# EE512A – Advanced Inference in Graphical Models

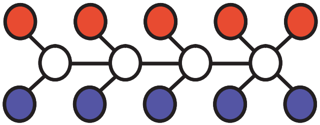
## — Fall Quarter, Lecture 11 —

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 5th, 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

# 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
- L13 (11/10):
- 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.

# Approximation: Two general approaches

- exact solution to approximate problem - approximate problem
  1. 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
  2. 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
  1. Message or other form of propagation, variational approaches, LP relaxations, loopy belief propagation (LBP)
  2. 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

$$\mu_{i\to j}(x_j) = \sum_{x_i} \psi_{i,j}(x_i, x_j) \prod_{k\in\delta(i)\setminus\{j\}} \mu_{k\to i}(x_i) \tag{11.5}$$

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

$$p(x_i) \propto \prod_{j\in\delta(i)} \mu_{j\to i}(x_i) \tag{11.6}$$

and

$$p(x_i, x_j) \propto \psi_{i,j}(x_i, x_j) \prod_{k\in\delta(i)\setminus\{j\}} \mu_{k\to i}(x_i) \prod_{\ell\in\delta(j)\setminus\{i\}} \mu_{\ell\to j}(x_j) M \tag{11.7}$$

# Choices for dealing with higher order factors in MRFs

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

1. 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.
2. Alternatively, we can define BP on factor graphs.
3. 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 $\{\mu_{i\to j}(x_j)\}_{i,j}$ 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

$$\mu_{i\to j}(x_j) \propto \sum_{x_i} \psi_{i,j}(x_i,x_j)\psi_i(x_i) \prod_{k\in\delta(i)\setminus\{j\}} \mu_{k\to i}(x_i) \quad (11.9)$$

$$= \sum_{x_i} \psi'_{i,j}(x_i,x_j)\mu_{\neg j\to i}(x_i) \quad (11.10)$$

$$= (\psi'_{i,j})^T \mu_{\neg j\to i} \quad (11.11)$$

- Here, $\psi'_{i,j}$ is a matrix and $\mu_{\neg j\to i}$ 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 $\mu_{\neg j\to i}$ 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.

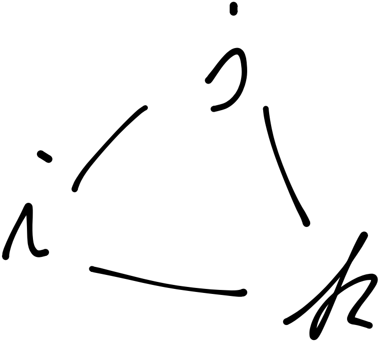
# Belief Propagation and Cycles

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., $\mu_{i \to j}(x_j) = 1$ 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.

# Belief Propagation, Cycles, and Oscillation

- Consider odd length cycle (e.g., $C_3$, $C_5$, etc.), $C_3$ is sufficient $i$—$j$—$k$—$i$



# Belief Propagation, Cycles, and Oscillation

- Consider odd length cycle (e.g., $C_3$, $C_5$, etc.), $C_3$ is sufficient
  $i$—$j$—$k$—$i$
- Assume all messages start out at state $\mu_{i \to j} = [1, 0]^T$.

# Belief Propagation, Cycles, and Oscillation

- Consider odd length cycle (e.g., $C_3$, $C_5$, etc.), $C_3$ is sufficient $i$—$j$—$k$—$i$
- Assume all messages start out at state $\mu_{i\to j} = [1,0]^T$.
- Consider (pairwise) edge functions, for each $i, j$

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

# Belief Propagation, Cycles, and Oscillation

- Consider odd length cycle (e.g., $C_3$, $C_5$, etc.), $C_3$ is sufficient $i$—$j$—$k$—$i$
- Assume all messages start out at state $\mu_{i \to j} = [1, 0]^T$.
- Consider (pairwise) edge functions, for each $i, j$

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

- then we have

$$\mu_{j \to k}(x_k) = \sum_{x_j} \psi_{j,k}(x_j, x_k) \mu_{i \to j}(x_j) \tag{11.2}$$

# Belief Propagation, Cycles, and Oscillation

- Consider odd length cycle (e.g., $C_3$, $C_5$, etc.), $C_3$ is sufficient $i$—$j$—$k$—$i$
- Assume all messages start out at state $\mu_{i\to j} = [1, 0]^T$.
- Consider (pairwise) edge functions, for each $i, j$

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

- then we have

$$\mu_{j\to k}(x_k) = \sum_{x_j} \psi_{j,k}(x_j, x_k)\mu_{i\to j}(x_j) \tag{11.2}$$

- or in matrix form

$$\mu_{j\to k} = (\psi_{j,k})^T \mu_{i\to j} \tag{11.3}$$

# Belief Propagation, Cycles, and Oscillation

- Let $\mu_{i \to j}^t$ be the $t^{\text{th}}$ formed message, with $\mu_{i \to j}^0$ being the starting state at $[1, 0]^T$.

# Belief Propagation, Cycles, and Oscillation

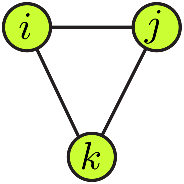
- Let $\mu^t_{i \to j}$ be the $t^{\text{th}}$ formed message, with $\mu^0_{i \to j}$ being the starting state at $[1, 0]^T$.
- Then $\mu^1_{i \to j} = [0, 1]^T$, $\mu^2_{i \to j} = [1, 0]^T$, $\mu^3_{i \to j} = [0, 1]^T$, and so on, never converging. In fact,

# Belief Propagation, Cycles, and Oscillation

- Let $\mu^t_{i\to j}$ be the $t^{\text{th}}$ formed message, with $\mu^0_{i\to j}$ being the starting state at $[1,0]^T$.
- Then $\mu^1_{i\to j} = [0,1]^T$, $\mu^2_{i\to j} = [1,0]^T$, $\mu^3_{i\to j} = [0,1]^T$, and so on, never converging. In fact,

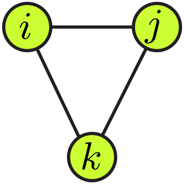
# Belief Propagation, Cycles, and Oscillation

- Let $\mu^t_{i\to j}$ be the $t^{\text{th}}$ formed message, with $\mu^0_{i\to j}$ being the starting state at $[1,0]^T$.
- Then $\mu^1_{i\to j} = [0,1]^T$, $\mu^2_{i\to j} = [1,0]^T$, $\mu^3_{i\to j} = [0,1]^T$, and so on, never converging. In fact,



# Belief Propagation, Cycles, and Oscillation

- Let $\mu^t_{i\to j}$ be the $t^{\text{th}}$ formed message, with $\mu^0_{i\to j}$ being the starting state at $[1,0]^T$.
- Then $\mu^1_{i\to j} = [0,1]^T$, $\mu^2_{i\to j} = [1,0]^T$, $\mu^3_{i\to j} = [0,1]^T$, and so on, never converging. In fact,




$$
\begin{aligned}
\mu^{t+1}_{i\to j} &= (\psi_{i,j})^T \mu^t_{k\to i} && (11.4)\\
&= (\psi_{i,j})^T (\psi_{k,i})^T \mu^t_{j\to k} && (11.5)\\
&= (\psi_{i,j})^T (\psi_{k,i})^T (\psi_{j,k})^T \mu^t_{i\to j} && (11.6)\\
&= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}^3 \mu^t_{i\to j} && (11.7)\\
&= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \mu^t_{i\to j} && (11.8)
\end{aligned}
$$

# Belief Propagation, Cycles, and Oscillation

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

# Belief Propagation, Cycles, and Oscillation

- 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

$$\mu_{i \to j}^{t} \leftarrow \gamma \mu_{i \to j}^{t} + (1 - \gamma) \mu_{i \to j}^{t-1} \tag{11.9}$$

# Belief Propagation, Cycles, and Oscillation

- 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

$$\mu^t_{i \to j} \leftarrow \gamma \mu^t_{i \to j} + (1 - \gamma)\mu^{t-1}_{i \to j} \tag{11.9}$$

- 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.

# Belief Propagation, Cycles, and Oscillation

- 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

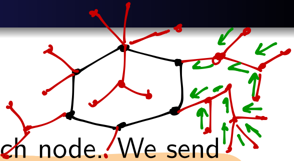
$$\mu^t_{i \to j} \leftarrow \gamma \mu^t_{i \to j} + (1 - \gamma) \mu^{t-1}_{i \to j} \tag{11.9}$$

- 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 $\mu^0_{i\to j} = [0.5, 0.5]^T$ then $\mu^{t+1}_{i\to j} = \mu^t_{i\to j}$.
- 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 $\mu^t_{i\to j}$ as $t$ increases. w.l.o.g. consider $\mu^t_{\ell\to 1}$
- Let the cycle be nodes $(1, 2, 3, \ldots, \ell, 1)$


$$\mu^{t+1}_{\ell\to 1} = \left(\prod_{i=1}^{\ell-1} (\psi_{i,i+1})^T\right) \mu^t_{\ell\to 1} \tag{11.10}$$

$$= M \mu^t_{\ell\to 1} \tag{11.11}$$

- 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.** *$\mu_{\ell \to 1}$ 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 $\beta$ 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 *guaranteed* property of convergence since we don't have that $|\lambda_1| > |\lambda_2|$.

# Belief Propagation, arbitrary graph

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

# Graphical Models, Exponential Families, and Variational Inference

- We're going to start covering our book:
  Wainwright and Jordan *Graphical Models, Exponential Families, and Variational Inference* http://www.nowpublishers.com/product.aspx?product=MAL&doi=2200000001
- 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 conficts 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 $d = |\mathcal{I}|$.

# 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 $C_\alpha \subseteq V(G)$.

$m = |V(G)|$ $x \in \mathbb{R}^m$

$$\rho(x_1, x_2, x_3) = \frac{1}{Z} \psi_{12}(x_1, x_2) \psi_{23}(x_2, x_3) \psi_{31}(x_3, x_1)$$

$$\rho(x_1, x_2) = \sum_{x_3} \rho(x_1, x_2, x_3)$$

# 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 $d = |\mathcal{I}|$.

# 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

$$p_\theta(x) = \exp\left[\overbrace{\langle \theta, \phi(x) \rangle}^{\sum_{i \in \mathcal{I}} \theta_i \phi_i(x)} - A(\theta)\right] \tag{11.12}$$

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

# 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

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

  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 $x$. $\phi : \mathsf{D}_X{}^m \to \mathbb{R}^d$.

# exponential family models and clique features

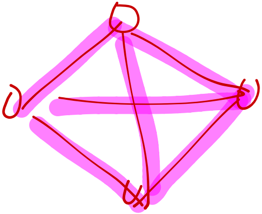
- Given a graph $G = (V, E)$ we have a set of cliques $\mathcal{C}$ of the graph.

# exponential family models and clique features

- Given a graph $G = (V, E)$ we have a set of cliques $\mathcal{C}$ of the graph.
- In order to respect the graph, we have to make sure that $\alpha \in \mathcal{I}$ respects the cliques.

## exponential family models and clique features

- Given a graph $G = (V, E)$ we have a set of cliques $\mathcal{C}$ 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 $C_\alpha \subseteq C$.



# exponential family models and clique features

- Given a graph $G = (V, E)$ we have a set of cliques $\mathcal{C}$ 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.

# exponential family models and clique features

- Given a graph $G = (V, E)$ we have a set of cliques $\mathcal{C}$ 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 $C \in \mathcal{C}$.

# exponential family models and clique features

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

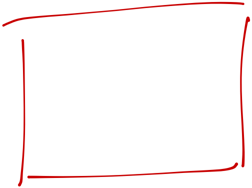
$$C_\alpha = C_{\alpha'} \quad \forall \alpha, \alpha' \in \mathcal{I}$$

# exponential family models and clique features

- Example: single scalar discrete random variable $X \in \{1, 2, \ldots, k\}$ might have indicator feature for all possible values $\alpha_i(x) \triangleq \mathbf{1}(x = i)$ — in this case $|C_\alpha| = 1$ for all $\alpha \in \mathcal{I}$.
- Could even think of $\{C_\alpha\}_{\alpha \in \mathcal{I}}$ as cliques of some graph, but not necessarily maxcliques.

# exponential family models and clique features

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



# exponential family models and clique features

- Example: single scalar discrete random variable $X \in \{1, 2, \ldots, k\}$ might have indicator feature for all possible values $\alpha_i(x) \triangleq \mathbf{1}(x = i)$ — in this case $|C_\alpha| = 1$ for all $\alpha \in \mathcal{I}$.
- 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: $p \in \mathcal{F}(G, \mathcal{M}^{(\mathrm{f})})$ by Hammersley-Clifford theorem,

# exponential family models and clique features

- Example: single scalar discrete random variable $X \in \{1, 2, \dots, k\}$ might have indicator feature for all possible values $\alpha_i(x) \triangleq \mathbf{1}(x = i)$ — in this case $|C_\alpha| = 1$ for all $\alpha \in \mathcal{I}$.
- 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: $p \in \mathcal{F}(G, \mathcal{M}^{(\mathrm{f})})$ by Hammersley-Clifford theorem,
  - where $G = (V, E)$ where $V$ is the nodes corresponding to vector $x$,

# exponential family models and clique features

- Example: single scalar discrete random variable $X \in \{1, 2, \ldots, k\}$ might have indicator feature for all possible values $\alpha_i(x) \triangleq \mathbf{1}(x = i)$ — in this case $|C_\alpha| = 1$ for all $\alpha \in \mathcal{I}$.
- 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: $p \in \mathcal{F}(G, \mathcal{M}^{(\mathsf{f})})$ 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: $\exists$ an $\alpha \in \mathcal{I}$ such that $u, v \in C_\alpha$ where $u, v \in V(G) \Leftrightarrow$ 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 $p \in \mathcal{F}((V, E), \mathcal{M}^{(\mathsf{f})})$ implies $p \in \mathcal{F}((V, E + E_1), \mathcal{M}^{(\mathsf{f})})$ 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) $p(x) = \left[\prod_{\alpha} \exp(\phi_\alpha(x))\right] \exp(A(\theta))$
- Exponential models can represent any factorization, given any factorization in terms ~~of $\phi$, we can do $\exp(\log \phi)$ to get~~ potentials.
- We can often make them log-linear models as well with the right potential functions which won't increase tree-width of the graph.
- Moreover, exponential family models are incredibly flexible and have a number of desirable properties (e.g., aspects of the log partition function which we will see)

# 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.

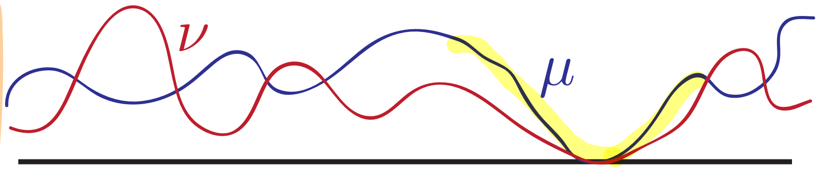
$$\sum_{i \in \text{Dom}} f(x_i)$$

# 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. $\nu$

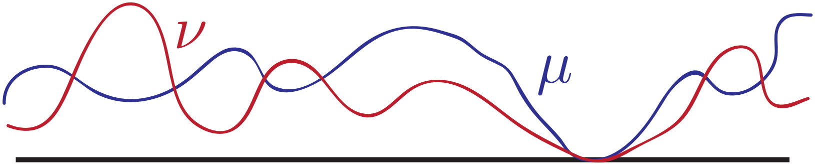
# absolutely continuous

- 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 $\nu \ll \mu$.



## 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

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

$$A(\theta) = \log Z(\theta)$$

# exponential family models

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

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

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

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

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

# exponential family models

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

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

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

$$A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \quad (11.14)$$

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 models

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

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

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

$$A(\theta) = \log \int_{\mathsf{D}_X} \exp \left( \langle \theta, \phi(x) \rangle \right) \nu(dx) \quad (11.14)$$

  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

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

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

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

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

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

where $\forall \alpha \in \mathcal{I}$

$$\mathbb{E}_p[\phi_\alpha(X)] = \int_{\mathsf{D}_X} \phi_\alpha(x) p(x) \nu(dx) \tag{11.18}$$

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

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

# Minimal Representation of Exponential Family

- Solution as form:

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{11.20}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{11.21}$$

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

# Minimal Representation of Exponential Family

- Solution as form:

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{11.20}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{11.21}$$

  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).

# Minimal Representation of Exponential Family

- Solution as form:

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{11.20}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{11.21}$$

  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 $\langle \gamma, \phi(x_1) \rangle \neq \langle \gamma, \phi(x_2) \rangle$.

# Minimal Representation of Exponential Family

- Solution as form:

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{11.20}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{11.21}$$

  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 $\gamma \in \mathbb{R}^D$, 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).

# Minimal Representation of Exponential Family

- Solution as form:

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{11.20}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{11.21}$$

  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 $\gamma \in \mathbb{R}^D$, 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

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{11.22}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{11.23}$$

- Overcomplete representation $d = |\mathcal{I}|$ higher than need be

# Overcomplete Representation

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{11.22}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{11.23}$$

- 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.

# Overcomplete Representation

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{11.22}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{11.23}$$

- 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

$$p_{\theta+\gamma}(x) = \exp(\langle (\theta + \gamma), \phi(x) \rangle - A(\theta + \gamma)) \tag{11.24}$$

$$= \exp(\langle \theta, \phi(x) \rangle + \langle \gamma, \phi(x) \rangle - A(\theta + \gamma)) \tag{11.25}$$

$$= \exp(\langle \theta, \phi(x) \rangle + c - A(\theta + \gamma)) \tag{11.26}$$

$$= \exp(\langle \theta, \phi(x) \rangle - A(\theta)) = p_\theta(x) \tag{11.27}$$

# Overcomplete Representation

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{11.22}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{11.23}$$

- 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

$$p_{\theta+\gamma}(x) = \exp(\langle (\theta + \gamma), \phi(x) \rangle - A(\theta + \gamma)) \tag{11.24}$$

$$= \exp(\langle \theta, \phi(x) \rangle + \langle \gamma, \phi(x) \rangle - A(\theta + \gamma)) \tag{11.25}$$

$$= \exp(\langle \theta, \phi(x) \rangle + c - A(\theta + \gamma)) \tag{11.26}$$

$$= \exp(\langle \theta, \phi(x) \rangle - A(\theta)) = p_\theta(x) \tag{11.27}$$

- True for any $\lambda\gamma$ with $\lambda \in \mathbb{R}$, so affine set of identical distributions!

# Overcomplete Representation

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{11.22}$$

$$\text{where } A(\theta) = \log \int_{\mathsf{D}_X} \exp\left(\langle \theta, \phi(x) \rangle\right) \nu(dx) \tag{11.23}$$

- 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

$$p_{\theta+\gamma}(x) = \exp(\langle (\theta + \gamma), \phi(x) \rangle - A(\theta + \gamma)) \tag{11.24}$$

$$= \exp(\langle \theta, \phi(x) \rangle + \langle \gamma, \phi(x) \rangle - A(\theta + \gamma)) \tag{11.25}$$

$$= \exp(\langle \theta, \phi(x) \rangle + c - A(\theta + \gamma)) \tag{11.26}$$

$$= \exp(\langle \theta, \phi(x) \rangle - A(\theta)) = p_\theta(x) \tag{11.27}$$

- 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.

# Exponential family models

- Minimal representation of Bernoulli distribution is

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{11.28}$$

# Exponential family models

- Minimal representation of Bernoulli distribution is

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{11.28}$$

- overcomplete rep of Bernoulli dist.

$$p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \tag{11.29}$$

$$= \exp(\theta_0(1-x) + \theta_1 x - A(\gamma)) \tag{11.30}$$

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

# Exponential family models

- Minimal representation of Bernoulli distribution is

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{11.28}$$

- overcomplete rep of Bernoulli dist.

$$p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \tag{11.29}$$
$$= \exp(\theta_0(1 - x) + \theta_1 x - A(\gamma)) \tag{11.30}$$

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

- Is there a vector $a$ s.t. $\langle a, \phi(x) \rangle = c$ for all $x$, $\nu$-a.e.?

# Exponential family models

- Minimal representation of Bernoulli distribution is

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{11.28}$$

- overcomplete rep of Bernoulli dist.

$$p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \tag{11.29}$$

$$= \exp(\theta_0(1-x) + \theta_1 x - A(\gamma)) \tag{11.30}$$

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

- Is there a vector $a$ s.t. $\langle a, \phi(x) \rangle = c$ for all $x$, $\nu$-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

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{11.28}$$

- overcomplete rep of Bernoulli dist.

$$p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \tag{11.29}$$
$$= \exp(\theta_0(1 - x) + \theta_1 x - A(\gamma)) \tag{11.30}$$

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

$$p(x|\gamma) = \exp(\gamma x - A(\gamma)) \tag{11.28}$$

- overcomplete rep of Bernoulli dist.

$$\begin{aligned} p(x|\theta_0, \theta_1) &= \exp(\langle \theta, \phi(x) \rangle) & (11.29) \\ &= \exp(\theta_0(1-x) + \theta_1 x - A(\gamma)) & (11.30) \end{aligned}$$

  where $\theta = (\theta_0, \theta_1)$ and $\phi(x) = (1-x, x)$.
- Is there a vector $a$ s.t. $\langle a, \phi(x) \rangle = c$ for all $x$, $\nu$-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 of form $\theta_1 - \theta_0 = \gamma$ gives same distribution.

# Famous Example - Ising Model

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

$$p_\theta(x) = \exp\left\{ \sum_{v \in V} \theta_v x_v + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \right\}, \tag{11.31}$$

with

$$A(\theta) = \log \sum_{x \in \{0,1\}^m} \exp\left\{ \sum_{v \in V} \theta_v x_v + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \right\} \tag{11.32}$$

- 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 $\{C_\alpha\}_\alpha = \Big\{\{x_i\}_i, \{x_i x_j\}_{(i,j)\in E}\Big\}$.
- We can easily generalize this via a set system. I.e., consider $(V, \mathcal{V})$, where $\mathcal{V} = \big\{V_1, V_2, \dots, V_{|\mathcal{V}|}\big\}$ and where $\forall i, V_i \subseteq V$.
- We can form sufficient statistic set via $\{C_\alpha\}_\alpha = \big\{\{x_V\}_{V\in\mathcal{V}}\big\}$.
- Higher order factors/interaction functions/potential functions/sufficient statistics.

# Multivalued variables

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

$$\mathbf{1}_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{else} \end{cases} \tag{11.33}$$

and

$$\mathbf{1}_{st;jk}(x_s, x_t) = \begin{cases} 1 & \text{if } x_s = j \text{ and } x_t = k, \\ 0 & \text{else} \end{cases} \tag{11.34}$$

- Model becomes

$$p_\theta(x) = \exp\left\{ \sum_{v \in V} \sum_{i=0}^{r-1} \theta_{v;j} \mathbf{1}_{s;j}(x_v) + \sum_{(s,t) \in E} \sum_{j,k} \theta_{st;ij} \mathbf{1}_{st;jk}(x_s, x_t) - A(\theta) \right. \tag{11.35}$$

- Is this overcomplete?

# Multivalued variables

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

$$\mathbf{1}_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{else} \end{cases} \tag{11.33}$$

and

$$\mathbf{1}_{st;jk}(x_s, x_t) = \begin{cases} 1 & \text{if } x_s = j \text{ and } x_t = k, \\ 0 & \text{else} \end{cases} \tag{11.34}$$

- Model becomes

$$p_\theta(x) = \exp\left\{ \sum_{v \in V} \sum_{i=0}^{r-1} \theta_{v;j} \mathbf{1}_{s;j}(x_v) + \sum_{(s,t) \in E} \sum_{j,k} \theta_{st;ij} \mathbf{1}_{st;jk}(x_s, x_t) - A(\theta) \right. \tag{11.35}$$

- Is this overcomplete? Yes. Why?

# Multivariate Gaussian

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

$$p_\theta(x) = \exp\left\{ \langle \theta, x \rangle + \frac{1}{2} \langle\langle \Theta, xx^\intercal \rangle\rangle - A(\theta, \Theta) \right\} \tag{11.36}$$

# Multivariate Gaussian

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

$$p_\theta(x) = \exp\left\{ \langle \theta, x \rangle + \frac{1}{2} \langle\langle \Theta, xx^\intercal \rangle\rangle - A(\theta, \Theta) \right\} \tag{11.36}$$

- $\langle\langle \Theta, xx^\intercal \rangle\rangle = \sum_{ij} \Theta_{ij} x_i x_j$ is Frobenius norm.

# Multivariate Gaussian

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

$$p_\theta(x) = \exp\left\{ \langle \theta, x \rangle + \frac{1}{2} \langle\langle \Theta, xx^\intercal \rangle\rangle - A(\theta, \Theta) \right\} \tag{11.36}$$

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

# Multivariate Gaussian

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

$$p_\theta(x) = \exp\left\{ \langle \theta, x \rangle + \frac{1}{2} \langle\langle \Theta, xx^\intercal \rangle\rangle - A(\theta, \Theta) \right\} \tag{11.36}$$

- $\langle\langle \Theta, xx^\intercal \rangle\rangle = \sum_{ij} \Theta_{ij} x_i x_j$ is Frobenius norm.
- 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).

# Multivariate Gaussian

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

$$p_\theta(x) = \exp\left\{ \langle \theta, x \rangle + \frac{1}{2} \langle\!\langle \Theta, xx^\intercal \rangle\!\rangle - A(\theta, \Theta) \right\} \tag{11.36}$$

- $\langle\!\langle \Theta, xx^\intercal \rangle\!\rangle = \sum_{ij} \Theta_{ij} x_i x_j$ is Frobenius norm.
- 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).
- Any other constraints on $\Theta$?

# Multivariate Gaussian

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

$$p_\theta(x) = \exp\left\{ \langle \theta, x \rangle + \frac{1}{2} \langle\langle \Theta, x x^\intercal \rangle\rangle - A(\theta, \Theta) \right\} \tag{11.36}$$

- $\langle\langle \Theta, x x^\intercal \rangle\rangle = \sum_{ij} \Theta_{ij} x_i x_j$ is Frobenius norm.
- 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).
- Any other constraints on $\Theta$? negative definite

# Multivariate Gaussian

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

$$p_\theta(x) = \exp\left\{ \langle \theta, x \rangle + \frac{1}{2} \langle\!\langle \Theta, xx^\intercal \rangle\!\rangle - A(\theta, \Theta) \right\} \tag{11.36}$$

- $\langle\!\langle \Theta, xx^\intercal \rangle\!\rangle = \sum_{ij} \Theta_{ij} x_i x_j$ is Frobenius norm.
- 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).
- Any other constraints on $\Theta$? 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 - key thing is to place the hard constraints in the $\nu$ 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 $\mu_\alpha$ associated with statistic $\phi_\alpha$ defined as:

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x)p(x)\nu(dx) \tag{11.37}$$

# Mean Parameters, Convex Cores

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

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x)p(x)\nu(dx) \tag{11.37}$$

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

# Mean Parameters, Convex Cores

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

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \nu(dx) \tag{11.37}$$

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

$$\mathcal{M}(\phi) = \mathcal{M} \triangleq \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)],\ \forall \alpha \in \mathcal{I} \right\} \tag{11.38}$$

# Mean Parameters, Convex Cores

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

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \nu(dx) \tag{11.37}$$

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

$$\mathcal{M}(\phi) = \mathcal{M} \triangleq \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)],\ \forall \alpha \in \mathcal{I} \right\} \tag{11.38}$$

- We don't say $p$ was necessarily exponential family

# Mean Parameters, Convex Cores

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

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \nu(dx) \tag{11.37}$$

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

$$\mathcal{M}(\phi) = \mathcal{M} \triangleq \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)],\ \forall \alpha \in \mathcal{I} \right\} \tag{11.38}$$

- We don't say $p$ was necessarily exponential family
- $\mathcal{M}$ is convex since expected value is a linear operator. So convex combinations of $p$ and $p'$ will lead to convex combinations of $\mu$ and $\mu'$

# Mean Parameters, Convex Cores

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

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x)p(x)\nu(dx) \tag{11.37}$$

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

$$\mathcal{M}(\phi) = \mathcal{M} \triangleq \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)],\ \forall \alpha \in \mathcal{I} \right\} \tag{11.38}$$

- We don't say $p$ was necessarily exponential family
- $\mathcal{M}$ is convex since expected value is a linear operator. So convex combinations of $p$ and $p'$ will lead to convex combinations of $\mu$ and $\mu'$
- $\mathcal{M}$ is like a "convex core" of all distributions expressed via $\phi$.

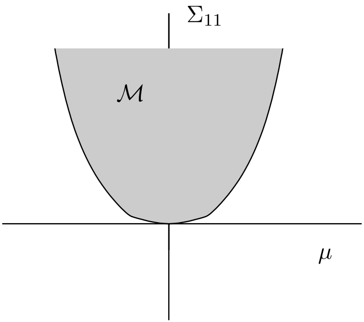
# Mean Parameters and Gaussians

- 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 $\mu$, 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

$$\mathcal{M} = \left\{(\mu, C) \in \mathbb{R}^m \times \mathcal{S}_+^m | C - \mu\mu^\intercal \succeq 0\right\} \tag{11.39}$$

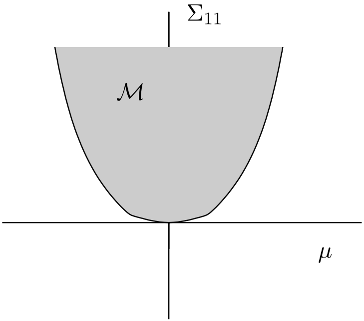
where $\mathcal{S}_+^m$ is the set of symmetric positive semi-definite matrices.

# 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."

# 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."

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

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^b : \mu = \sum_x \phi(x)p(x) \text{ for some } p \in \mathcal{U} \right\} \tag{11.40}$$

$$= \operatorname{conv}\{\phi(x), x \in \mathsf{D}_X \text{ (that are } \nu\text{-measurable)},\} \tag{11.41}$$

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

# Mean Parameters and Polytopes

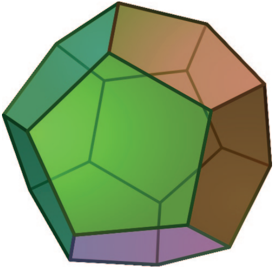
- When $X$ is discrete, we get a polytope since

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^b : \mu = \sum_x \phi(x)p(x) \text{ for some } p \in \mathcal{U} \right\} \tag{11.40}$$

$$= \operatorname{conv}\{\phi(x), x \in \mathsf{D}_X \text{ (that are } \nu\text{-measurable),}\} \tag{11.41}$$

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

- So we have a convex polytope

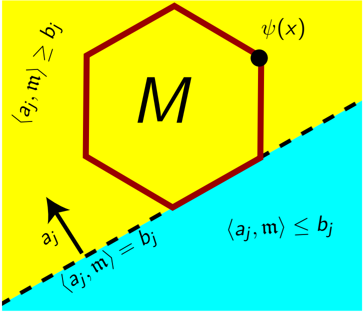


# Mean Parameters and Polytopes

- 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

$$M = \left\{ \mu \in \mathbb{R}^d : A\mu \geq b \right\} \quad = \left\{ \mu \in \mathbb{R}^d : \langle a_j, \mu \rangle \geq b_j, \forall j \in J \right\} \tag{11.42}$$

with $A$ having rows $a_j$.



# Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

$$\phi(x) = \{x_s, s \in V; x_s x_t, (s,t) \in E(G)\} \in \mathbb{R}^{|V|+|E|} \tag{11.43}$$

we get

$$\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \;\; \forall v \in V \tag{11.44}$$

$$\mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \;\; \forall (s,t) \in E(G) \tag{11.45}$$

# Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

$$\phi(x) = \{x_s, s \in V; x_s x_t, (s,t) \in E(G)\} \in \mathbb{R}^{|V|+|E|} \tag{11.43}$$

we get

$$\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \;\; \forall v \in V \tag{11.44}$$

$$\mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \;\; \forall (s,t) \in E(G) \tag{11.45}$$

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

# Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

$$\phi(x) = \{x_s, s \in V; x_s x_t, (s,t) \in E(G)\} \in \mathbb{R}^{|V|+|E|} \tag{11.43}$$

we get

$$\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \;\; \forall v \in V \tag{11.44}$$

$$\mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \;\; \forall (s,t) \in E(G) \tag{11.45}$$

- 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

$$\phi(x) = \{x_s, s \in V; x_s x_t, (s,t) \in E(G)\} \in \mathbb{R}^{|V|+|E|} \tag{11.43}$$

we get

$$\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \;\; \forall v \in V \tag{11.44}$$

$$\mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \;\; \forall (s,t) \in E(G) \tag{11.45}$$

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

$$\phi(x) = \{x_s, s \in V; x_s x_t, (s,t) \in E(G)\} \in \mathbb{R}^{|V|+|E|} \tag{11.43}$$

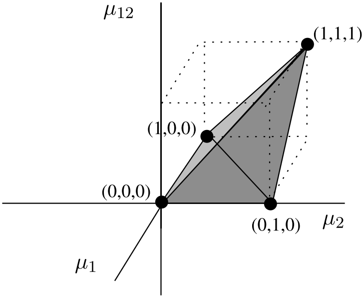
  we get

$$\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \;\; \forall v \in V \tag{11.44}$$

$$\mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \;\; \forall (s,t) \in E(G) \tag{11.45}$$

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

$$\forall v \in V(G), j \in \{0 \ldots r-1\}, \text{ define } \phi_{v,j}(x_v) \triangleq \mathbf{1}(x_v = j) \tag{11.46}$$

$$\forall (s,t) \in E(G), j,k \in \{0 \ldots r-1\}, \text{ we define:} \tag{11.47}$$

$$\phi_{st,jk}(x_s, x_t) \triangleq \mathbf{1}(x_s = j, x_t = k) = \mathbf{1}(x_s = j)\mathbf{1}(x_t = k) \tag{11.48}$$

- 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

$$\mu_{v,j} = \mathbb{E}_p[\mathbf{1}(x_v = j)] = p(x_v = j) \tag{11.49}$$

$$\mu_{st,jk} = \mathbb{E}_p[\mathbf{1}(x_s = j, x_t = k)] = p(x_s = j, x_t = k) \tag{11.50}$$

- Such an $\mathcal{M}$ is called the *marginal polytope*. Any $\mu$ must live in the polytope that corresponds to node and edge true marginals!!
- We can also associate such a polytope with a graph $G$, where we take only $(s,t) \in E(G)$. Denote this as $\mathbb{M}(G)$.
- This polytope can help us to characterize when BP converges (there might be an outer bound of this polytope), or it might characterize the result of a mean-field approximation (an inner bound of this polytope) as we'll see.

# 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.

# 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.
- 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 $\theta$ 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 $\theta$ to the point in the marginal polytope, called forward mapping, moving from $\theta \in \Omega$ to $\mu \in \mathcal{M}$.
- We can view the (maximum likelihood) learning problem as moving from a point in the polytope (given by the empirical distribution) to the canonical parameters. Called backwards mapping

# Learning is the dual of Inference

- We can view the inference problem as moving from the canonical parameters $\theta$ to the point in the marginal polytope, called forward mapping, moving from $\theta \in \Omega$ to $\mu \in \mathcal{M}$.
- We can view the (maximum likelihood) learning problem as moving from a point in the polytope (given by the empirical distribution) to the canonical parameters. Called backwards mapping
- graph structure (e.g., tree-width) makes this easy or hard, and also characterizes the polytope (how complex it is in terms of number of faces).

# Learning is the dual of Inference

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

$$\ell(\theta, \mathbf{D}) = \frac{1}{M} \sum_{i=1}^{M} \log p_\theta(\bar{x}^{(i)}) = \langle \theta, \hat{\mu} \rangle - A(\theta) \tag{11.51}$$

where empirical means given by

$$\hat{\mu} = \hat{\mathbb{E}}[\phi(X)] = \frac{1}{M} \sum_{i=1}^{M} \phi(\bar{x}^{(i)}) \tag{11.52}$$

# Learning is the dual of Inference

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

$$\ell(\theta, \mathbf{D}) = \frac{1}{M} \sum_{i=1}^{M} \log p_\theta(\bar{x}^{(i)}) = \langle \theta, \hat{\mu} \rangle - A(\theta) \tag{11.51}$$

where empirical means given by

$$\hat{\mu} = \hat{\mathbb{E}}[\phi(X)] = \frac{1}{M} \sum_{i=1}^{M} \phi(\bar{x}^{(i)}) \tag{11.52}$$

- By taking derivatives of the above, it is easy to see that solution is the point $\hat{\theta}$ such that (empirical matches expected means)

$$\mathbb{E}_{\hat{\theta}}[\phi(X)] = \hat{\mu} \tag{11.53}$$

this is the the *backward mapping problem*, going from $\mu$ to $\theta$.

# Learning is the dual of Inference

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

$$\ell(\theta, \mathbf{D}) = \frac{1}{M} \sum_{i=1}^{M} \log p_\theta(\bar{x}^{(i)}) = \langle \theta, \hat{\mu} \rangle - A(\theta) \tag{11.51}$$

where empirical means given by

$$\hat{\mu} = \hat{\mathbb{E}}[\phi(X)] = \frac{1}{M} \sum_{i=1}^{M} \phi(\bar{x}^{(i)}) \tag{11.52}$$

- By taking derivatives of the above, it is easy to see that solution is the point $\hat{\theta}$ such that (empirical matches expected means)

$$\mathbb{E}_{\hat{\theta}}[\phi(X)] = \hat{\mu} \tag{11.53}$$

this is the the *backward mapping problem*, going from $\mu$ to $\theta$.

- This is identical to the maximum entropy problem.

# 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

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \quad (11.54)$$

- 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

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{11.54}$$

- If we know the log partition function, we know a lot for an exponential family model. In particular, we know
- $A(\theta)$ is convex in $\theta$ (strictly so if minimal representation).

# Log partition (or cumulant) function

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{11.54}$$

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

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] = \int \phi_\alpha(X) p_\theta(x) \nu(dx) = \mu_\alpha \tag{11.55}$$

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

# Log partition (or cumulant) function

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{11.54}$$

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

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] = \int \phi_\alpha(X) p_\theta(x) \nu(dx) = \mu_\alpha \tag{11.55}$$

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

$$\frac{\partial^2 A}{\partial \theta_{\alpha_1} \partial \theta_{\alpha_2}}(\theta) = \mathbb{E}_\theta[\phi_{\alpha_1}(X)\phi_{\alpha_2}(X)] - \mathbb{E}_\theta[\phi_{\alpha_1}(X)]\mathbb{E}_\theta[\phi_{\alpha_2}(X)] \tag{11.56}$$

# Log partition (or cumulant) function

$$A(\theta) = \log \int_{\mathsf{D}_X} \langle \theta, \phi(x) \rangle \, \nu(dx) \tag{11.54}$$

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

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] = \int \phi_\alpha(X) p_\theta(x) \nu(dx) = \mu_\alpha \tag{11.55}$$

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

$$\frac{\partial^2 A}{\partial \theta_{\alpha_1} \partial \theta_{\alpha_2}}(\theta) = \mathbb{E}_\theta[\phi_{\alpha_1}(X)\phi_{\alpha_2}(X)] - \mathbb{E}_\theta[\phi_{\alpha_1}(X)]\mathbb{E}_\theta[\phi_{\alpha_2}(X)] \tag{11.56}$$

- Proof given in book.

# Log partition function

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

# Log partition function

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

# 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 $\mu$ and $\theta$.

# Log partition function

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\}$.
- For minimal exponential family models, this mapping is one-to-one, that is there is a unique pairing between $\mu$ and $\theta$.
- For non-minimal exponential families, more than one $\theta$ for a given $\mu$ (not surprising since multiple $\theta$'s can yield the same distribution).

# Log partition function

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\}$.
- For minimal exponential family models, this mapping is one-to-one, that is there is a unique pairing between $\mu$ and $\theta$.
- For non-minimal exponential families, more than one $\theta$ for a given $\mu$ (not surprising since multiple $\theta$'s can yield the same distribution).
- For non-exponential families, other distributions can yield $\mu$, but the exponential family one is the one that has maximum entropy.

# Log partition function

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\}$.
- For minimal exponential family models, this mapping is one-to-one, that is there is a unique pairing between $\mu$ and $\theta$.
- For non-minimal exponential families, more than one $\theta$ for a given $\mu$ (not surprising since multiple $\theta$'s can yield the same distribution).
- For non-exponential families, other distributions can yield $\mu$, but the exponential family one is the one that has maximum entropy. ex1: Gaussian, a distribution with maximum entropy amongst all other distributions with same mean and covariance.

# 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 $\mu$ and $\theta$.
- For non-minimal exponential families, more than one $\theta$ for a given $\mu$ (not surprising since multiple $\theta$'s can yield the same distribution).
- For non-exponential families, other distributions can yield $\mu$, but the exponential family one is the one that has maximum entropy. ex1: Gaussian, a distribution with maximum entropy amongst all other distributions with same mean and covariance. ex2: Consider the maximum entropy optimization problem, yields a distribution with exactly this property.

# Log partition function

- So $\nabla A : \Omega \to \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\}$.
- For minimal exponential family models, this mapping is one-to-one, that is there is a unique pairing between $\mu$ and $\theta$.
- For non-minimal exponential families, more than one $\theta$ for a given $\mu$ (not surprising since multiple $\theta$'s can yield the same distribution).
- For non-exponential families, other distributions can yield $\mu$, but the exponential family one is the one that has maximum entropy. ex1: Gaussian, a distribution with maximum entropy amongst all other distributions with same mean and covariance. ex2: Consider the maximum entropy optimization problem, yields a distribution with exactly this property.
- Key point: all mean parameters are realizable by member of exp. family.

# Mappings - one-to-one

In fact, we have

**Theorem 11.6.1**

*The gradient map $\nabla A$ is one-to-one iff the exponential representation is minimal.*

# Mappings - one-to-one

In fact, we have

**Theorem 11.6.1**

*The gradient map $\nabla A$ is one-to-one iff the exponential representation is minimal.*

- Proof basically uses property that if representation is non-minimal, and $\langle a, \phi(x) \rangle = c$ for all $x$, then we can form an affine set of equivalent parameters $\theta + \gamma a$.

## Mappings - one-to-one

In fact, we have

**Theorem 11.6.1**

*The gradient map $\nabla A$ is one-to-one iff the exponential representation is minimal.*

- Proof basically uses property that if representation is non-minimal, and $\langle a, \phi(x) \rangle = c$ for all $x$, then we can form an affine set of equivalent parameters $\theta + \gamma a$.
- Other direction, uses strict convexity.

# Mappings - onto

Moreover,

**Theorem 11.6.2**

*In a minimal exponential family, the gradient map $\nabla A$ is onto the interior of $\mathcal{M}$ (denoted $\mathcal{M}^\circ$). Consequently, for each $\mu \in \mathcal{M}^\circ$, there exists some $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.*

# Mappings - onto

Moreover,

**Theorem 11.6.2**

*In a minimal exponential family, the gradient map $\nabla A$ is onto the interior of $\mathcal{M}$ (denoted $\mathcal{M}^\circ$). Consequently, for each $\mu \in \mathcal{M}^\circ$, there exists some $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.*

- Example: consider, for example, a Gaussian.

# Mappings - onto

Moreover,

**Theorem 11.6.2**

*In a minimal exponential family, the gradient map $\nabla A$ is onto the interior of $\mathcal{M}$ (denoted $\mathcal{M}^\circ$). Consequently, for each $\mu \in \mathcal{M}^\circ$, there exists some $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.*

- Example: consider, for example, a 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

Moreover,

**Theorem 11.6.2**

*In a minimal exponential family, the gradient map $\nabla A$ is onto the interior of $\mathcal{M}$ (denoted $\mathcal{M}^\circ$). Consequently, for each $\mu \in \mathcal{M}^\circ$, there exists some $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.*

- Example: consider, for example, a 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 an exponential family model with those moments).

# Mappings - onto

Moreover,

**Theorem 11.6.2**

*In a minimal exponential family, the gradient map $\nabla A$ is onto the interior of $\mathcal{M}$ (denoted $\mathcal{M}^\circ$). Consequently, for each $\mu \in \mathcal{M}^\circ$, there exists some $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.*

- Example: consider, for example, a 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 an exponential family model with those moments).
- The theorem here is more general and applies for any set of sufficient statistics.

# Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \operatorname*{argmax}_{\theta} \left( \langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{11.57}$$

# Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \operatorname*{argmax}_{\theta} \left( \langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{11.57}$$

- Convex conjugate dual of $A(\theta)$ is defined as:

$$A^*(\mu) \triangleq \sup_{\theta \in \Omega} \left( \langle \theta, \mu \rangle - A(\theta) \right) \tag{11.58}$$

# Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \underset{\theta}{\operatorname{argmax}} \left( \langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{11.57}$$

- Convex conjugate dual of $A(\theta)$ is defined as:

$$A^*(\mu) \stackrel{\Delta}{=} \sup_{\theta \in \Omega} \left( \langle \theta, \mu \rangle - A(\theta) \right) \tag{11.58}$$

- So dual is optimal value of the ML problem, when $\mu \in \mathcal{M}$

# Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \underset{\theta}{\operatorname{argmax}} \left( \langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{11.57}$$

- Convex conjugate dual of $A(\theta)$ is defined as:

$$A^*(\mu) \triangleq \sup_{\theta \in \Omega} \left( \langle \theta, \mu \rangle - A(\theta) \right) \tag{11.58}$$

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

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu \tag{11.59}$$

# Conjugate Duality

- Consider maximum likelihood problem for exp. family

$$\theta^* \in \operatorname*{argmax}_{\theta} \left( \langle \theta, \hat{\mu} \rangle - A(\theta) \right) \tag{11.57}$$

- Convex conjugate dual of $A(\theta)$ is defined as:

$$A^*(\mu) \triangleq \sup_{\theta \in \Omega} \left( \langle \theta, \mu \rangle - A(\theta) \right) \tag{11.58}$$

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

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu \tag{11.59}$$

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

# Conjugate Duality

### Theorem 11.6.3 (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:*

$$A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \in \bar{\mathcal{M}} \end{cases} \tag{11.60}$$

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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \tag{11.61}$$

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

$$\mu = \int_{\mathsf{D}_X} \phi(x) p_\theta(x) \nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta) \tag{11.62}$$

# Conjugate Duality

- Note that $A*$ isn't exactly entropy, only entropy sometimes, and depends on matching parameters to $\mu$ via the matching mapping $\theta(\mu)$ which achieves

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \mu \tag{11.63}$$

# Conjugate Duality

- Note that $A*$ isn't exactly entropy, only entropy sometimes, and depends on matching parameters to $\mu$ via the matching mapping $\theta(\mu)$ which achieves

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \mu \tag{11.63}$$

- $A(\theta)$ in Equation 11.61 is the "inference" problem (dual of the dual) for a given $\theta$, 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 $\mu$ via the matching mapping $\theta(\mu)$ which achieves

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \mu \qquad (11.63)$$

- $A(\theta)$ in Equation 11.61 is the "inference" problem (dual of the dual) for a given $\theta$, since computing it involves computing the desired node/edge marginals.
- Whenever $\mu \notin \mathcal{M}$, then $A^*(\mu)$ returns $\infty$ which can't be the resulting sup, so Equation 11.61 need only consider $\mathcal{M}$.

# Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{11.61}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{11.61}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{11.61}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{11.61}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{11.61}$$

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

# Conjugate Duality

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{11.61}$$

- 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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \tag{11.61}$$

- 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). ☺☺

# 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`