Logistics

Announcements

- Read chapters 1, 2, and 3 in this book
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, mean params and polytopes
- **L12 (11/10):**
- **L13 (11/12):**
- **L14 (11/17):**
- **L15 (11/19):**
- **L16 (11/24):**
- **L17 (11/26):**
- **L18 (12/1):**
- **L19 (12/3):**
- **L20 (12/6):** Final Presentations: (12/10):

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

Prof. Jeff Bilmes

Approximation: Two general approaches

- **exact solution to approximate problem** - approximate problem
  - learning with or using a model with a structural restriction, structure learning, using a $k$-tree for a lower $k$ than one knows is true. Make sure $k$ is small enough so that exact inference can be performed, and make sure that, in that low tree-width model, one has best possible graph
  - Functional restrictions to the model (i.e., use factors or potential functions that obey certain properties). Then certain fast algorithms (e.g., graph-cut) can be performed.

- **approximate solution to exact problem** - approximate inference
  - Message or other form of propagation, variational approaches, LP relaxations, loopy belief propagation (LBP)
  - sampling (Monte Carlo, MCMC, importance sampling) and pruning (e.g., search based A*, score based, number of hypothesis based) procedures
  - Both methods only guaranteed approximate quality solutions.
  - No longer in the achievable region in time-space tradoff graph, new set of time/space tradeoffs to achieve a particular accuracy.
Belief Propagation: message definition

Generic message definition

\[ \mu_{i \rightarrow j}(x_j) = \sum_{x_i} \psi_{i,j}(x_i, x_j) \prod_{k \in \delta(i) \setminus \{j\}} \mu_{k \rightarrow i}(x_i) \quad (11.5) \]

- If graph is a tree, and if we obey MPP order, then we will reach a point where we’ve got marginals. I.e.,

\[ p(x_i) \propto \prod_{j \in \delta(i)} \mu_{j \rightarrow i}(x_i) \quad (11.6) \]

and

\[ p(x_i, x_j) \propto \psi_{i,j}(x_i, x_j) \prod_{k \in \delta(i) \setminus \{j\}} \mu_{k \rightarrow i}(x_i) \prod_{\ell \in \delta(j) \setminus \{i\}} \mu_{\ell \rightarrow j}(x_j)M \quad (11.7) \]

Choices for dealing with higher order factors in MRFs

So, to deal with MRFs with higher order factors, we can:

- transform MRF to have only pairwise interactions, add more variables, we can keep using BP on MRF edges (as done above), makes the math a bit easier, does not change fundamental computational cost. Possible since for any given \( p \), we know the interaction terms.

- Alternatively, we can define BP on factor graphs.

- Alternatively, could define BP directly on the maxcliques of the MRF (but maxcliques are not easy to get in a MRF when not triangulated).

For the remainder of this term, we’ll assume we’ve done the pair-wise transformation (i.e., option 1 above).
State representation

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

Messages as matrix multiply

\[
\mu_{i \rightarrow j}(x_j) \propto \sum_{x_i} \psi_{i,j}(x_i, x_j) \psi_i(x_i) \prod_{k \in \delta(i) \setminus \{j\}} \mu_{k \rightarrow i}(x_i) \\
= \sum_{x_i} \psi'_{i,j}(x_i, x_j) \mu_{j \rightarrow i}(x_i) \\
= (\psi'_{i,j})^T \mu_{j \rightarrow i}
\]

- Here, \( \psi'_{i,j} \) is a matrix and \( \mu_{j \rightarrow i} \) is a column vector.
- Going from state \( \mu^t \) to \( \mu^{t+1} \) is like matrix-vector multiply — group messages from \( \mu^t \) together into one vector representing \( \mu_{j \rightarrow i} \) for each \( (i, j) \in E \), do the matrix-vector update, and store result in new state vector \( \mu^{t+1} \).
- If \( G \) is tree, \( \mu^t \) will converged after \( D \) steps.
Belief Propagation and Cycles

What if graph has cycles?

- MPP causes deadlock since there is no way to start sending messages
- Like before, we can assume that messages have an initial state, e.g.,
  \( \mu_{i\rightarrow j}(x_j) = 1 \) for all \((i, j) \in E(G)\) - note this is bi-directional. This breaks deadlock.
- We can then start sending messages. Will we converge after \( D \) steps? What does \( D \) even mean here?
- No, in fact we could oscillate forever.

Consider odd length cycle (e.g., \( C_3, C_5 \), etc.), \( C_3 \) is sufficient

Assume all messages start out at state \( \mu_{i\rightarrow j} = [1, 0]^T \).

Consider (pairwise) edge functions, for each \( i, j \)

\[
\psi_{ij}(x_i, x_j) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}
\] (11.1)

then we have

\[
\mu_{j\rightarrow k}(x_k) = \sum_{x_j} \psi_{j,k}(x_j, x_k) \mu_{i\rightarrow j}(x_j)
\] (11.2)

or in matrix form

\[
\mu_{j\rightarrow k} = (\psi_{j,k})^T \mu_{i\rightarrow j}
\] (11.3)
Belief Propagation, Cycles, and Oscillation

- Let $\mu_{i \rightarrow j}^t$ be the $t^{th}$ formed message, with $\mu_{i \rightarrow j}^0$ being the starting state at $[1, 0]^T$.
- Then $\mu_{i \rightarrow j}^1 = [0, 1]^T$, $\mu_{i \rightarrow j}^2 = [1, 0]^T$, $\mu_{i \rightarrow j}^3 = [0, 1]^T$, and so on, never converging. In fact, this follows from:

$$
\begin{align*}
\mu_{i \rightarrow j}^{t+1} &= (\psi_{i,j})^T \mu_{k \rightarrow i}^t \\
&= (\psi_{i,j})^T (\psi_{k,i})^T \mu_{j \rightarrow k}^t \\
&= (\psi_{i,j})^T (\psi_{k,i})^T (\psi_{j,k})^T \mu_{i \rightarrow j}^t \\
&= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \mu_{i \rightarrow j}^t \\
&= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \mu_{i \rightarrow j}^t
\end{align*}
$$

Thus, each time we go around the loop in the cycle, the message configuration for each $(i, j)$ will flip, thereby never converging.

Damping the messages? I.e., Let $0 \leq \gamma < 1$ and treat messages as

$$
\mu_{i \rightarrow j}^t \leftarrow \gamma \mu_{i \rightarrow j}^t + (1 - \gamma) \mu_{i \rightarrow j}^{t-1}
$$

Empirical Folklore - if we converge quickly without damping, the quality of the resulting marginals might be good. If we don’t converge quickly, w/o damping, might indicate some problem.

Ways out of this problem: Other message schedules, other forms of the interaction matrices, and other initializations.
Belief Propagation, Cycles, and Oscillation

- If we initialize messages differently, things will turn out better.
- If \( \mu_{i \rightarrow j}^0 = [0.5, 0.5]^T \) then \( \mu_{i \rightarrow j}^{t+1} = \mu_{i \rightarrow j}^t \).
- Damping the messages appropriately will also end up at this configuration.
- Is there a better way to characterize this?

Belief Propagation, Single Cycle

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

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

\[
= M \mu_{\ell \rightarrow 1}^t
\]

- Will this converge to anything?
Belief Propagation, Single Cycle

Theorem 11.3.1 (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 11.3.2

1. $\mu_1 \rightarrow 1$ converges to the principle eigenvector of $M$.
2. $\mu_2 \rightarrow 1$ converges to the principle eigenvector of $M^T$.
3. The convergence rate is determined by the ratio of the 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 $\beta$ 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.
What’s going on with our oscillating example?

- We had \( M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \) which has row-eigenvector matrix
  \[ \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \]
  with corresponding eigenvalues \(-1\) and \(1\).
- Note that any uniform vector will be “converged”, i.e., any vector of the form \([aa]\).
- However, we don’t have the \textit{guaranteed} property of convergence since we don’t have that \(|\lambda_1| > |\lambda_2|\).

Belief Propagation, arbitrary graph

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

- We start with chapter 3 (we assume you will read chapters 1 and 2 on your own).
- We’ll follow the Wainwright and Jordan notation, will point out where it conflicts a bit with the current notation we’ve been using.

---

**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 $\phi_\alpha$ 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 $C_\alpha \subseteq V(G)$.
- $\theta$ 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

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \quad (11.12)$$

where $\langle \theta, \phi(x) \rangle = \sum_\alpha \theta_\alpha \phi_\alpha(x)$. Note that we’re using $\phi$ 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 $x$. $\phi: \mathbb{D}_X^m \rightarrow \mathbb{R}^d$. 

exponential family models and clique features

- Given a graph $G = (V, E)$ we have a set of cliques $C$ of the graph.
- In order to respect the graph, we have to make sure that $\alpha \in \mathcal{I}$ respects the cliques.
- That is, for any $\alpha \in \mathcal{I}$, and feature function $\phi_\alpha(x_{C_\alpha})$ there must be a clique $C \in \mathcal{C}$ such that $C_\alpha \subseteq C$.
- On the other hand, by having a different index set $\mathcal{I}$ we can have more than one feature (sufficient statistic) for a given clique.
- That is, for any given $C \in \mathcal{C}$ we might have multiple $\alpha_1, \alpha_2 \in \mathcal{I}$ such that $C_{\alpha_1} = C_{\alpha_2} = C$ for some clique $C \in \mathcal{C}$.

Example: single scalar discrete random variable $X \in \{1, 2, \ldots, r\}$ might have indicator feature for all possible values $\alpha_i(x) \triangleq 1(x = i)$ — in this case $|C_\alpha| = 1$ for all $\alpha \in \mathcal{I}$ and $C_\alpha = C_{\alpha'}$ for all $\alpha, \alpha' \in \mathcal{I}$.

Could even think of $\{C_\alpha\}_{\alpha \in \mathcal{I}}$ as cliques of some graph, but not necessarily maxcliques.

Likely not dealing with triangulated models. Could be based on cliques, or cliques and subsets of cliques (consider 4-cycle with edges and vertices).

Key: $p \in \mathcal{F}(G, \mathcal{M}(\mathcal{I}))$ by Hammersley-Clifford theorem,
- where $G = (V, E)$ where $V$ is the nodes corresponding to vector $x$,
- and $E$ is formed by using $\{C_\alpha\}_{\alpha \in \mathcal{I}}$ as an edge clique cover: $\exists \alpha \in \mathcal{I}$ such that $u, v \in C_\alpha$ where $u, v \in V(G) \iff$ there is an edge $(u, v) \in E(G)$.
exponential family models

- Exponential models are in our sense sufficient to deal with the computational aspects graphical models.
- We can have \( p \in \mathcal{F}((V, E), \mathcal{M}(f)) \) implies \( p \in \mathcal{F}((V, E + E_1), \mathcal{M}(f)) \) but in some sense, for any \( G \), we want to deal with the models for which \( G \) is tight (we don’t want to use overly complex graph to deal with family that is simpler) – Approx sol. to exact prob.
- Exponential models can represent any factorization, given any factorization in terms of \( \phi \), i.e., \( p(x) = \exp(A(\theta)) \prod_\alpha \exp(\phi_\alpha(x)) \)
- We can often make them log-linear models as well with the right potential functions which won’t increase tree-width of the graph.
- Moreover, exponential family models are incredibly flexible and have a number of desirable properties (e.g., aspects of the log partition function which we will see).

absolutely continuous

- Underlying base measure \( \nu \), so that \( \int f(x)\nu(dx) \) corresponds to \( \sum_i f(x_i) \) for a counting measure, or \( \int f(x)dx \) if not.
- Underlying base measure \( \nu \), \( p \) is absolutely continuous w.r.t. \( \nu \)
- A measure \( \nu \) is absolutely continuous with respect to \( \mu \) if for each \( A \in \mathcal{F} \), \( \mu(A) = 0 \) implies \( \nu(A) = 0 \). In this case \( \nu \) is also said to be dominated by \( \mu \) (if \( \mu \) goes to zero, so must \( \nu \)), and the relation is indicated by \( \nu \ll \mu \).

- If \( \nu \ll \mu \) and \( \mu \ll \nu \), the measures are equivalent, indicated by \( \nu \equiv \nu \).
exponential family models

- Based on underlying set of parameters $\theta$, we have family of models

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{\alpha \in I} \theta_\alpha \phi_\alpha(x) \right\} = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \quad (11.13)$$

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

$$A(\theta) = \log \int_{D_X} \exp (\langle \theta, \phi(x) \rangle) \nu(dx) \quad (11.14)$$

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

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

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

exponential family models

- Based on underlying set of parameters $\theta$, we have family of models

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{\alpha \in I} \theta_\alpha \phi_\alpha(x) \right\} = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \quad (11.15)$$

- Family can arise for a number of reasons, e.g., distribution having maximum entropy but that satisfies certain (moment) constraints.

- Given data $D = \{x^{(i)}_E\}_{i=1}^M$, form the expected statistics (requirements) of a model, with $\bar{x}^{(i)} \sim p(x)$

$$\hat{\mu}_\alpha = \frac{1}{M} \sum_{i=1}^M \phi_\alpha(\bar{x}^{(i)}) \quad (11.16)$$

Thus, $\lim_{M \to \infty} \hat{\mu}_\alpha = E_p[\phi_\alpha(X)] = \mu_\alpha$
Exponential family models

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

\[ p^* \in \arg\max_{p \in \mathcal{U}} H(p) \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \hat{\mu}_\alpha \quad \forall \alpha \in \mathcal{I} \quad (11.17) \]

where \( \forall \alpha \in \mathcal{I} \)

\[ \mathbb{E}_p[\phi_\alpha(X)] = \int_{\mathcal{D}X} \phi_\alpha(x)p(x)\nu(dx) \quad (11.18) \]

- \( \mathbb{E}_p[\phi_\alpha(X)] \) 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 Eq. 11.15, by finding \( \theta \) that solves

\[ E_{p_\theta}[\phi_\alpha(X)] = \hat{\mu}_\alpha \text{ for all } \alpha \in \mathcal{I} \quad (11.19) \]

Minimal Representation of Exponential Family

- Solution has form:

\[ p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \quad (11.20) \]

where \( A(\theta) = \log \int_{\mathcal{D}X} \exp(\langle \theta, \phi(x) \rangle) \nu(dx) \quad (11.21) \)

Exercise: show that solution to Eqn (11.17) has this form.

- 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 \( \nu \)-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

\[ p_{\theta}(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \quad (11.22) \]

where

\[ A(\theta) = \log \int_{D_X} \exp(\langle \theta, \phi(x) \rangle) \nu(dx) \quad (11.23) \]

- 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 \( \theta \), we have, we have

\[ p_{\theta+\gamma}(x) = \exp(\langle (\theta + \gamma), \phi(x) \rangle - A(\theta + \gamma)) \quad (11.24) \]
\[ = \exp(\langle \theta, \phi(x) \rangle + \langle \gamma, \phi(x) \rangle - A(\theta + \gamma)) \quad (11.25) \]
\[ = \exp(\langle \theta, \phi(x) \rangle + c - A(\theta + \gamma)) \quad (11.26) \]
\[ = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) = p_{\theta}(x) \quad (11.27) \]

- 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.

Sources for Today’s Lecture