{\bar{x}^{(i)}\}_{i=1}^M \), defines an empirical distribution

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

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

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

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

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

- 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

\[ A(\theta) = \log \int_{D_X} \exp \langle \theta, \phi(x) \rangle \nu(dx) \quad (13.20) \]

- 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 \( \theta \) (strictly so if minimal representation).
- It yields cumulants of the random vector \( \phi(X) \)

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

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

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

(13.22)


Logistic

Review

Log partition function: properties

- 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. Why do we want bounds? We shall soon see.
- 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

\[ p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \]  
(13.1)

with

\[ A(\theta) = \log \int_{D_X} \langle \theta, \phi(x) \rangle \nu(dx) \]  
(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
  \[ \mathcal{M} = \{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\} \].
- 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 \( \mu \) 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 that are realizable by some dist. are also realizable by member of exp. family.
Expanding on one of the previous properties, . . .

**Theorem 13.3.1**

*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 of $A(\theta)$

**Mappings - onto**

**Theorem 13.3.2**

*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 $E_{\theta}[\phi(X)] = \mu$.*

- Ex: 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 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
  \[ \theta^* \in \arg\max_{\theta} (\langle \theta, \hat{\mu} \rangle - A(\theta)) \] (13.3)

- Compare this to convex conjugate dual (also sometimes Fenchel-Legendre dual or transform) of \( A(\theta) \) is defined as:
  \[ A^*(\mu) \overset{\Delta}{=} \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) \] (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
\[ \mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu \] (13.5)

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

---

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

(b) Partition function has variational representation (dual of dual)
\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \] (13.7)

(c) For \( \theta \in \Omega \), sup occurs at \( \mu \in \mathcal{M}^\circ \) of moment matching conditions
\[ \mu = \int_{D_X} \phi(x)p_{\theta}(x)\nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta) \] (13.8)
Conjugate Duality, and Inference

- 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

$$E_{\theta(\mu)}[\phi(X)] = \mu \quad (13.9)$$

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

- Whenever $\mu \notin \mathcal{M}$, then $A^*(\mu)$ returns $\infty$ which can’t be the resulting sup in Equation 13.7, so Equation 13.7 need only consider $\mathcal{M}$.

Conjugate Duality, Good and Bad News

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

- 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, Avenues to Approximation

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \]  
(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 semidefinite 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.
- 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 \( \theta_v \) and \( \theta_{st} \).

\[
\theta_v(x_v) \triangleq \sum_i \theta_{v,i} \mathbf{1}(x_v = i) \quad \text{and} \quad \theta_{st}(x_s, x_t) \triangleq \sum_{i,j} \theta_{st,ij} \mathbf{1}(x_s = i, x_t = j)
\]

(13.10)  
(13.11)
Marginal notation, and graph
Marginal polytope

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

\[
\mu_u(x_v) \triangleq \sum_{i \in D_{X_v}} \mu_{v,i} 1(x_v = i), \text{ for } u \in V(G)
\]

\[
\mu_{st}(x_s, x_t) \triangleq \sum_{(j,k) \in D_{X_{\{s,t\}}}} \mu_{st,jk} 1(x_s = j, x_t = k), \text{ for } (s, t) \in E(G)
\]

- And \(\mathbb{M}(G)\) corresponds to the set of all singleton and pairwise marginals that can be jointly realized by some underlying probability distribution \(p \in \mathcal{F}(G, \mathcal{M}^{(f)})\) 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 (tree outer bound) polytope

- An “outer bound” of \(\mathbb{M}\) consists of a set that contains \(\mathbb{M}\). If formed from a \text{\textbf{subset}} of the linear inequalities (subset of the rows of matrix module \((A, b)\)), then it is a polyhedral outer bound.

- A simple way to form outer bound: require only local consistency, i.e., consider set \(\{\tau_v, v \in V(G)\} \cup \{\tau_{s,t}, (s, t) \in E(G)\}\) that is, always non-negative, and that satisfies normalization

\[
\sum_{x_v} \tau_v(x_v) = 1
\]

and pair-node marginal consistency constraints

\[
\sum_{x_t'} \tau_{s,t}(x_s, x_t') = \tau_s(x_s)
\]  

\[
\sum_{x_s'} \tau_{s,t}(x_s', x_t) = \tau_t(x_t)
\]
Local consistency (tree outer bound) polytope: properties

- 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.

Pseudo-marginals

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

- 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).
Bethe Entropy Approximation

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \]  
(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, let’s use an approximate entropy based on \( \mathbb{L} \) being those distributions that factor w.r.t. a tree.
- When \( p \in \mathcal{F}(G, \mathbb{M}(\mathcal{F})) \) and \( G \) is a tree \( T \), then we can write \( p \) as:

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

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

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

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

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

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

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

\[
= \sum_{v \in V(T)} H_v(\mu_v) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st})
\]
(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.
- 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.,

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

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

- 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

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

Approximate variational representation of log partition function

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

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

- 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 $\mathcal{L}$), 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
  \[ C_{vv}(\tau) = 1 - \sum_{x_v} \tau_v(x_v) \]  
  (13.29)

- Lagrangian constraints for local consistency
  \[ C_{ts}(x_s; \tau) = \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t) \]  
  (13.30)

- Yields following Lagrangian
  \[ L(\tau, \lambda; \theta) = \langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau) + \sum_{v \in V} \lambda_{vv} C_{vv}(\tau) \]  
  \[ + \sum_{(s,t) \in E(G)} \left[ \sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s; \tau) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t; \tau) \right] \]  
  (13.31)

- Fixed points: Variational Problem and LBP
  \[ \text{Theorem 13.5.1} \]
  
  **LBP updates are Lagrangian method for attempting to solve Bethe variational problem:**
  
  (a) For any \( G \), any LBP fixed point specifies a pair \((\tau^*, \lambda^*)\) s.t.
  \[ \nabla_\tau L(\tau^*, \lambda^*; \theta) = 0 \text{ and } \nabla_\lambda L(\tau^*, \lambda^*; \theta) = 0 \]  
  (13.33)

  (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.

  - 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 $\lambda_{st}$ end up being exactly the messages that we have defined. I.e., we get

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

(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? Recall Max. Likelihood

$$
\theta^* \in \arg \max_{\theta} (\langle \theta, \hat{\mu} \rangle - A(\theta))
$$

(13.3)

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

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

(13.4)

- Recall again the expression for the partition function

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

(13.7)

and some approximation to $A(\theta)$, say $A_{\text{approx}}(\theta)$.
- Due to sup in Eq. (13.3), 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.
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
\[
p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))
\]  

(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
\[
p(x_A | x_B) = \frac{p(x_{A \cup B})}{p(x_B)} = \frac{\sum_{x_{V \setminus (A \cup B)}} p(x)}{\sum_{x_{V \setminus B}} p(x)}
\]  

(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.

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

Example of inaccuracy (example 4.2 from book), consider a 4-clique
\[
\mu_s(x_s) = [0.5 \ 0.5] \quad \text{for} \ s = 1, 2, 3, 4
\]  

(13.37a)

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

(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
\[
H_{\text{Bethe}}(\mu) = 4 \log 2 - 6 \log 2 = -2 \log 2 < 0
\]  

(13.38)

which obviously can’t be a true entropy since we must have \( H > 0 \) for discrete distributions.

True \(-A^*(\mu) = \log 2 > 0\).
Sources for Today’s Lecture