EE512A – Advanced Inference in Graphical Models

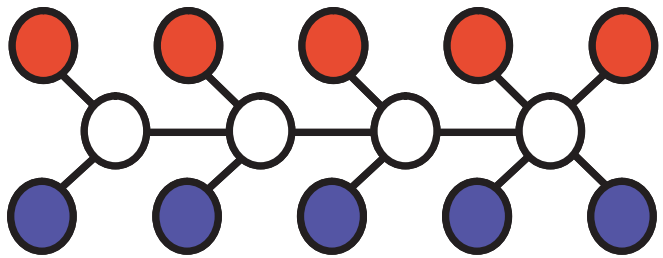
— Fall Quarter, Lecture 12 —

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 10th, 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,
- L13 (11/10): exponential models, mean params and polytopes, tree outer bound
- L14 (11/12):
- L15 (11/17):
- L16 (11/19):
- L17 (11/24):
- L18 (11/26):
- L19 (12/1):
- L20 (12/3):
- Final Presentations: (12/10):

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

Power method lemma

Theorem 12.2.1 (Power method lemma)

Let A be a matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$ (sorted in decreasing order) and corresponding eigenvectors $x_1, x_2, \ldots, x_n$. If $|\lambda_1| > |\lambda_2|$ (strict), then the update $x^{t+1} = \alpha A x^t$ 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 12.2.1

1. *$\mu_{\ell \to 1}$ converges to the principle eigenvector of M.*
2. *$\mu_{2 \to 1}$ converges to the principle eigenvector of M^T.*
3. *The convergence rate is determined by the ratio of the largest 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. □

exponential family models

- $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ is a collection of functions known as potential functions, sufficient statistics, or features. $\mathcal{I}$ is an index set of size $d = |\mathcal{I}|$.
- Each ϕ_α is a function of x, $\phi_\alpha(x)$ but it usually does not use all of x (only a subset of elements). Notation $\phi_\alpha(x_{C_\alpha})$ assumed implicitly understood, where $C_\alpha \subseteq V(G)$.
- θ is a vector of canonical parameters (same length, $|\mathcal{I}|$). $\theta \in \Omega \subseteq \mathbb{R}^d$ where $d = |\mathcal{I}|$.
- We can define a family as

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.12}$$

where $\langle \theta, \phi(x) \rangle = \sum_\alpha \theta_\alpha \phi_\alpha(x)$. Note that we're using ϕ here in the exponent, before we were using it out of the exponent.

- Note that $\phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_{|\mathcal{I}|})$ where again each $\phi_i(x)$ might use only some of the elements in vector x. $\phi : \mathsf{D}_X{}^m \to \mathbb{R}^d$.

Log partition (cumulant) function

- Based on underlying set of parameters θ, we have family of models

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp\left\{\sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(x)\right\} = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \quad (12.12)$$

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

$$A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \quad (12.13)$$

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

- $A(\theta)$ is convex function of θ, so Ω is convex.
- Exponential family for which Ω is open is called regular

Maximum entropy estimation

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

$$p^* \in \operatorname*{argmax}_{p \in \mathcal{U}} H(p) \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \hat{\mu}_\alpha \ \forall \alpha \in \mathcal{I} \tag{12.14}$$

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

$$\mathbb{E}_p[\phi_\alpha(X)] = \int_{\mathsf{D}_X} \phi_\alpha(x) p(x) \nu(dx). \tag{12.15}$$

- $\mathbb{E}_p[\phi_\alpha(X)]$ 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

$$E_{p_\theta}[\phi_\alpha(X)] = \hat{\mu}_\alpha \text{ for all } \alpha \in \mathcal{I}. \tag{12.16}$$

Maximum entropy solution

- Solution to maxent problem

$$p^* \in \operatorname*{argmax}_{p \in \mathcal{U}} H(p) \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \hat{\mu}_\alpha \;\; \forall \alpha \in \mathcal{I} \tag{12.14}$$

has the form of an exponential model:

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.15}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{12.16}$$

- Exercise: show that solution to Eqn (**??**) has this form.

Minimal Representation of Exponential Family

- Minimal representation - Does not exist a nonzero vector $\gamma \in \mathbb{R}^d$ for which $\langle \gamma, \phi(x) \rangle$ is constant $\forall x$ (that are ν-measurable).
- I.e., guarantee that, for all non-zero $\gamma \in \mathbb{R}^d$, there exists $x_1 \neq x_2$, with $\nu(x_1), \nu(x_2) > 0$, such that $\langle \gamma, \phi(x_1) \rangle \neq \langle \gamma, \phi(x_2) \rangle$.
- essential idea: that for a set of sufficient stats $\mathcal{I}$, there is not a lower-dimensional vector $|\mathcal{I}'| < |\mathcal{I}|$ that is also sufficient (a min suf stat is a function of all other suf stats).
- We can't reduce the dimensionality d without changing the family.

Overcomplete Representation

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.14}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{12.15}$$

- Overcomplete representation $d = |\mathcal{I}|$ higher than need be
- I.e., $\exists \gamma \neq 0$ s.t. $\langle \gamma, \phi(x) \rangle = c$, $\forall x$ where $c =$ constant.
- I.e., Exists affine hyperplane of different parameters that induce exactly same distribution. Assume overcomplete, given $\gamma \neq 0$ s.t., $\langle \gamma, \phi(x) \rangle = c$ and some other parameters θ, we have , we have

$$p_{\theta+\gamma}(x) = \exp(\langle (\theta + \gamma), \phi(x) \rangle - A(\theta + \gamma)) \tag{12.16}$$

$$= \exp(\langle \theta, \phi(x) \rangle + \langle \gamma, \phi(x) \rangle - A(\theta + \gamma)) \tag{12.17}$$

$$= \exp(\langle \theta, \phi(x) \rangle + c - A(\theta + \gamma)) \tag{12.18}$$

$$= \exp(\langle \theta, \phi(x) \rangle - A(\theta)) = p_\theta(x) \tag{12.19}$$

- True for any $\lambda\gamma$ with $\lambda \in \mathbb{R}$, so affine set of identical distributions!
- We'll see later, this useful in understanding BP algorithm.

Exponential family models

- Minimal representation of Bernoulli distribution is

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{12.1}$$

So $p(X = 1) = 1 - p(X = 0) = \exp(\gamma - A(\gamma))$ and $p(X = 0) = \exp(-A(\gamma))$.

Exponential family models

- Minimal representation of Bernoulli distribution is

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{12.1}$$

So $p(X = 1) = 1 - p(X = 0) = \exp(\gamma - A(\gamma))$ and $p(X = 0) = \exp(-A(\gamma))$.

- overcomplete rep of Bernoulli dist.

$$p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \tag{12.2}$$

$$= \exp(\theta_0(1-x) + \theta_1 x - A(\gamma)) \tag{12.3}$$

where $\theta = (\theta_0, \theta_1)$ and $\phi(x) = (1 - x, x)$.

Exponential family models

- Minimal representation of Bernoulli distribution is

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{12.1}$$

  So $p(X=1) = 1 - p(X=0) = \exp(\gamma - A(\gamma))$ and $p(X=0) = \exp(-A(\gamma))$.
- overcomplete rep of Bernoulli dist.

$$p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \tag{12.2}$$

$$= \exp(\theta_0(1-x) + \theta_1 x - A(\gamma)) \tag{12.3}$$

  where $\theta = (\theta_0, \theta_1)$ and $\phi(x) = (1-x, x)$.
- Is there a non-zero vector a s.t. $\langle a, \phi(x) \rangle = c$ for all x, ν-a.e.?

Exponential family models

- Minimal representation of Bernoulli distribution is

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{12.1}$$

So $p(X=1) = 1 - p(X=0) = \exp(\gamma - A(\gamma))$ and $p(X=0) = \exp(-A(\gamma))$.

- overcomplete rep of Bernoulli dist.

$$p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \tag{12.2}$$

$$= \exp(\theta_0(1-x) + \theta_1 x - A(\gamma)) \tag{12.3}$$

where $\theta = (\theta_0, \theta_1)$ and $\phi(x) = (1-x, x)$.

- Is there a non-zero vector a s.t. $\langle a, \phi(x) \rangle = c$ for all x, ν-a.e.?
- If $a = (1,1)$ then $\langle a, \phi(x) \rangle = (1-x) + x = 1$

Exponential family models

- Minimal representation of Bernoulli distribution is

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{12.1}$$

So $p(X=1) = 1 - p(X=0) = \exp(\gamma - A(\gamma))$ and $p(X=0) = \exp(-A(\gamma))$.

- overcomplete rep of Bernoulli dist.

$$p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \tag{12.2}$$

$$= \exp(\theta_0(1-x) + \theta_1 x - A(\gamma)) \tag{12.3}$$

where $\theta = (\theta_0, \theta_1)$ and $\phi(x) = (1-x, x)$.

- Is there a non-zero vector a s.t. $\langle a, \phi(x) \rangle = c$ for all x, ν-a.e.?
- If $a = (1,1)$ then $\langle a, \phi(x) \rangle = (1-x) + x = 1$
- This is overcomplete since there is a linear combination of feature functions that are constant.

Exponential family models

- Minimal representation of Bernoulli distribution is

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{12.1}$$

So $p(X=1) = 1 - p(X=0) = \exp(\gamma - A(\gamma))$ and $p(X=0) = \exp(-A(\gamma))$.

- overcomplete rep of Bernoulli dist.

$$p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.2}$$

$$= \exp(\theta_0(1-x) + \theta_1 x - A(\theta)) \tag{12.3}$$

where $\theta = (\theta_0, \theta_1)$ and $\phi(x) = (1-x, x)$.

- Is there a non-zero vector a s.t. $\langle a, \phi(x) \rangle = c$ for all x, ν-a.e.?
- If $a = (1,1)$ then $\langle a, \phi(x) \rangle = (1-x) + x = 1$
- This is overcomplete since there is a linear combination of feature functions that are constant.
- Since $\theta_0(1-x) + \theta_1 x = \theta_0 + x(\theta_1 - \theta_0)$, any parameters θ_1, θ_2 such that $\theta_1 - \theta_0 = \gamma$ gives same distribution determined by γ.

Famous Example - Ising Model

- Famous example is the Ising model in statistical physics. We have a grid network with pairwise interactions, each variable is 0/1-valued binary, and parameters associated with pairs being both on. Model becomes

$$p_\theta(x) = \exp\left\{\sum_{v\in V}\theta_v x_v + \sum_{(s,t)\in E}\theta_{st}x_s x_t - A(\theta)\right\}, \qquad (12.4)$$

with

$$A(\theta) = \log \sum_{x\in\{0,1\}^m} \exp\left\{\sum_{v\in V}\theta_v x_v + \sum_{(s,t)\in E}\theta_{st}x_s x_t - \cdots\right\} \qquad (12.5)$$

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

Ising Model and Immediate Generalization

- Note, in this case $\mathcal{I}$ is all singletons (unaries) and all pairs, so that $\{C_\alpha\}_\alpha = \left\{\{x_i\}_i, \{x_i x_j\}_{(i,j)\in E}\right\}$.
- We can easily generalize this via a set system. I.e., consider $(V, \mathcal{V})$, where $\mathcal{V} = \{V_1, V_2, \ldots, V_{|\mathcal{V}|}\}$ and where $\forall i, V_i \subseteq V$.
- We can form sufficient statistic set via $\{C_\alpha\}_\alpha = \{\{x_V\}_{V\in\mathcal{V}}\}$.
- Could have, for example that $\phi_\alpha = \prod_{i\in C_\alpha} x_i$.
- Hence, it is possible to generalize with higher order factors (which are also called "interaction functions", "potential functions", or "sufficient statistics").

Multivalued variables

- Variables need not binary, instead $\mathsf{D}_X = \{0, 1, \ldots, r-1\}$ for $r > 2$.

Multivalued variables

- Variables need not binary, instead $\mathsf{D}_X = \{0, 1, \ldots, r-1\}$ for $r > 2$.
- We can define a set of indicator functions constituting sufficient statistics. That is

$$\mathbf{1}_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{else} \end{cases} \tag{12.6}$$

and

$$\mathbf{1}_{st;jk}(x_s, x_t) = \begin{cases} 1 & \text{if } x_s = j \text{ and } x_t = k, \\ 0 & \text{else} \end{cases} \tag{12.7}$$

Multivalued variables

- Variables need not binary, instead $\mathsf{D}_X = \{0, 1, \ldots, r-1\}$ for $r > 2$.
- We can define a set of indicator functions constituting sufficient statistics. That is

$$\mathbf{1}_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{else} \end{cases} \tag{12.6}$$

and

$$\mathbf{1}_{st;jk}(x_s, x_t) = \begin{cases} 1 & \text{if } x_s = j \text{ and } x_t = k, \\ 0 & \text{else} \end{cases} \tag{12.7}$$

- Model becomes

$$p_\theta(x) = \exp\left\{ \sum_{v \in V} \sum_{i=0}^{r-1} \theta_{v;j} \mathbf{1}_{s;j}(x_v) + \sum_{(s,t) \in E} \sum_{j,k} \theta_{st;ij} \mathbf{1}_{st;jk}(x_s, x_t) - A(\theta) \right\} \tag{12.8}$$

Multivalued variables

- Variables need not binary, instead $\mathsf{D}_X = \{0, 1, \ldots, r-1\}$ for $r > 2$.
- We can define a set of indicator functions constituting sufficient statistics. That is

$$\mathbf{1}_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{else} \end{cases} \tag{12.6}$$

and

$$\mathbf{1}_{st;jk}(x_s, x_t) = \begin{cases} 1 & \text{if } x_s = j \text{ and } x_t = k, \\ 0 & \text{else} \end{cases} \tag{12.7}$$

- Model becomes

$$p_\theta(x) = \exp\left\{ \sum_{v \in V} \sum_{i=0}^{r-1} \theta_{v;j} \mathbf{1}_{s;j}(x_v) + \sum_{(s,t) \in E} \sum_{j,k} \theta_{st;ij} \mathbf{1}_{st;jk}(x_s, x_t) - A(\theta) \right\} \tag{12.8}$$

- Is this overcomplete?

Multivalued variables

- Variables need not binary, instead $\mathsf{D}_X = \{0, 1, \ldots, r-1\}$ for $r > 2$.
- We can define a set of indicator functions constituting sufficient statistics. That is

$$\mathbf{1}_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{else} \end{cases} \tag{12.6}$$

and

$$\mathbf{1}_{st;jk}(x_s, x_t) = \begin{cases} 1 & \text{if } x_s = j \text{ and } x_t = k, \\ 0 & \text{else} \end{cases} \tag{12.7}$$

- Model becomes

$$p_\theta(x) = \exp\left\{ \sum_{v \in V} \sum_{i=0}^{r-1} \theta_{v;j} \mathbf{1}_{s;j}(x_v) + \sum_{(s,t) \in E} \sum_{j,k} \theta_{st;ij} \mathbf{1}_{st;jk}(x_s, x_t) - A(\theta) \right\} \tag{12.8}$$

- Is this overcomplete? Yes. Why?

Multivariate Gaussian

- Usually, multivariate Gaussian is parameterized via mean and covariance matrix. For canonical exponential form, we use mean and correlation matrix. I.e.

$$p_\theta(x) = \exp\left\{\langle\theta, x\rangle + \frac{1}{2}\langle\langle\Theta, xx^\intercal\rangle\rangle - A(\theta, \Theta)\right\} \tag{12.9}$$

Multivariate Gaussian

- Usually, multivariate Gaussian is parameterized via mean and covariance matrix. For canonical exponential form, we use mean and correlation matrix. I.e.

$$p_\theta(x) = \exp\left\{\langle\theta, x\rangle + \frac{1}{2}\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle - A(\theta,\Theta)\right\} \tag{12.9}$$

- $\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle = \sum_{ij}\Theta_{ij}x_ix_j$ is Frobenius inner product.

Multivariate Gaussian

- Usually, multivariate Gaussian is parameterized via mean and covariance matrix. For canonical exponential form, we use mean and correlation matrix. I.e.

$$p_\theta(x) = \exp\left\{\langle\theta, x\rangle + \frac{1}{2}\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle - A(\theta, \Theta)\right\} \tag{12.9}$$

- $\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle = \sum_{ij}\Theta_{ij}x_ix_j$ is Frobenius inner product.
- So sufficient statistics are $(x_i)_{i=1}^n$ and $(x_ix_j)_{i,j}$

Multivariate Gaussian

- Usually, multivariate Gaussian is parameterized via mean and covariance matrix. For canonical exponential form, we use mean and correlation matrix. I.e.

$$p_\theta(x) = \exp\left\{\langle\theta, x\rangle + \frac{1}{2}\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle - A(\theta, \Theta)\right\} \tag{12.9}$$

- $\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle = \sum_{ij}\Theta_{ij}x_ix_j$ is Frobenius inner product.
- So sufficient statistics are $(x_i)_{i=1}^n$ and $(x_ix_j)_{i,j}$
- $\Theta_{s,t} = 0$ means identical to missing edge in corresponding graph (marginal independence). Θ is negative inverse covariance matrix.

Multivariate Gaussian

- Usually, multivariate Gaussian is parameterized via mean and covariance matrix. For canonical exponential form, we use mean and correlation matrix. I.e.

$$p_\theta(x) = \exp\left\{\langle\theta, x\rangle + \frac{1}{2}\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle - A(\theta, \Theta)\right\} \tag{12.9}$$

- $\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle = \sum_{ij}\Theta_{ij}x_ix_j$ is Frobenius inner product.
- So sufficient statistics are $(x_i)_{i=1}^n$ and $(x_ix_j)_{i,j}$
- $\Theta_{s,t} = 0$ means identical to missing edge in corresponding graph (marginal independence). Θ is negative inverse covariance matrix.
- Any other constraints on Θ?

Multivariate Gaussian

- Usually, multivariate Gaussian is parameterized via mean and covariance matrix. For canonical exponential form, we use mean and correlation matrix. I.e.

$$p_\theta(x) = \exp\left\{\langle\theta, x\rangle + \frac{1}{2}\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle - A(\theta, \Theta)\right\} \tag{12.9}$$

- $\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle = \sum_{ij}\Theta_{ij}x_ix_j$ is Frobenius inner product.
- So sufficient statistics are $(x_i)_{i=1}^n$ and $(x_ix_j)_{i,j}$
- $\Theta_{s,t} = 0$ means identical to missing edge in corresponding graph (marginal independence). Θ is negative inverse covariance matrix.
- Any other constraints on Θ? negative definite

Multivariate Gaussian

- Usually, multivariate Gaussian is parameterized via mean and covariance matrix. For canonical exponential form, we use mean and correlation matrix. I.e.

$$p_\theta(x) = \exp\left\{\langle\theta, x\rangle + \frac{1}{2}\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle - A(\theta, \Theta)\right\} \tag{12.9}$$

- $\langle\langle\Theta, xx^\mathsf{T}\rangle\rangle = \sum_{ij}\Theta_{ij}x_ix_j$ is Frobenius inner product.
- So sufficient statistics are $(x_i)_{i=1}^n$ and $(x_ix_j)_{i,j}$
- $\Theta_{s,t} = 0$ means identical to missing edge in corresponding graph (marginal independence). Θ is negative inverse covariance matrix.
- Any other constraints on Θ? negative definite
- Mixtures of Gaussians can also be parameterized in exponential form (but note, key is that it is the joint distribution $p_{\theta_s}(y_s, x_s)$).

Other examples

A few other examples in the book

- Mixture models

Other examples

A few other examples in the book

- Mixture models
- Latent Dirichlet Allocation, and general hierarchical Bayesian models. Key here is that it is for one expansion, not variable.

Other examples

A few other examples in the book

- Mixture models
- Latent Dirichlet Allocation, and general hierarchical Bayesian models. Key here is that it is for one expansion, not variable.
- Models with hard constraints, or having zero probabilities — key thing is to place the hard constraints in the ν measure. Sufficient statistics become easy if complexity is encoded in the measure. Alternative is to allow features over extended reals (i.e., a feature can provide $-\infty$ but this leads to certain technical difficulties that they would rather not deal with).

Mean Parameters, Convex Cores

- Consider quantities μ_α associated with statistic ϕ_α defined as:

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \nu(dx) \tag{12.10}$$

Mean Parameters, Convex Cores

- Consider quantities μ_α associated with statistic ϕ_α defined as:

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \nu(dx) \tag{12.10}$$

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

Mean Parameters, Convex Cores

- Consider quantities μ_α associated with statistic ϕ_α defined as:

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \nu(dx) \tag{12.10}$$

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

$$\mathcal{M}(\phi) = \mathcal{M} \triangleq \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)], \forall \alpha \in \mathcal{I} \right\} \tag{12.11}$$

$d = |\mathcal{I}|$

Mean Parameters, Convex Cores

- Consider quantities μ_α associated with statistic ϕ_α defined as:

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \nu(dx) \tag{12.10}$$

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

$$\mathcal{M}(\phi) = \mathcal{M} \stackrel{\Delta}{=} \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)], \ \forall \alpha \in \mathcal{I} \right\} \tag{12.11}$$

- We don't say p was necessarily exponential family

Mean Parameters, Convex Cores

- Consider quantities μ_α associated with statistic ϕ_α defined as:

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \nu(dx) \tag{12.10}$$

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

$$\mathcal{M}(\phi) = \mathcal{M} \stackrel{\Delta}{=} \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)],\ \forall \alpha \in \mathcal{I} \right\} \tag{12.11}$$

- 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'

Mean Parameters, Convex Cores

- Consider quantities μ_α associated with statistic ϕ_α defined as:

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \nu(dx) \tag{12.10}$$

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

$$\mathcal{M}(\phi) = \mathcal{M} \stackrel{\Delta}{=} \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)], \ \forall \alpha \in \mathcal{I} \right\} \tag{12.11}$$

- 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 Gaussians

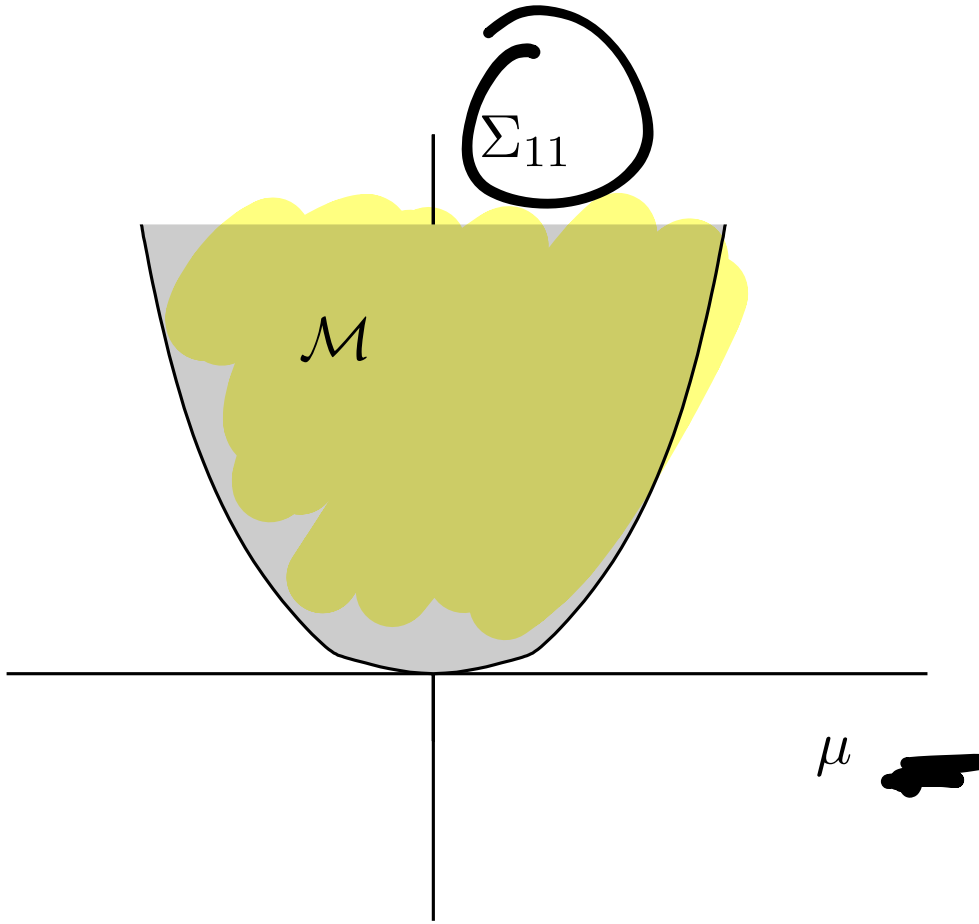
- Here, we have $\mathbb{E}[XX^\mathsf{T}] = C$ and $\mu = \mathbb{E}X$. Question is, how to define $\mathcal{M}$?
- Given definition of C and μ, then $C - \mu\mu^\mathsf{T}$ must be valid covariance matrix (since this is $\mathbb{E}[X - \mathbb{E}X][X - \mathbb{E}X]^\mathsf{T} = C - \mu\mu^\mathsf{T}$).
- Thus, $C - \mu\mu^\mathsf{T} \succeq 0$, thus p.s.d. matrix.
- On the other hand, if this is true, we can form a Gaussian using $C - \mu\mu^\mathsf{T}$ as the covariance matrix.
- Thus, for Gaussian MRFs, $\mathcal{M}$ has the form

$$\mathcal{M} = \left\{(\mu, C) \in \mathbb{R}^m \times \mathcal{S}_+^m | C - \mu\mu^\mathsf{T} \succeq 0\right\} \tag{12.12}$$

where $\mathcal{S}_+^m$ is the set of symmetric positive semi-definite matrices.

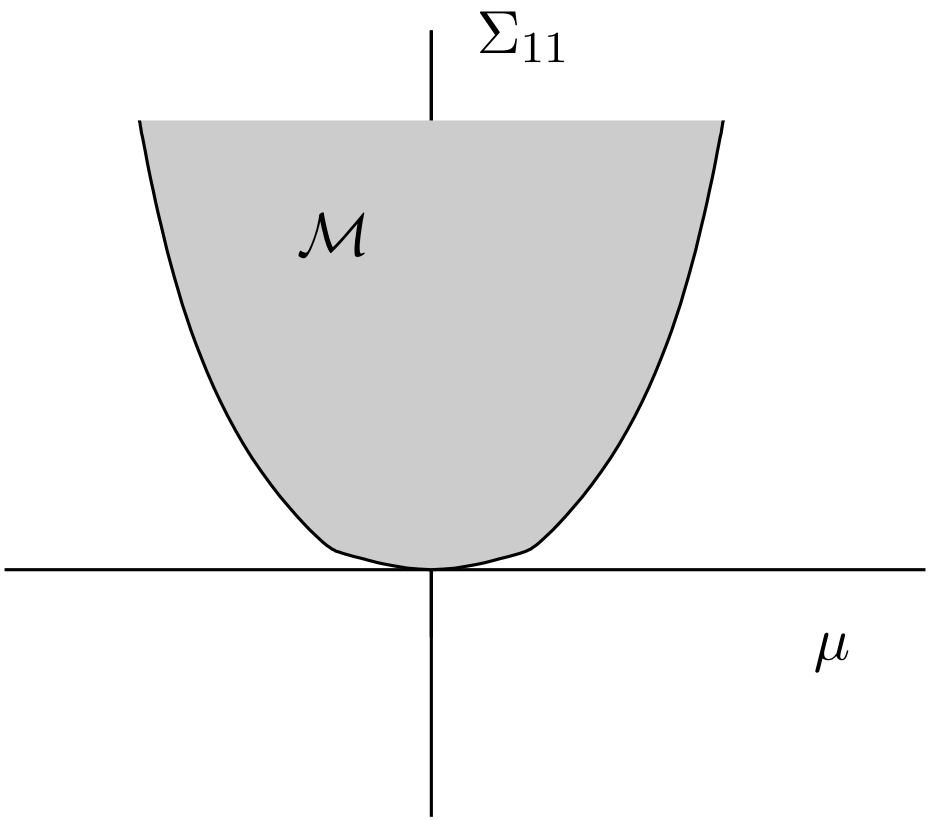
Mean Parameters and Gaussians

- "Illustration of the set $\mathcal{M}$ for a scalar Gaussian: the model has two mean parameters $\mu = \mathbb{E}[X]$ and $\Sigma_{11} = \mathbb{E}[X^2]$, which must satisfy the quadratic contraint $\Sigma_{11} - \mu^2 \geq 0$. Notice that $\mathcal{M}$ is convex, which is a general property." but is not a polytope.



Mean Parameters and Gaussians

- "Illustration of the set $\mathcal{M}$ for a scalar Gaussian: the model has two mean parameters $\mu = \mathbb{E}[X]$ and $\Sigma_{11} = \mathbb{E}[X^2]$, which must satisfy the quadratic contraint $\Sigma_{11} - \mu^2 \geq 0$. Notice that $\mathcal{M}$ is convex, which is a general property." but is not a polytope.
- Also, don't confuse the "mean parameters" with the means of a Gaussian. The typical means of Gaussians are means in this new sense, but those means are not all of the means. ☺



Mean Parameters and Polytopes

- When X is discrete, we get a polytope since

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^b : \mu = \sum_x \phi(x) p(x) \text{ for some } p \in \mathcal{U} \right\} \tag{12.13}$$

$$= \text{conv}\left\{ \phi(x), x \in \mathsf{D}_X \text{ (that are } \nu\text{-measurable),} \right\} \tag{12.14}$$

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

Mean Parameters and Polytopes

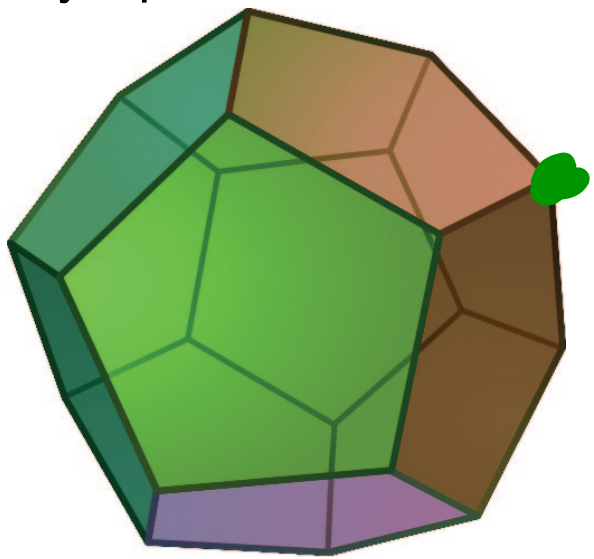
- When X is discrete, we get a polytope since

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^b : \mu = \sum_x \phi(x) p(x) \text{ for some } p \in \mathcal{U} \right\} \quad (12.13)$$

$$= \text{conv}\left\{ \phi(x), x \in \mathsf{D}_X \text{ (that are } \nu\text{-measurable),} \right\} \quad (12.14)$$

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

- So we have a convex polytope

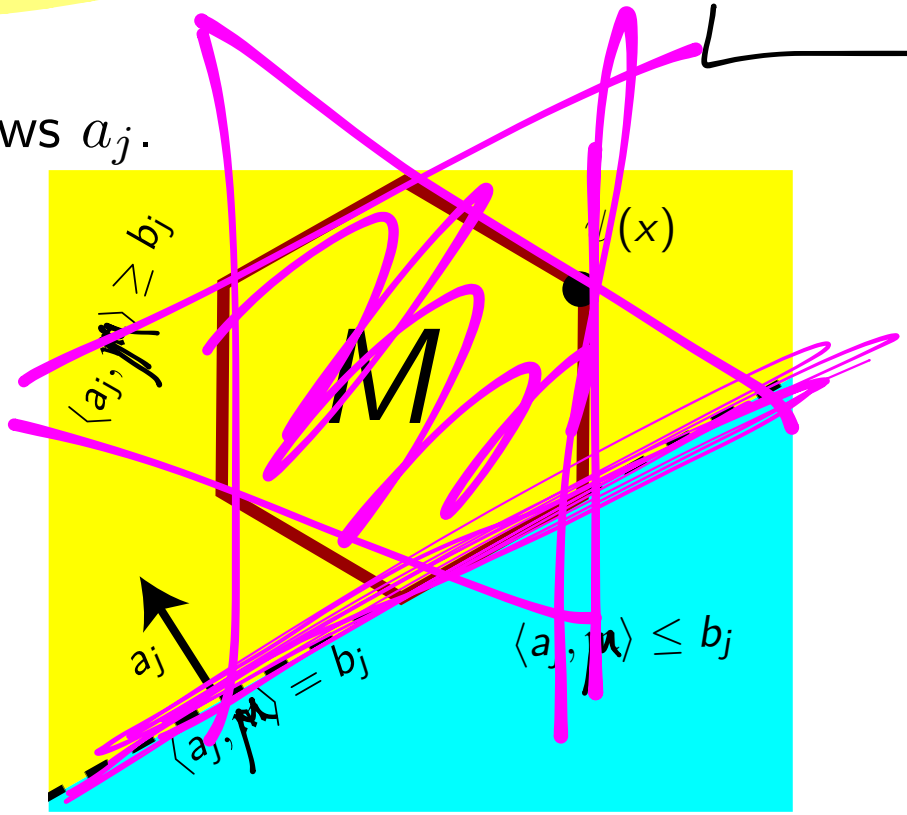


Mean Parameters and Polytopes

- 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

$$M = \left\{ \mu \in \mathbb{R}^d : A\mu \geq b \right\} = \left\{ \mu \in \mathbb{R}^d : \langle a_j, \mu \rangle \geq b_j, \forall j \in J \right\} \quad (12.15)$$

with A having rows a_j.



Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

$$\phi(x) = \{x_s, s \in V; x_s x_t, (s,t) \in E(G)\} \in \mathbb{R}^{|V|+|E|} \tag{12.16}$$

we get

$$\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \ \ \forall v \in V \tag{12.17}$$

$$\mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \ \ \forall (s,t) \in E(G) \tag{12.18}$$

Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

$$\phi(x) = \{x_s, s \in V; x_s x_t, (s,t) \in E(G)\} \in \mathbb{R}^{|V|+|E|} \tag{12.16}$$

we get

$$\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \;\; \forall v \in V \tag{12.17}$$

$$\mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \;\; \forall (s,t) \in E(G) \tag{12.18}$$

- 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 $= \text{conv}\{\phi(x), x \in \{0,1\}^m\}$.

Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

$$\phi(x) = \{x_s, s \in V; x_s x_t, (s,t) \in E(G)\} \in \mathbb{R}^{|V|+|E|} \tag{12.16}$$

we get

$$\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \;\; \forall v \in V \tag{12.17}$$
$$\mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \;\; \forall (s,t) \in E(G) \tag{12.18}$$

- 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 $= \text{conv}\{\phi(x), x \in \{0,1\}^m\}$.
- 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.

Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

$$\phi(x) = \{x_s, s \in V; x_s x_t, (s,t) \in E(G)\} \in \mathbb{R}^{|V|+|E|} \tag{12.16}$$

we get

$$\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \;\; \forall v \in V \tag{12.17}$$

$$\mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \;\; \forall (s,t) \in E(G) \tag{12.18}$$

- 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 $= \text{conv}\{\phi(x), x \in \{0,1\}^m\}$.
- 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.

Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

$$\phi(x) = \{x_s, s \in V; x_s x_t, (s,t) \in E(G)\} \in \mathbb{R}^{|V|+|E|} \tag{12.16}$$

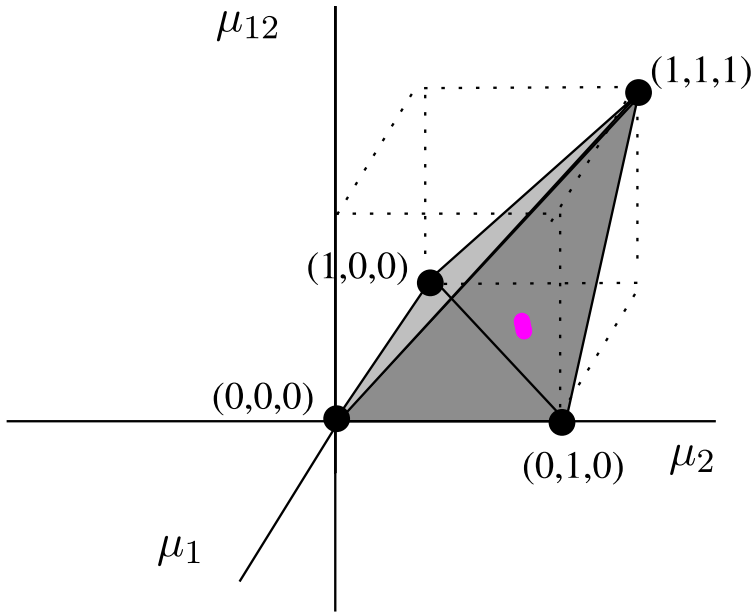
we get

$$\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \;\; \forall v \in V \tag{12.17}$$

$$\mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \;\; \forall (s,t) \in E(G) \tag{12.18}$$

- 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 $= \text{conv}\{\phi(x), x \in \{0,1\}^m\}$.
- 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* $\mu_1 = \mathbb{E}[X_1]$*,* $\mu_2 = \mathbb{E}[X_2]$*,* $\mu_{12} = \mathbb{E}[X_2 X_2]$*, 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)

$$\forall v \in V(G), j \in \{0 \dots r-1\}, \text{ define } \phi_{v,j}(x_v) \triangleq \mathbf{1}(x_v = j) \quad (12.19)$$

$$\forall (s,t) \in E(G), j,k \in \{0 \dots r-1\}, \text{ we define:} \quad (12.20)$$

$$\phi_{st,jk}(x_s, x_t) \triangleq \mathbf{1}(x_s = j, x_t = k) = \mathbf{1}(x_s = j)\mathbf{1}(x_t = k) \quad (12.21)$$

- 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 (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

$$\mu_{v,j} = \mathbb{E}_p[\mathbf{1}(x_v = j)] = p(x_v = j) \tag{12.22}$$

$$\mu_{st,jk} = \mathbb{E}_p[\mathbf{1}(x_s = j, x_t = k)] = p(x_s = j, x_t = k) \tag{12.23}$$

- 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.

Marginal Polytopes and Facet complexity

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

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.

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.

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.

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 in k

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 in k
- 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 θ to the point in the marginal polytope, called forward mapping, moving from $\theta \in \Omega$ to $\mu \in \mathcal{M}$.

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

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).

Review: Maximum Entropy Estimation

The next slide is (again) a repeat from lecture 11.

Maximum entropy estimation

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

$$p^* \in \operatorname*{argmax}_{p \in \mathcal{U}} H(p) \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \hat{\mu}_\alpha \ \forall \alpha \in \mathcal{I} \tag{12.14}$$

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

$$\mathbb{E}_p[\phi_\alpha(X)] = \int_{\mathsf{D}_X} \phi_\alpha(x) p(x) \nu(dx). \tag{12.15}$$

- $\mathbb{E}_p[\phi_\alpha(X)]$ 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

$$E_{p_\theta}[\phi_\alpha(X)] = \hat{\mu}_\alpha \text{ for all } \alpha \in \mathcal{I}. \tag{12.16}$$

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

$$\ell(\theta, \mathbf{D}) = \frac{1}{M} \sum_{i=1}^{M} \log p_\theta(\bar{x}^{(i)}) = \langle \theta, \hat{\mu} \rangle - A(\theta) \tag{12.24}$$

where empirical means given by

$$\hat{\mu} = \hat{\mathbb{E}}[\phi(X)] = \frac{1}{M} \sum_{i=1}^{M} \phi(\bar{x}^{(i)}) \tag{12.25}$$

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

$$\ell(\theta, \mathbf{D}) = \frac{1}{M} \sum_{i=1}^{M} \log p_\theta(\bar{x}^{(i)}) = \langle \theta, \hat{\mu} \rangle - A(\theta) \tag{12.24}$$

  where empirical means given by

$$\hat{\mu} = \hat{\mathbb{E}}[\phi(X)] = \frac{1}{M} \sum_{i=1}^{M} \phi(\bar{x}^{(i)}) \tag{12.25}$$

- 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:

$$\mathbb{E}_{\hat{\theta}}[\phi(X)] = \hat{\mu} \tag{12.26}$$

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

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

$$\ell(\theta, \mathbf{D}) = \frac{1}{M} \sum_{i=1}^{M} \log p_\theta(\bar{x}^{(i)}) = \langle \theta, \hat{\mu} \rangle - A(\theta) \tag{12.24}$$

  where empirical means given by

$$\hat{\mu} = \hat{\mathbb{E}}[\phi(X)] = \frac{1}{M} \sum_{i=1}^{M} \phi(\bar{x}^{(i)}) \tag{12.25}$$

- 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:

$$\mathbb{E}_{\hat{\theta}}[\phi(X)] = \hat{\mu} \tag{12.26}$$

  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: $H(p) = -\int p(x) \log p(x) \nu(dx)$

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

$$\hat{p}(x) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}(x = \bar{x}^{(i)}) \tag{12.27}$$

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

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

$$\hat{p}(x) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}(x = \bar{x}^{(i)}) \tag{12.27}$$

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

$$\ell(\theta(\hat{u}), \mathbf{D}) = \langle \theta(\hat{u}), \hat{\mu} \rangle - A(\theta(\hat{u})) = \frac{1}{M} \sum_{i=1}^{M} \log p_{\theta(\hat{u})}(\bar{x}^{(i)}) \tag{12.28}$$

$$= \int \hat{p}(x) \log p_{\theta(\hat{\mu})}(x) \nu(dx) = \mathbb{E}_{\hat{p}}[\log p_{\theta(\hat{\mu})}(x)] \tag{12.29}$$

$$= -H_{\hat{p}}[p_{\theta(\hat{\mu})}(x)] = -H_{p_{\theta(\hat{\mu})}}[p_{\theta(\hat{\mu})}(x)] \tag{12.30}$$

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

$$\hat{p}(x) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}(x = \bar{x}^{(i)}) \tag{12.27}$$

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

$$\ell(\theta(\hat{u}), \mathbf{D}) = \langle \theta(\hat{u}), \hat{\mu} \rangle - A(\theta(\hat{u})) = \frac{1}{M} \sum_{i=1}^{M} \log p_{\theta(\hat{u})}(\bar{x}^{(i)}) \tag{12.28}$$

$$= \int \hat{p}(x) \log p_{\theta(\hat{\mu})}(x) \nu(dx) = \mathbb{E}_{\hat{p}}[\log p_{\theta(\hat{\mu})}(x)] \tag{12.29}$$

$$= -H_{\hat{p}}[p_{\theta(\hat{\mu})}(x)] = -H_{p_{\theta(\hat{\mu})}}[p_{\theta(\hat{\mu})}(x)] \tag{12.30}$$

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

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.

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.

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.

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?

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.

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).

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.

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

$$A(\theta) = \log \int_{\mathsf{D}_X} \exp \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{12.31}$$

- If we know the log partition function, we know a lot for an exponential family model. In particular, we know

Log partition (or cumulant) function

$$A(\theta) = \log \int_{\mathsf{D}_X} \exp \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{12.31}$$

- 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).

Log partition (or cumulant) function

$$A(\theta) = \log \int_{\mathsf{D}_X} \exp \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{12.31}$$

- 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 $\phi(X)$

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] = \int \phi_\alpha(X) p_\theta(x) \nu(dx) = \mu_\alpha \tag{12.32}$$

in general, derivative of log part. function is expected value of feature

$$\frac{1}{\int \exp\langle \theta, \phi(x) \rangle dx} \cdot \exp\langle \theta, \phi(x) \rangle \cdot \phi_\alpha$$

Log partition (or cumulant) function

$$A(\theta) = \log \int_{\mathsf{D}_X} \exp \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{12.31}$$

- 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 $\phi(X)$

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] = \int \phi_\alpha(X) p_\theta(x) \nu(dx) = \mu_\alpha \tag{12.32}$$

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

$$\frac{\partial^2 A}{\partial \theta_{\alpha_1} \partial \theta_{\alpha_2}}(\theta) = \mathbb{E}_\theta[\phi_{\alpha_1}(X)\phi_{\alpha_2}(X)] - \mathbb{E}_\theta[\phi_{\alpha_1}(X)]\mathbb{E}_\theta[\phi_{\alpha_2}(X)] \tag{12.33}$$

Log partition (or cumulant) function

$$A(\theta) = \log \int_{\mathsf{D}_X} \exp \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{12.31}$$

- 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 $\phi(X)$

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] = \int \phi_\alpha(X) p_\theta(x) \nu(dx) = \mu_\alpha \tag{12.32}$$

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

$$\frac{\partial^2 A}{\partial \theta_{\alpha_1} \partial \theta_{\alpha_2}}(\theta) = \mathbb{E}_\theta[\phi_{\alpha_1}(X)\phi_{\alpha_2}(X)] - \mathbb{E}_\theta[\phi_{\alpha_1}(X)]\mathbb{E}_\theta[\phi_{\alpha_2}(X)] \tag{12.33}$$

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

Log partition function

- 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).

Log partition function

- 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

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

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

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

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

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

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

Mappings - one-to-one

In fact, we have

Theorem 12.4.1

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

Mappings - one-to-one

In fact, we have

Theorem 12.4.1

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

In fact, we have

Theorem 12.4.1

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 12.4.2

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 12.4.2

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 12.4.2

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 12.4.2

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

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{12.34}$$

Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{12.34}$$

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

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \tag{12.35}$$

Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{12.34}$$

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

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \tag{12.35}$$

- 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

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{12.34}$$

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

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \tag{12.35}$$

- 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*

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu \tag{12.36}$$

Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \operatorname*{argmax}_{\theta} \left(\langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{12.34}$$

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

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left(\langle \theta, \mu \rangle - A(\theta) \right) \tag{12.35}$$

- 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*

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu \tag{12.36}$$

- 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 12.4.3 (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:*

$$A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \in \bar{\mathcal{M}} \end{cases} \tag{12.37}$$

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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.38}$$

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

$$\mu = \int_{\mathsf{D}_X} \phi(x) p_\theta(x) \nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta) \tag{12.39}$$

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

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \mu \tag{12.40}$$

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

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \mu \tag{12.40}$$

- $A(\theta)$ in Equation 12.38 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

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \mu \tag{12.40}$$

- $A(\theta)$ in Equation 12.38 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 12.38, so Equation 12.38 need only consider $\mathcal{M}$.

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.38}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.38}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.38}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.38}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.38}$$

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

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.38}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.38}$$

- 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). ☺☺

Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.38}$$

- 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). ☺☺
- Much of the rest of the class will be above approaches to the above correspond 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}.

$$\theta_v(x_v) \triangleq \sum_i \theta_{v,i} \mathbf{1}(x_v = i) \text{ and} \tag{12.41}$$

$$\theta_{s,t}(x_s, x_t) \triangleq \sum_{i,j} \theta_{st,ij} \mathbf{1}(x_s = i, x_t = j) \tag{12.42}$$

Marginal notation, and graph

Marginal polytope

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

$$\mu_v(x_v) \triangleq \sum_{i \in \mathsf{D}_{X_v}} \mu_{v,i} \mathbf{1}(x_v = i), \text{ for } u \in V(G) \tag{12.43}$$

$$\mu_{st}(x_s, x_t) \triangleq \sum_{(j,k) \in \mathsf{D}_{X_{\{s,t\}}}} \mu_{st,jk} \mathbf{1}(x_s = j, x_t = k), \text{ for } (s,t) \in E(G) \tag{12.44}$$

Marginal notation, and graph

Marginal polytope

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

$$\mu_v(x_v) \stackrel{\Delta}{=} \sum_{i \in \mathsf{D}_{X_v}} \mu_{v,i} \mathbf{1}(x_v = i), \text{ for } u \in V(G) \tag{12.43}$$

$$\mu_{st}(x_s, x_t) \stackrel{\Delta}{=} \sum_{(j,k) \in \mathsf{D}_{X_{\{s,t\}}}} \mu_{st,jk} \mathbf{1}(x_s = j, x_t = k), \text{ for } (s,t) \in E(G) \tag{12.44}$$

- 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.

$$\mu_v(x_v) \stackrel{\Delta}{=} \sum_{i \in \mathsf{D}_{X_v}} \mu_{v,i} \mathbf{1}(x_v = i), \text{ for } u \in V(G) \tag{12.43}$$

$$\mu_{st}(x_s, x_t) \stackrel{\Delta}{=} \sum_{(j,k) \in \mathsf{D}_{X_{\{s,t\}}}} \mu_{st,jk} \mathbf{1}(x_s = j, x_t = k), \text{ for } (s,t) \in E(G) \tag{12.44}$$

- 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.

$$\mu_v(x_v) \stackrel{\Delta}{=} \sum_{i \in \mathsf{D}_{X_v}} \mu_{v,i} \mathbf{1}(x_v = i), \text{ for } u \in V(G) \tag{12.43}$$

$$\mu_{st}(x_s, x_t) \stackrel{\Delta}{=} \sum_{(j,k) \in \mathsf{D}_{X_{\{s,t\}}}} \mu_{st,jk} \mathbf{1}(x_s = j, x_t = k), \text{ for } (s,t) \in E(G) \tag{12.44}$$

- 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.
- $\mathbb{M}$ 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 $\mathbb{L}$.

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 $\{\tau_v, v \in V(G)\} \cup \{\tau_{s,t}, (s,t) \in E(G)\}$ that is non-negative and satisfies normalization

$$\sum_{x_v} \tau_v(x_v) = 1 \tag{12.45}$$

and pair-node marginal consistency constraints

$$\sum_{x'_t} \tau_{s,t}(x_s, x'_t) = \tau_s(x_s) \tag{12.46a}$$

$$\sum_{x'_s} \tau_{s,t}(x'_s, x_t) = \tau_t(x_t) \tag{12.46b}$$

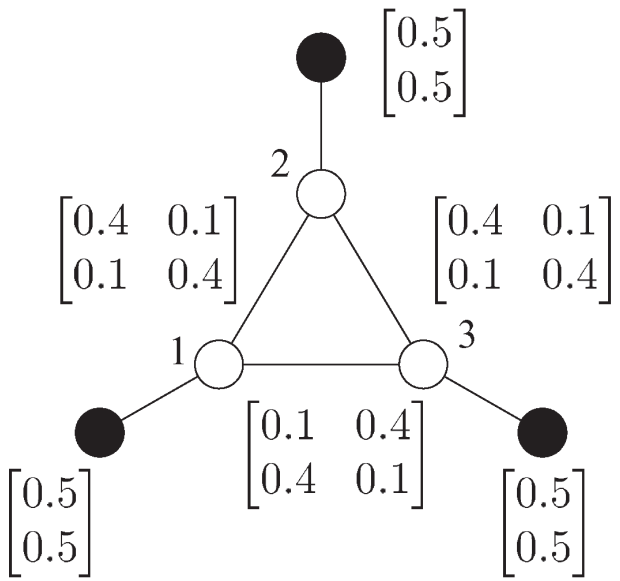
Local consistency polytope

- Define $\mathbb{L}(G)$ to be the (locally consistent) polytope that obeys the constraints in Equations 12.45 and 12.46.
- 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.

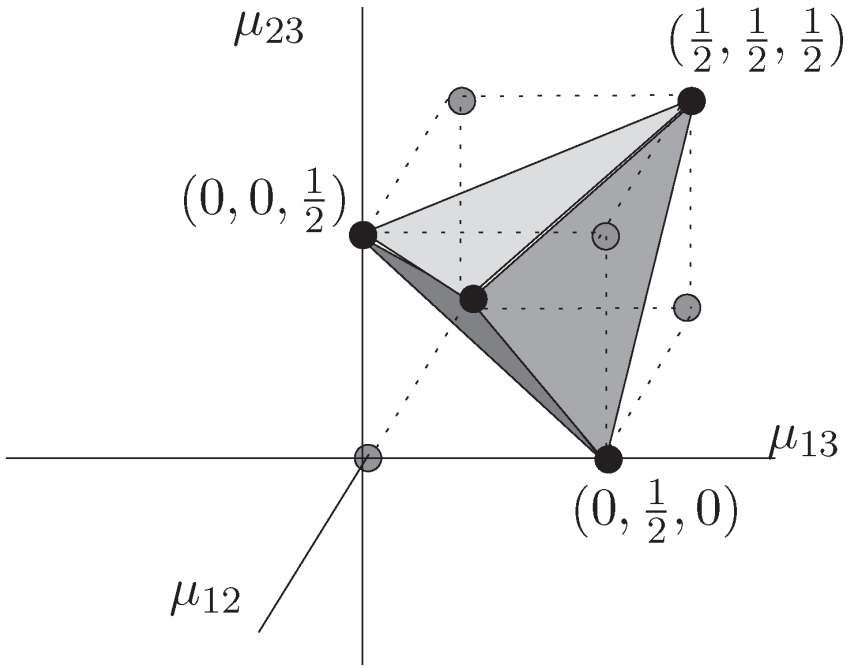
Pseudo-marginals

$$\tau_v(x_v) = [0.5, 0.5], \text{ and } \tau_{s,t}(x_s, x_t) = \begin{bmatrix} \beta_{st} & .5 - \beta_{st} \\ .5 - \beta_{st} & \beta_{st} \end{bmatrix} \quad (12.47)$$

- 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)

Exponential Family: Recap

- Exponential Family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.48}$$

with

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{12.49}$$

Exponential Family: Recap

- Exponential Family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.48}$$

with

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{12.49}$$

- $A(\theta)$ is key.

Exponential Family: Recap

- Exponential Family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.48}$$

with

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{12.49}$$

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

Exponential Family: Recap

- Exponential Family

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.48}$$

with

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{12.49}$$

- $A(\theta)$ 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

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.48}$$

with

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{12.49}$$

- $A(\theta)$ 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.

Bethe Entropy Approximation

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.38}$$

- So inference corresponds to Equation 12.38, 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:

$$p(x_1, \ldots, x_N) = \prod_{v \in V(T)} p_v(x_v) \prod_{(i,j) \in E(T)} \frac{p_{ij}(x_i, x_j)}{p_i(x_i) p_j(x_j)} \tag{12.50}$$

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

$$p_\mu(x) = \prod_{s \in V(T)} \mu_s(x_s) \prod_{(s,t) \in E(T)} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{12.51}$$

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

$$p_\mu(x) = \prod_{s \in V(T)} \mu_s(x_s) \prod_{(s,t) \in E(T)} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{12.51}$$

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

$$-A^*(\mu) = H(p_\mu) = \sum_{v \in V(T)} H(X_v) - \sum_{(s,t) \in E(T)} I(X_s; X_t) \tag{12.52}$$

$$= \sum_{v \in V(T)} H_v(\mu_v) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \tag{12.53}$$

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

$$p_\mu(x) = \prod_{s \in V(T)} \mu_s(x_s) \prod_{(s,t) \in E(T)} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{12.51}$$

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

$$-A^*(\mu) = H(p_\mu) = \sum_{v \in V(T)} H(X_v) - \sum_{(s,t) \in E(T)} I(X_s; X_t) \tag{12.52}$$

$$= \sum_{v \in V(T)} H_v(\mu_v) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \tag{12.53}$$

- 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.,

$$-A^*(\tau) \approx H_{\text{Bethe}}(\tau) \triangleq \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \quad (12.54)$$

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.,

$$-A^*(\tau) \approx H_{\text{Bethe}}(\tau) \triangleq \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \quad (12.54)$$

- 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.,

$$-A^*(\tau) \approx H_{\text{Bethe}}(\tau) \triangleq \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \quad (12.54)$$

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

$$I_{st}(\tau_{st}) = I_{st}(\tau_{st}(x_s, x_t)) \quad (12.55)$$

$$= \sum_{x_s, x_t} \tau_{st}(x_s, x_t) \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)} \quad (12.56)$$

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.57}$$

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.57}$$

Approximate variational representation of log partition function

$$A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{L}} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\} \tag{12.58}$$

$$= \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \tag{12.59}$$

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.57}$$

Approximate variational representation of log partition function

$$A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{L}} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\} \tag{12.58}$$

$$= \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \tag{12.59}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.57}$$

Approximate variational representation of log partition function

$$A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{L}} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\} \tag{12.58}$$

$$= \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \tag{12.59}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{12.57}$$

Approximate variational representation of log partition function

$$A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{L}} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\} \tag{12.58}$$

$$= \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \tag{12.59}$$

- 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

$$C_{vv}(\tau) = 1 - \sum_{x_v} \tau_v(x_v) \tag{12.60}$$

- Lagrangian constraints for local consistency

$$C_{ts}(x_s; \tau) = \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t) \tag{12.61}$$

- Yields following Lagrangian

$$\mathcal{L}(\tau, \lambda; \theta) = \langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau) + \sum_{v \in V} \lambda_{vv} C_{vv}(\tau) \tag{12.62}$$

$$+ \sum_{(s,t) \in E(G)} \left[\sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s; \tau) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t; \tau) \right] \tag{12.63}$$

Fixed points: Variational Problem and LBP

Theorem 12.6.1

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.*

$$\nabla_\tau \mathcal{L}(\tau^*, \lambda^*; \theta) = 0 \text{ and } \nabla_\lambda \mathcal{L}(\tau^*, \lambda^*; \theta) = 0 \tag{12.64}$$

(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

$$\lambda_{st}(x_t) = \mu_{s \to t}(x_t) = \sum_{x_s} \psi_{s,t}(x_s, x_t) \prod_{k \in \delta(s) \setminus \{t\}} \mu_{k \to s}(x_s) \quad (12.65)$$

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

$$\lambda_{st}(x_t) = \mu_{s \to t}(x_t) = \sum_{x_s} \psi_{s,t}(x_s, x_t) \prod_{k \in \delta(s) \setminus \{t\}} \mu_{k \to s}(x_s) \qquad (12.65)$$

- 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

$$\lambda_{st}(x_t) = \mu_{s\to t}(x_t) = \sum_{x_s} \psi_{s,t}(x_s, x_t) \prod_{k\in\delta(s)\setminus\{t\}} \mu_{k\to s}(x_s) \qquad (12.65)$$

- 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

$$\lambda_{st}(x_t) = \mu_{s \to t}(x_t) = \sum_{x_s} \psi_{s,t}(x_s, x_t) \prod_{k \in \delta(s) \setminus \{t\}} \mu_{k \to s}(x_s) \qquad (12.65)$$

- 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

$$\lambda_{st}(x_t) = \mu_{s \to t}(x_t) = \sum_{x_s} \psi_{s,t}(x_s, x_t) \prod_{k \in \delta(s) \setminus \{t\}} \mu_{k \to s}(x_s) \quad (12.65)$$

- 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

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

Bounds on A

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

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

Bounds on A

- Moreover, this does not mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why 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

- Moreover, this does not mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\} \qquad (12.38)$$

Bounds on A

- Moreover, this does not mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\} \qquad (12.38)$$

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

Bounds on A

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

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

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

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

- Recall again the expression for the partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\} \tag{12.38}$$

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

- Due to $\sup$, we might want, an upper bound $A_{\text{approx}}(\theta) \geq A(\theta)$,

Bounds on A

- Moreover, this does not mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \qquad (12.38)$$

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

- Due to $\sup$, we might want, an 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

- Moreover, this does not mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it. Why 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\} \qquad (12.38)$$

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

- Due to $\sup$, we might want, an 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

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.66}$$

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

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.66}$$

- 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

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.66}$$

- 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

$$p(x_A|x_B) = \frac{p(x_{A\cup B})}{p(x_B)} = \frac{\sum_{x_{V\setminus(A\cup B)}} p(x)}{\sum_{x_{V\setminus B}} p(x)} \tag{12.67}$$

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

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{12.66}$$

- 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

$$p(x_A|x_B) = \frac{p(x_{A\cup B})}{p(x_B)} = \frac{\sum_{x_{V\setminus(A\cup B)}} p(x)}{\sum_{x_{V\setminus B}} p(x)} \tag{12.67}$$

  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

$$\mu_s(x_s) = [0.5 \;\; 0.5] \qquad \text{for } s = 1, 2, 3, 4 \tag{12.68a}$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \forall (s,t) \in E(G) \tag{12.68b}$$

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

$$\mu_s(x_s) = [0.5 \;\; 0.5] \qquad \text{for } s = 1, 2, 3, 4 \tag{12.68a}$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \forall (s,t) \in E(G) \tag{12.68b}$$

- 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

$$\mu_s(x_s) = [0.5 \;\; 0.5] \qquad \text{for } s = 1, 2, 3, 4 \tag{12.68a}$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \forall (s,t) \in E(G) \tag{12.68b}$$

- 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

$$H_{\text{Bethe}}(\mu) = 4 \log 2 - 6 \log 2 = -2 \log 2 < 0 \tag{12.69}$$

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

$$\mu_s(x_s) = [0.5 \;\; 0.5] \qquad \text{for } s = 1,2,3,4 \tag{12.68a}$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \forall (s,t) \in E(G) \tag{12.68b}$$

- 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

$$H_{\text{Bethe}}(\mu) = 4\log 2 - 6\log 2 = -2\log 2 < 0 \tag{12.69}$$

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

$$\mu_s(x_s) = [0.5 \;\; 0.5] \qquad \text{for } s = 1,2,3,4 \tag{12.68a}$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \forall (s,t) \in E(G) \tag{12.68b}$$

- 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

$$H_{\text{Bethe}}(\mu) = 4\log 2 - 6\log 2 = -2\log 2 < 0 \tag{12.69}$$

  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 12.6.2

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

$$p_{\tau^*}(x) \triangleq \frac{1}{Z(\tau^*)} \prod_{s \in V} \tau_s^*(x_s) \prod_{(s,t) \in E(G)} \frac{\tau_{st}^*(x_s, x_t)}{\tau_s^*(x_s)\tau_t^*(x_t)} \tag{12.70}$$

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)

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

- Consider

$$\theta_s(x_s) = \log \tau_s(x_s) = \log 0.5 \;\; 0.5] \qquad \text{for } s = 1, 2, 3, 4 \tag{12.71a}$$

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

$$= \log 4 \begin{bmatrix} \beta_{st} & 0.5 - \beta_{st} \\ 0.5 - \beta_{st} & \beta_{st} \end{bmatrix} \; \forall (s,t) \in E(G) \tag{12.71b}$$

- 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. ☹

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`