

EE512 – Graphical Models – Fall 2009

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
Fall Quarter, 2009

<http://ssli.ee.washington.edu/~bilmes/ee512fa09>

Lecture 18 - Dec 3rd, 2009

Last updated \$Id: lec18.tex,v 1.5 2009/12/04 22:19:24 bilmes Exp \$

Class Road Map

- L1 (10/1): intro, cond. indep., ex. GMs
- L2 (10/6): GMs and MRFs
- L3 (10/8): MRFs, mobius, FGs
- L4 (10/13): Sem BNs
- L5 (10/15): Sem BNs 2.
- L6 (10/20): Evidence
- L7 (10/22): Inf. Trees 1
- L8 (10/27): Inf. Trees 2
- L9 (10/29): Inf. Trees 3
- L10 (11/03): Inf. Trees 4
- L11 (11/05): Inf. Trees 5
- L12 (11/10): Inf. Trees 6
- L13 (11/12): Inf. Trees 7
- L14 (11/17): intr grphs, JT inf
- L15 (11/19): trees, rings, Conditioning
- L16 (11/24): Stru Lear
- L17 (12/01): Approx Inf. I
- L18 (12/03): *Approx Inf. II
- L19 (12/08): (video) Approx Inf. III
- L20 (12/10): (video) Approx Inf. IV
- L21 (12/18): Friday, 2:30-4:20pm, final presentations

Readings

- Read Wainwright/Jordan book chaps 3/4/5
<http://dx.doi.org/10.1561/22000000001>
- Read “tree_inference.pdf”
- Read “evidence.pdf”
- Read “dgms.pdf”
- Read “ugms.pdf” .
- Read “intro.pdf” .
- Optionally (but encouraged): read chapters 1 through 10 in Jordan text.
- Read relevant chapters in Koller/Friedman text.

Homework/Project

- Updated project proposals this evening (midnight).
- What to turn in: **One page** proposals that describe what you plan to do for a project, again email me PDF files (only PDF accepted). I'll comment on the PDF and mail it back to you.
- only **one page** please.
- Final project will be a 4-page conference-style research paper due on Thursday Dec. 17th.
- Project should ideally be on some aspect of the material we have learnt.
- It could also be an implementation (i.e., a fast implementation of the JT algorithm, or loopy BP, and some reporting and experiences that you have had in doing this).
- While it ideally should be research-oriented, it is not acceptable to propose whatever machine learning task you are currently working on (e.g., “An application of SVMs to protein folding” would not be acceptable).

Belief Propagation (BP)

- Message passing: tree + MPP + all messages \Rightarrow convergence.
- Convergence = *reparameterization* of edge/node functions as true marginals.
- k -way interaction transformable into pairwise interactions – similar to factor graph representation, same complexity.
- Also possible to define BP directly on factor graphs — not that different.
- Notion of *parametrization* and marginal parametrization as a goal, not always possible (4-cycle).
- Parallel message passing, after D steps converged for tree.
- State representation and matrix multiply

Belief Propagation (BP)

- If graph has cycles, un-damped oscillation is possible.
- Possible to damp each message (mix with previous ones) to stamp out oscillation.

State representation

- Consider the set of messages $\{\mu_{i \rightarrow j}(x_j)\}_{i,j}$ as a large state vector μ^t with $2|E(G)|r$ scalar elements.
- Each sent message moves the state vector from μ^t at time t to μ^{t+1} at next time step.
- A parallel message moves the state vector as well.
- Convergence means that any set or subset of messages sent in parallel is such that $\mu^{t+1} = \mu^t$.

Messages as matrix multiply

$$\mu_{i \rightarrow j}(x_j) \propto \sum_{x_i} \psi_{i,j}(x_i, x_j) \psi_i(x_i) \prod_{k \in \delta(i) \setminus \{j\}} \mu_{k \rightarrow i}(x_i) \quad (1)$$

$$= \sum_{x_i} \psi'_{i,j}(x_i, x_j) \mu_{j \rightarrow i}(x_i) \quad (2)$$

$$= (\psi'_{i,j})^T \mu_{j \rightarrow i} \quad (3)$$

- Here, $\psi'_{i,j}$ is a matrix and $\mu_{j \rightarrow i}$ is a column vector.
- Going from state μ^t to μ^{t+1} is like matrix-vector multiply — group messages from μ^t together into one vector representing $\mu_{j \rightarrow i}$ for each $(i, j) \in E$, do the matrix-vector update, and store result in new state vector μ^{t+1} .
- if $|\delta(i) \setminus \{j\}| = 1$ then no grouping necessary.
- If G is tree, μ^t will converged after D steps.

Belief Propagation, Single Cycle

- Consider a graph with a single cycle C_ℓ .
- It could be a cycle with trees hanging off of each node. We send messages from the leaves of those dangling trees to the cycle (root) nodes, leaving only a cycle remaining.
- Consider what happens to $\mu_{i \rightarrow j}^t$ as t increases. w.l.o.g. consider $\mu_{\ell \rightarrow 1}^t$
- Let the cycle be nodes $(1, 2, 3, \dots, \ell, 1)$

$$\mu_{\ell \rightarrow 1}^{t+1} = \left(\prod_{i=1}^{\ell-1} (\psi_{i,i+1})^T \right) \mu_{\ell \rightarrow 1}^t \quad (4)$$

$$= M \mu_{\ell \rightarrow 1}^t \quad (5)$$

- Will this converge to anything?

Belief Propagation, Single Cycle

Theorem (Power method lemma)

Let A be a matrix with eigenvalues $\lambda_1, \dots, \lambda_n$ (sorted in decreasing order) and corresponding eigenvectors x_1, x_2, \dots, x_n . If $|\lambda_1| > |\lambda_2|$ (strict), then the update $x^{t+1} = \alpha Ax^t$ converges to a multiple of x_1 starting from any initial vector $x^0 = \sum_i \beta_i x_i$ provided that $\beta_1 \neq 0$. The convergence rate factor is given by $|\lambda_2/\lambda_1|$.

Belief Propagation, Single Cycle

From this, we the following theorem follows almost immediately.

Theorem

1. $\mu_{\ell \rightarrow 1}$ converges to the principle eigenvector of M .
2. $\mu_{2 \rightarrow 1}$ converges to the principle eigenvector of M^T .
3. The convergence rate is determined by the ratio of the larges and second largest eigenvalue of M .
4. The diagonal elements of M correspond to correct marginal $p(x_1)$
5. The steady state “pseudo-marginal” $b(x_1)$ is related to the true marginal by $b(x_1) = \beta p(x_1) + (1 - \beta)q(x_1)$ where β is the ratio of the largest eigenvalue of M to the sum of all eigenvalues, and $q(x_1)$ depends on the eigenvectors of M .

Proof.

See Weiss2000. □

Belief Propagation, arbitrary graph

- This works for a graph with a single cycle, or a graph that contains a single cycle
- It still does not tell us that we end up with correct marginals, rather we get “pseudo-marginals”, which are locally normalized, but might not be the correct marginals.
- Moreover, they might not be the correct marginals for any probability distribution.
- Also, we'd like a characterization of LBP's convergence (if it happens) for more general graphs, with an arbitrary number of loops.

exponential family models

- $\psi = (\psi_C, C \in \mathcal{C})$ is a collection of functions known as potential functions or sufficient statistics, where \mathcal{C} is the index set.
- Each ψ_C is a function of x_C , $\psi_C(x_C)$
- In the past, \mathcal{C} are the cliques of some graph (still the case here), but not nec. maxcliques. Not nec. dealing with triangulated models.
- θ is a vector of canonical parameters (same length, $|\mathcal{C}|$).
 $\theta \in \Omega \subseteq \mathbb{R}^d$ where $d = |\mathcal{C}|$.
- The family is defined as

$$p_\theta(x) = \exp(\langle \theta, \psi(x) \rangle - A(\theta)) \quad (6)$$

Note that we're using ψ here in the exponent, before we were using it out of the exponent.

- $p \in \mathcal{F}(G, R^{(f)})$ by Hammersley-Clifford theorem, where G is formed by using \mathcal{C} as an edge clique cover.

exponential family models

- exponential models are in our sense sufficient to deal with the computational aspects graphical models.
- We can have $p \in \mathcal{F}((V, E), R^{(f)})$ implies $p \in \mathcal{F}((V, E + E_1), R^{(f)})$ but in some sense, for any G , we want to deal with the models for which G is tight (we don't want to use overly complex graph to deal with family that is simpler)
- Exponential models can represent any factorization, given any factorization, we can do $\exp(\log \psi)$ to get potentials.
- We can often make them log-linear models as well with the right potential functions which won't increase tree-width of the graph.
- Moreover, exponential family models are incredibly flexible and have a number of desirable properties (e.g., aspects of the log partition function which we will see)

exponential family models

- Underlying base measure ν , p is absolutely continuous w.r.t. ν
- Based on underlying set of parameters θ , we have family of models

$$p_{\theta}(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{C \in \mathcal{C}} \theta_C \psi_C(x) \right\} = \exp(\langle \theta, \psi(x) \rangle - A(\theta)) \quad (7)$$

- this family can arise for a number of reasons, one of them is that it is the distribution having maximum entropy but that satisfies certain (moment) constraints.
- Given data $\mathbf{D} = \{\bar{x}_E^{(i)}\}_{i=1}^M$ of size M , form the expected statistics (requirements) of a model

$$\hat{m}_C = \frac{1}{M} \sum_{i=1}^M \psi_C(\bar{x}^{(i)}) \quad (8)$$

Exponential family models

- Goal is to find

$$p^* \in \operatorname{argmax}_{p \in \mathcal{U}} H(p) \text{ s.t. } \mathbb{E}_p[\psi_C(X)] = \hat{m}_C \quad \forall C \in \mathcal{C} \quad (9)$$

where $\forall C \in \mathcal{C}$

$$\mathbb{E}_p[\psi_C(X)] = \int_{D_X} \psi_C(x) p(x) \nu(dx) \quad (10)$$

- This is solved by a distribution in the form of Eq. 7, by finding θ to solve

$$E_{p_\theta}[\psi_C(X)] = \hat{m}_C \text{ for all } C \in \mathcal{C} \quad (11)$$

Exponential family models

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

$$A(\theta) = \log \int_{\mathcal{D}_x} \exp(\langle \theta, \psi(x) \rangle) \nu(dx) \quad (12)$$

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

- $A(\theta)$ is convex function of θ
- Minimal representation - not exist a nonzero vector $\gamma \in \mathbb{R}^d$ for which $\langle \gamma, \psi(x) \rangle$ is constant – essential idea is that for a set of sufficient stats \mathcal{C} , there shouldn't be any lower-dim vector $|\mathcal{C}'| < |\mathcal{C}|$ that is also sufficient (a min suf stat is a function of all other suf stats). We can't reduce the dimensionality d without reducing the family.
- Overcomplete representation d higher than need be - exists affine hyperplane of different parameters that induce exactly same distribution — useful in understanding BP algorithm.

Exponential family models

- Minimal representation of Bernoulli distribution is

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \quad (13)$$

- overcomplete rep of Bernoulli dist.

$$p(x|\theta_0, \theta_1) = \exp(\theta_0(1-x) + \theta_1 x - A(\gamma)) \quad (14)$$

- overcomplete since there is a linear combination of feature functions that are constant, i.e., any $(1-x) + x = 1$. Parameters of form $\theta_1 - \theta_0 = \gamma$ give same distribution.

Famous Example - Ising Model

- Famous example is the Ising model in statistical physics. We have a grid network with pairwise interactions, each variable is binary, and parameters associated with pairs being both on. Model becomes

$$p_{\theta}(x) = \exp \left\{ \sum_{v \in V} \theta_v x_v + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \right\}, \quad (15)$$

with

$$A(\theta) = \log \sum_{x \in \{0,1\}^m} \exp \left\{ \sum_{v \in V} \theta_v x_v + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \right\} \quad (16)$$

- Note that this is in minimal form. Any change to parameters will result in different distribution

Mean Parameters, Convex Cores

- Consider quantities \mathbf{m}_C associated with statistic ψ_C defined as:

$$\mathbf{m}_C = \mathbb{E}_p[\psi_C(X)] = \int_{D_X} \psi_C(x) p(x) \nu(dx) \quad (17)$$

- this defines a vector of “mean parameters” $(\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_d)$ with $d = |\mathcal{C}|$.
- Define all the possible such vectors

$$\mathcal{M} \triangleq \left\{ \mathbf{m} \in \mathbb{R}^d : \exists p \text{ s.t. } \mathbf{m}_C = \mathbb{E}_p[\psi_C(X)] \forall C \in \mathcal{C} \right\} \quad (18)$$

- didn't say p was 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 \mathbf{m} and \mathbf{m}'
- \mathcal{M} is like a “convex core” of all distributions expressed via ψ .

Mean Parameters and Polytopes

- When X is discrete, we get a polytope since

$$\mathcal{M} = \left\{ \mathbf{m} \in \mathbb{R}^b : \mu = \sum_x \psi(x)p(x) \text{ for some } p \in \mathcal{U} \right\} \quad (19)$$

$$= \text{convexhull} \{ \psi(x), x \in D_X \} \quad (20)$$

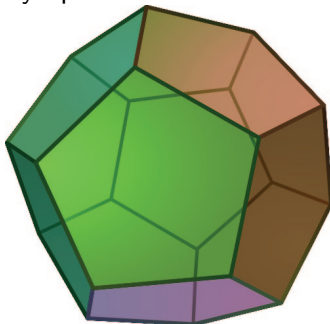
Mean Parameters and Polytopes

- When X is discrete, we get a polytope since

$$\mathcal{M} = \left\{ \mathbf{m} \in \mathbb{R}^b : \mu = \sum_x \psi(x)p(x) \text{ for some } p \in \mathcal{U} \right\} \quad (19)$$

$$= \text{convexhull} \{ \psi(x), x \in D_X \} \quad (20)$$

- So we have a polytope

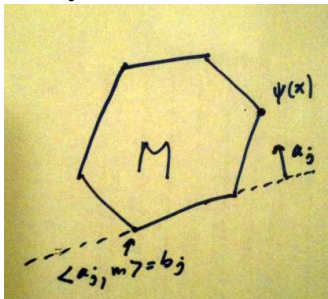


Mean Parameters and Polytopes

- 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

$$M = \left\{ \mathbf{m} \in \mathbb{R}^d : A\mathbf{m} \geq \mathbf{b} \right\} = \left\{ \mathbf{m} \in \mathbb{R}^d : \langle \mathbf{a}_j, \mathbf{m} \rangle \geq b_j, \forall j \in J \right\} \quad (21)$$

with A having rows \mathbf{a}_j .



Mean Parameters and Polytopes

- Example: Ising mean parameters

$$\mathbf{m}_v = \mathbb{E}_p[X_s] = p(X_s = 1) \quad \forall v \in V \quad (22)$$

$$\mathbf{m}_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \quad \forall (s, t) \in E(G) \quad (23)$$

- In this case, the 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.

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)

$$\forall v \in V(G), j \in \{1 \dots r\}, \text{ define } \psi_{v,j}(x_v) = \mathbf{1}(x_v = j) \quad (24)$$

$$\forall (s, t) \in E(G), j, k \in \{1 \dots r\}, \quad (25)$$

$$\text{define } \psi_{st,jk}(x_s, x_t) = \mathbf{1}(x_s = j, x_t = k) \quad (26)$$

- 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.,
 $\mathbf{m}_v(j) = p(x_v = j)$ and $\mathbf{m}_{st}(j, k) = p(x_s = j, x_t = k)$ since

$$\mathbf{m}_{v,j} = \mathbb{E}_p[\mathbf{1}(x_v = j)] = p(x_v = j) \quad (27)$$

$$\mathbf{m}_{st,jk} = \mathbb{E}_p[\mathbf{1}(x_s = j, x_t = k)] = p(x_s = j, x_t = k) \quad (28)$$

- Such an \mathcal{M} is called the *marginal polytope*. Any \mathbf{m} must live in the polytope that corresponds to node and edge true marginals!!
- 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 $\mathfrak{m} \in \mathcal{M}$.
- We can view the (maximum likelihood) learning problem as moving from a point in the polytope (empirical distribution) to the canonical parameters.
- 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).

Learning is the dual of Inference

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

$$\ell(\theta, \mathbf{D}) = \frac{1}{M} \sum_{i=1}^M \log p_{\theta}(\bar{x}^{(i)}) = \langle \theta, \hat{\mathbf{m}} \rangle - A(\theta) \quad (29)$$

where

$$\hat{\mathbf{m}} = \hat{\mathbb{E}}[\psi(\mathbf{X})] = \frac{1}{M} \sum_{i=1}^M \psi(\bar{x}^{(i)}) \quad (30)$$

- By taking derivatives of the above, it is easy to see that solution is point θ such that

$$\mathbb{E}_{\theta}[\psi(\mathbf{X})] = \hat{\mathbf{m}} \quad (31)$$

this is the the *backward mapping problem*, going from \mathbf{m} to θ .

Learning is the dual of Inference

- In other words, the solution to the maximum likelihood problem is one that satisfies the moment constraints and has the exponential model form. The exponential model form is exactly the equation that arises when we find the maximum entropy distribution over those distributions satisfying the moment constraints.
- This shows that maximum entropy learning under a set of constraints (given by $\mathbb{E}_{\theta}[\psi(X)] = \hat{m}$) is the same as maximum likelihood learning of an exponential model form.
- If we do maximum entropy learning, where does the exp pop up? From the entropy function. I.e., the exponential form is the distribution that has maximum entropy having those constraints.

Log partition function

$$A(\theta) = \log \int_{\mathcal{D}_X} \langle \theta, \psi(x) \rangle \nu(dx) \quad (32)$$

- 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
- It yields cumulants of the random vector $\psi(X)$

$$\frac{\partial A}{\partial \theta_C}(\theta) = \mathbb{E}_\theta[\psi_C(X)] = \int_{\mathcal{D}_X} \psi_C(X) p_\theta(x) \nu(dx) \quad (33)$$

$$\frac{\partial^2 A}{\partial \theta_{C_1} \partial \theta_{C_2}}(\theta) = \mathbb{E}_\theta[\psi_{C_1}(X) \psi_{C_2}(X)] - \mathbb{E}_\theta[\psi_{C_1}(X)] \mathbb{E}_\theta[\psi_{C_2}(X)] \quad (34)$$

Log partition function

- So derivative of log partition function w.r.t. θ is equal to our mean parameter m in the discrete case.
- Given $A(\theta)$, we can recover the marginals for each potential function ψ_C , $C \in \mathcal{C}$ (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

- $\nabla A : \Omega \rightarrow \mathcal{M}'$ where $\mathcal{M}' \subseteq \mathcal{M}$,
- for minimal exponential family models, this mapping is one-to-one, that is $\mathcal{M}' = \mathcal{M}$ and there is a unique pairing between \mathfrak{m} and θ .
- For non-minimal exponential families, we have more than one θ (not surprising since multiple θ 's can yield the same distribution).
- For non-exponential families, other distributions can yield \mathfrak{m} , but the exponential family one is the one that has maximum entropy (consider a Gaussian, which is the distribution having a given mean and covariance having maximum entropy amongst all other distributions).
- Key point: all mean parameters are realizable by member of exp. family.
- There exists some technical points about interior of \mathcal{M} we won't discuss here.

Conjugate Duality

- Maximum likelihood problem for exp. family

$$\theta^* \in \operatorname{argmax}_{\theta} (\langle \theta, \hat{\mathbf{m}} \rangle - A(\theta)) \quad (35)$$

- Conjugate dual

$$A^*(\mathbf{m}) \triangleq \sup_{\theta \in \Omega} (\langle \theta, \mathbf{m} \rangle - A(\theta)) \quad (36)$$

- dual is like ML when $\mathbf{m} \in \mathcal{M}$
- Key: when $\mathbf{m} \in \mathcal{M}$, dual is negative entropy of exp. model $p_{\theta(\mathbf{m})}$ where $\theta(\mathbf{m})$ is the unique set of canonical parameters satisfying this *matching condition*

$$\mathbb{E}_{\theta(\mathbf{m})}[\psi(X)] = \nabla A(\theta(\mathbf{m})) = \mathbf{m} \quad (37)$$

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

Conjugate Duality

Theorem

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

$$A^*(\mathbf{m}) = \begin{cases} -H(p_{\theta(\mathbf{m})}) & \text{if } \mathbf{m} \in \mathcal{M} \\ +\infty & \text{otherwise} \end{cases} \quad (38)$$

(b) Partition function has variational representation

$$A(\theta) = \sup_{\mathbf{m} \in \mathcal{M}} \{ \langle \theta, \mathbf{m} \rangle - A^*(\mathbf{m}) \} \quad (39)$$

(c) For $\theta \in \Omega$, sup occurs at $\mathbf{m} \in \mathcal{M}$ at matching conditions

$$\mathbf{m} = \int_{\mathcal{D}_X} \psi(x) p_{\theta}(x) \nu(dx) = \mathbb{E}_{\theta}[\psi(X)] = \nabla A(\theta) \quad (40)$$

Conjugate Duality

- Note that A^* isn't exactly entropy, but is only entropy sometimes
- $A(\theta)$ in previous expression is the “inference” problem (dual of the dual) for a given θ , whenever $m \notin \mathcal{M}$ we've got $-\infty$ which can't be sup, so need only consider \mathcal{M} .
- 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

- Some good news: The above form gives us new avenues to do approximation.
- Surprisingly, this is strongly related to belief propagation (i.e., the sum-product commutative semiring)!!

LBP and Bethe Approximation

- 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, or convertible, as we've mentioned, or can define things via a factor graph).
- Exponential 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} .

$$\theta_v(x_v) \triangleq \sum_i \theta_{v,i} \mathbf{1}(x_v = i) \text{ and} \quad (41)$$

$$\theta_{s,t}(x_s, x_t) \triangleq \sum_{i,j} \theta_{st,ij} \mathbf{1}(x_s = i, x_t = j) \quad (42)$$

LBP and Bethe Approximation

Marginal polytope

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

$$\mathbf{m}_v(x_v) \triangleq \sum_{i \in D_{X_v}} \mathbf{m}_{v,i} \mathbf{1}(x_v = i) \quad (43)$$

$$\mathbf{m}_{st}(x_s, x_t) \triangleq \sum_{(j,k) \in D_{X_{\{s,t\}}}} \mathbf{m}_{st,jk} \mathbf{1}(x_s = j, x_t = k) \quad (44)$$

$$(45)$$

- And $\mathbb{M}(G)$ corresponds to the set of all singleton and pairwise marginals that can be jointly realized by some underlying probability distribution $p \in \mathcal{F}(G, R^{(f)})$ that contains only pairwise interactions.
- \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 that contains \mathbb{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. Lets call this \mathbb{L} .
- Another way to form outer bound: require only consistency, i.e., consider set $\{b_v, v \in V(G)\} \cup \{b_{s,t}, (s, t) \in E(G)\}$ that is non-negative and satisfies normalization

$$\sum_{x_v} b_v(x_v) = 1 \quad (46)$$

and pair-node marginal consistency constraints

$$\sum_{x'_t} b_{s,t}(x_s, x'_t) = b_s(x_s) \quad (47)$$

$$\sum_{x'_s} b_{s,t}(x'_s, x_t) = b_t(x_t) \quad (48)$$

(49)

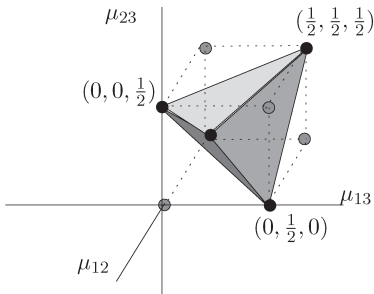
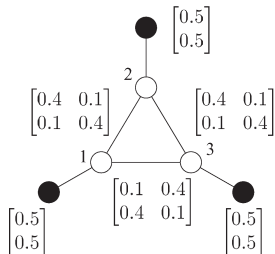
Local consistency polytope

- Define $\mathbb{L}(G)$ to be the (locally consistent) polytope that obeys these constraints.
- 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 issue is that members of \mathbb{L} might not be possible marginals for any distribution.

Pseudo-marginals

$$b_V(x_V) = [0.5, 0.5], \text{ and } b_{s,t}(x_s, x_t) = \begin{bmatrix} \beta_{st} & .5 - \beta_{st} \\ .5 - \beta_{st} & \beta_{st} \end{bmatrix} \quad (50)$$

- 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 $m_1(x_1 = 1) = m_2(x_2 = 1) = m_3(x_3 = 1) = 1/2$ and the $\mathbb{L}(C_3)$ outer bound.



Bethe Entropy Approximation

- Maybe it is hard to compute $A^*(\mathbf{m})$ but perhaps we can reasonably approximate it.
- In case when $-A^*(\mathbf{m})$ is the entropy, let's use an approximate entropy based on \mathbb{L} being those distributions that factor w.r.t. a tree.
- When $G = T$ is a tree, we have

$$-A^*(\mathbf{m}) = H(p_{\mathbf{m}}) = \sum_{v \in V(T)} H(X_v) - \sum_{(s,t) \in E(T)} I(X_s; X_t) \quad (51)$$

$$= \sum_{v \in V(T)} H_v(\mathbf{m}_v) - \sum_{(s,t) \in E(T)} I_{st}(\mathbf{m}_{st}) \quad (52)$$

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

$$-A^*(b) \approx H_{\text{Bethe}}(b) \triangleq \sum_{v \in V(G)} H_v(b_v) - \sum_{(s,t) \in E(G)} I_{st}(b_{st})$$

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mathbf{m} \in \mathcal{M}} \{ \langle \theta, \mathbf{m} \rangle - A^*(\mathbf{m}) \} \quad (53)$$

Approximate variational representation of log partition function

$$A_{\text{Bethe}}(\theta) = \sup_{b \in \mathbb{L}} \{ \langle \theta, b \rangle + H_{\text{Bethe}}(b) \} \quad (54)$$

$$= \sup_{b \in \mathbb{L}} \left\{ \langle \theta, b \rangle + \sum_{v \in V(G)} H_v(b_v) - \sum_{(s,t) \in E(G)} I_{st}(b_{st}) \right\} \quad (55)$$

- 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.

Bethe Variational Problem and LBP

- Lagrangian constraints for summing to unity at nodes

$$C_{vv}(b) = 1 - \sum_{x_v} b_v(x_v) \quad (56)$$

- Lagrangian constraints for local consistency

$$C_{ts}(x_s; b) = b_s(x_s) - \sum_{x_t} b_{st}(x_s, x_t) \quad (57)$$

- Yields following Lagrangian

$$\mathcal{L}(b, \lambda; \theta) = \langle \theta, b \rangle + H_{\text{Bethe}}(b) + \sum_{v \in V} \lambda_{vv} C_{vv}(b) \quad (58)$$

$$+ \sum_{(s,t) \in E(G)} \left[\sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s; b) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t; b) \right] \quad (59)$$

Fixed points: Variational Problem and LBP

Theorem

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

(a) *For any G , any LBP fixed point specifies a pair (b^*, λ^*) s.t.*

$$\nabla_b \mathcal{L}(b^*, \lambda^*; \theta) = 0 \text{ and } \nabla_\lambda \mathcal{L}(b^*, \lambda^*; \theta) = 0 \quad (60)$$

(b) *For tree MRFs, Lagrangian equations have unique solution (b^*, λ^*) where b^* 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.

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

$$\lambda_{st}(x_t) = \mu_{s \rightarrow t}(x_t) \quad (61)$$

- Proof is not too difficult. Just take derivatives, 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 which provide a lower bound on $A(\theta)$).
- For certain potential functions, we'll get $A_{\text{Bethe}}(\theta) \leq A(\theta)$