EE512A – Advanced Inference in Graphical Models

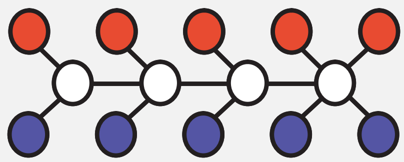
— Fall Quarter, Lecture 12 —

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,
- L12 (11/10): exponential models, mean params and polytopes, tree outer bound
- L13 (11/12): polytopes, tree outer bound, Bethe entropy approx.
- L14 (11/17):
- L15 (11/19):
- L16 (11/24):
- L17 (11/26):
- L18 (12/1):
- L19 (12/3):
- Final Presentations: (12/10):

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

Power method lemma

Theorem 12.2.1 (Power method lemma)

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

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

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

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

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

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

Maximum entropy solution

- Solution to maxent problem

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

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 (12.14) 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

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

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

- 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

$$\begin{aligned} p_{\theta+\gamma}(x) &= \exp(\langle (\theta + \gamma), \phi(x) \rangle - A(\theta + \gamma)) && (12.16) \\ &= \exp(\langle \theta, \phi(x) \rangle + \langle \gamma, \phi(x) \rangle - A(\theta + \gamma)) && (12.17) \\ &= \exp(\langle \theta, \phi(x) \rangle + c - A(\theta + \gamma)) && (12.18) \\ &= \exp(\langle \theta, \phi(x) \rangle - A(\theta)) = p_\theta(x) && (12.19) \end{aligned}$$

- 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

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \quad (12.1)$$

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

$$\begin{aligned} p(x|\theta_0, \theta_1) &= \exp(\langle \theta, \phi(x) \rangle - A(\theta)) && (12.2) \\ &= \exp(\theta_0(1 - x) + \theta_1 x - A(\theta)) && (12.3) \end{aligned}$$

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

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

with

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

- 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 $\mathcal{V} = \left\{V_1, V_2, \dots, V_{|\mathcal{V}|}\right\}$ and where $\forall i, V_i \subseteq V$.
- We can form sufficient statistic set via $\{C_\alpha\}_\alpha = \left\{\{x_V\}_{V\in\mathcal{V}}\right\}$.
- 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 $r > 2$.
- 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.

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

- $\langle\!\langle \Theta, xx^\intercal \rangle\!\rangle = \sum_{ij} \Theta_{ij} x_i x_j$ is Frobenius inner product.
- So sufficient statistics are $(x_i)_{i=1}^n$ and $(x_i x_j)_{i,j}$
- $\Theta_{s,t} = 0$ 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
- 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:

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

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

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

- 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

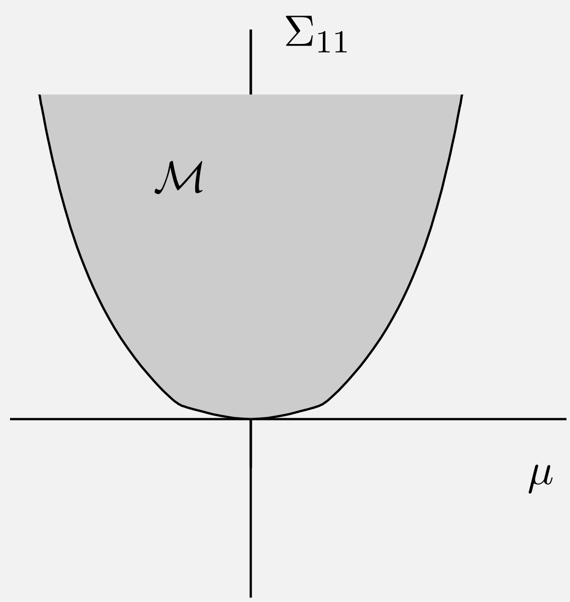
- Here, we have $\mathbb{E}[XX^\intercal] = C$ and $\mu = \mathbb{E}X$. Question is, how to define $\mathcal{M}$?
- Given definition of C and μ, then $C - \mu\mu^\intercal$ must be valid covariance matrix (since this is $\mathbb{E}[X - \mathbb{E}X][X - \mathbb{E}X]^\intercal = C - \mu\mu^\intercal$).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

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

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.
- 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. ☺

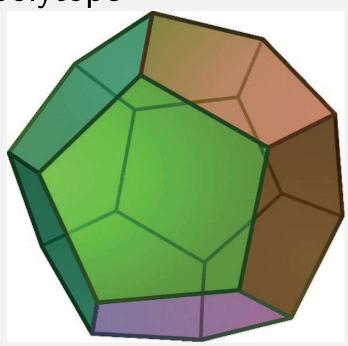


Mean Parameters and Polytopes

- When X is discrete, we get a polytope since

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^b : \mu = \sum_x \phi(x) p(x) \text{ for some } p \in \mathcal{U} \right\} \tag{12.13}$$
$$= \operatorname{conv}\left\{ \phi(x), x \in \mathsf{D}_X \text{ (that are } \nu\text{-measurable),} \right\} \tag{12.14}$$

where $\operatorname{conv}\{\cdot\}$ is the convex hull of the items in argument set.
- So we have a convex polytope

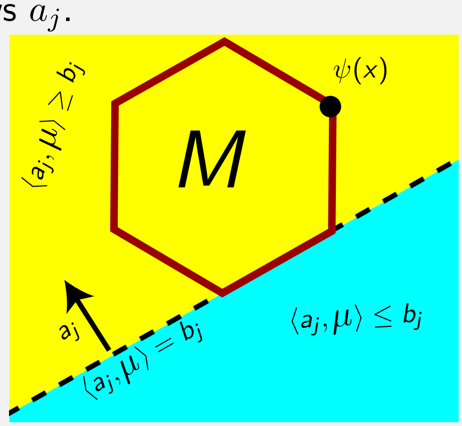


Mean Parameters and Polytopes

- 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

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

with A having rows a_j.



Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

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

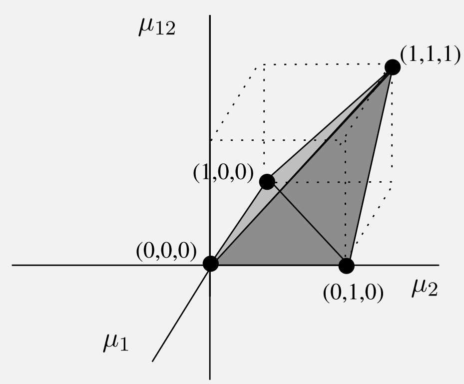
we get

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

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

- Mean parameters 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 $0 \le \mu_{12} \le \mu_i$ for $i = 1,2$, and $1 + \mu_{12} - \mu_1 - \mu_2 \ge 0$. 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

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

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

- 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.
- 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 $\mu \in \mathcal{M}$.
- 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

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

  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

$$\ell(\theta, \mathbf{D}) = \frac{1}{M}\sum_{i=1}^{M} \log p_\theta(\bar{x}^{(i)}) = \frac{1}{M}\sum_{i=1}^{M}\left(\left\langle \theta, \phi(\bar{x}^{(i)})\right\rangle - A(\theta)\right) \quad (12.24)$$

$$= \langle \theta, \hat{\mu}\rangle - A(\theta) \quad (12.25)$$

where empirical means are given by: $$\hat{\mu} = \hat{\mathbb{E}}[\phi(X)] = \frac{1}{M}\sum_{i=1}^{M}\phi(\bar{x}^{(i)}) \quad (12.26)$$

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

$$\mathbb{E}_{\hat{\theta}}[\phi(X)] = \hat{\mu} \quad (12.27)$$

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

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

Likelihood and negative entropy

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

$$\hat{p}(x) = \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}(x = \bar{x}^{(i)}) \quad (12.28)$$

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

$$\ell(\theta(\hat{u}), \mathbf{D}) = \langle \theta(\hat{u}), \hat{\mu}\rangle - A(\theta(\hat{u})) = \frac{1}{M}\sum_{i=1}^{M}\log p_{\theta(\hat{u})}(\bar{x}^{(i)}) \quad (12.29)$$

$$= \int \hat{p}(x)\log p_{\theta(\hat{\mu})}(x)\nu(dx) = \mathbb{E}_{\hat{p}}[\log p_{\theta(\hat{\mu})}(x)] \quad (12.30)$$

$$= \mathbb{E}_{p_{\theta(\hat{\mu})}}[\log p_{\theta(\hat{\mu})}(x)] = -H_{p_{\theta(\hat{\mu})}}[p_{\theta(\hat{\mu})}(x)] \quad (12.31)$$

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

Dual Mappings: Summary

Summarizing these relationships

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

Log partition (or cumulant) function: derivative offerings

$$A(\theta) = \log \int_{\mathsf{D}_X} \exp \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{12.32}$$

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

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] = \int \phi_\alpha(X) p_\theta(x) \nu(dx) = \mu_\alpha \tag{12.33}$$

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

- Also, we get

$$\frac{\partial^2 A}{\partial \theta_{\alpha_1} \partial \theta_{\alpha_2}}(\theta) = \mathbb{E}_\theta[\phi_{\alpha_1}(X)\phi_{\alpha_2}(X)] - \mathbb{E}_\theta[\phi_{\alpha_1}(X)]\mathbb{E}_\theta[\phi_{\alpha_2}(X)] \tag{12.34}$$

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

Log partition function: properties

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

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`