**Outstanding Reading**

- Read chapters 1, 2, and 3 in the Wainwright and Jordan book.
We need to find one makeup lecture this term.

- L1 (9/28): Introduction, Families, Semantics
- LX (9/30): No class
- L2 (10/5): Trees, exact inference
- L3 (10/7): More on trees and inference.
- L4 (10/12): To tree or not to tree.
- L5 (10/14): All models lead to trees
- L6 (10/19): Decomposable, JT
- L7 (10/21): Inference on JTs
- L8 (10/26): JT Inference, semi-rings,
- L9 (10/28): time-space tradeoff, conditioning, LBP
- L10 (11/2): LBP, exp. f. models
- L11 (11/4): exp. f. models, marg poly
- L12 (11/9): LBP and Bethe Approximation
- LXX (11/11): Veterans Day, no class
- L13 (11/16): LBP and Bethe Approximation
- L14 (11/18): LBP and Bethe Approximation
- L15 (11/23): LBP and Bethe Approximation
- LXX (11/25): Thanksgiving, no class
- L16 (11/30): LBP and Bethe Approximation
- L17 (12/2): LBP and Bethe Approximation
- L18 (12/7): LBP and Bethe Approximation
- L19: (12/9):
Exponential family models are rich, and lead to interesting approximation algorithms and umbrella many approximate inference methods.

We’ll be following the Wainwright & Jordan book pretty closely for a while.
Approximation: Two general approaches

- exact solution to approximate problem - approximate problem
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
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.
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
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
Approximation: Two general approaches

1. 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.

2. approximate solution to exact problem - approximate inference
   - Message or other form of propagation, variational approaches, LP relaxations
   - sampling

Both methods only guaranteed approximate quality solutions. No longer in the achievable region in time-space tradeoff graph, new set of time/space tradeoffs to achieve a particular accuracy.
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
  2. sampling
  3. etc.
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
  2. sampling
  3. etc.

- Both methods only guaranteed approximate quality solutions.
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
  2. Sampling
  3. etc.

Both methods only guaranteed approximate quality solutions.

No longer in the achievable region in time-space tradeoff graph, new set of time/space tradeoffs to achieve a particular accuracy.
Graphical Models, Exponential Families, and Variational Inference

- We’re going to start covering our book.
- Today we will start on 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

\( \phi = (\phi_\alpha, \alpha \in I) \) is a collection of functions known as potential functions, sufficient statistics, or features. \( I \) is an index set of size \( d = |I| \).
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 $C_\alpha \subseteq V(G)$.
exponential family models

- $\phi = (\phi_\alpha, \alpha \in I)$ is a collection of functions known as potential functions, sufficient statistics, or features. $I$ is an index set of size $d = |I|$.
- 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, $|I|$). $\theta \in \Omega \subseteq \mathbb{R}^d$ where $d = |I|$.
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

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \quad (1)$$

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

\[
p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{1}
\]

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 \( x \). \( \phi : D^m_X \rightarrow \mathbb{R}^d \).
Given a graph $G = (V, E)$ we have a set of cliques $C$ of the graph.
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 $\alpha \in I$ 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 $\phi_{\alpha}(x_{\mathcal{C}_\alpha})$ there must be a clique $\mathcal{C} \in \mathcal{C}$ such that $\mathcal{C}_\alpha \subseteq \mathcal{C}$. 
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 $\phi_\alpha(x_{\mathcal{C}_\alpha})$ 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.
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 $X \in \{1, 2, \ldots, k\}$ might have indicator feature for all possible values $\alpha_i(x) \triangleq 1(x = i)$. 
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 $\alpha_i(x) \triangleq 1(x = i)$.

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

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

Likely not dealing with triangulated models. Could be based on cliques or subsets of cliques.
Examples: single scalar discrete random variable $X \in \{1, 2, \ldots, k\}$ might have indicator feature for all possible values $\alpha_i(x) \triangleq 1(x = i)$.

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

Likely not dealing with triangulated models. Could be based on cliques or subsets of cliques.

Key: $p \in \mathcal{F}(G, R^{(f)})$ by Hammersley-Clifford theorem,
Example: single scalar discrete random variable $X \in \{1, 2, \ldots, k\}$ might have indicator feature for all possible values $\alpha_i(x) \triangleq 1(x = i)$.

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 subsets of cliques.

Key: $p \in \mathcal{F}(G, R^{(f)})$ by Hammersley-Clifford theorem,

- where $G = (V, E)$ where $V$ is the nodes corresponding to vector $x$,
Example: single scalar discrete random variable $X \in \{1, 2, \ldots, k\}$ might have indicator feature for all possible values $\alpha_i(x) \triangleq 1(x = i)$.

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 subsets of cliques.

Key: $p \in \mathcal{F}(G, R^{(f)})$ 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 $u, v \in V(G) \iff$ 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 \( p \in \mathcal{F}((V, E), R^{(f)}) \) implies \( p \in \mathcal{F}((V, E + E_1), R^{(f)}) \) but in some sense, for any \( G \), we want to deal with the models for which \( G \) is tight (we don’t want to use overly complex graph to deal with family that is simpler)
- Exponential models can represent any factorization, given any factorization 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)
Underlying base measure $\nu$, $p$ is absolutely continuous w.r.t. $\nu$
absolutely continuous

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

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(x) \right\} = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$$  \hspace{1cm} (2)
exponential family models

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

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{\alpha \in I} \theta_\alpha \phi_\alpha(x) \right\} = \exp (\langle \theta, \phi(x) \rangle - A(\theta)) \quad (2)$$

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

$$A(\theta) = \log \int_{D_X} \exp (\langle \theta, \phi(x) \rangle) \nu(dx) \quad (3)$$

with $\theta \in \Omega \define \{ \theta \in \mathbb{R}^d | A(\theta) < +\infty \}$
exponential family models

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

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(x) \right\} = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \quad (2)$$

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

$$A(\theta) = \log \int_{D_x} \exp (\langle \theta, \phi(x) \rangle) \nu(dx) \quad (3)$$

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

$$p_{\theta}(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{\alpha \in I} \theta_{\alpha} \phi_{\alpha}(x) \right\} = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \quad (2)$$

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

$$A(\theta) = \log \int_{D_X} \exp (\langle \theta, \phi(x) \rangle) \nu(dx) \quad (3)$$

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

- $A(\theta)$ 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

\[
p_\theta(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(x) \right\} = \exp(\langle \theta, \phi(x) \rangle - A(\theta)) \tag{4}
\]

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

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

\[
\hat{\mu}_\alpha = \frac{1}{M} \sum_{i=1}^M \phi_\alpha(\bar{x}^{(i)}) \tag{5}
\]
Exponential family models

- Goal is to find

\[ p^* \in \arg\max_{p \in \mathcal{U}} H(p) \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \hat{\mu}_\alpha \quad \forall \alpha \in \mathcal{I} \tag{6} \]

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

\[ \mathbb{E}_p[\phi_\alpha(X)] = \int_{\mathcal{D}_X} \phi_\alpha(x)p(x)\nu(dx) \tag{7} \]

- This is solved by taking a distribution in the form of Eq. 4, by finding \( \theta \) that solves

\[ E_{p_\theta}[\phi_\alpha(X)] = \hat{\mu}_\alpha \text{ for all } \alpha \in \mathcal{I} \tag{8} \]
Minimal Representation of Exponential Family

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

where

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

- 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

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

where \( A(\theta) = \log \int_{D_x} \exp (\langle \theta, \phi(x) \rangle) \nu(dx) \)  

- Overcomplete representation \( d = |\mathcal{I}| \) higher than need be
Overcomplete Representation

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

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

- Overcomplete representation \( d = |\mathcal{I}| \) higher than need be
- i.e., \( \exists \gamma \neq 0 \text{ s.t. } \langle \gamma, \phi(x) \rangle = c, \forall x \) where \( c = \text{constant} \).
Overcomplete Representation

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

(11)

where

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

(12)

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

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

(13)

\[ = \exp(\langle \theta, \phi(x) \rangle + \langle \gamma, \phi(x) \rangle - A(\theta + \gamma)) \]  

(14)

\[ = \exp(\langle \theta, \phi(x) \rangle + c - A(\theta + \gamma)) \]  

(15)

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

(16)
Overcomplete Representation

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

where \( A(\theta) = \log \int_{D_x} \exp (\langle \theta, \phi(x) \rangle) \nu(dx) \)

- Overcomplete representation \( d = |I| \) higher than need be
- I.e., \( \exists \gamma \neq 0 \) s.t. \( \langle \gamma, \phi(x) \rangle = c, \forall x \) where \( c = \text{constant} \).
- 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

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

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

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

(11)

where \( A(\theta) = \log \int_{D_x} \exp(\langle \theta, \phi(x) \rangle) \nu(dx) \)  

(12)

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

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

(13)

\[ = \exp(\langle \theta, \phi(x) \rangle + \langle \gamma, \phi(x) \rangle - A(\theta + \gamma)) \]  

(14)

\[ = \exp(\langle \theta, \phi(x) \rangle + c - A(\theta + \gamma)) \]  

(15)

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

(16)

- 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.

Prof. Jeff Bilmes
Exponential family models

- Minimal representation of Bernoulli distribution is

\[ p(x|\gamma) = \exp(\gamma x - A(\gamma)) \]  

(17)
Exponential family models

- Minimal representation of Bernoulli distribution is
  \[ p(x|\gamma) = \exp(\gamma x - A(\gamma)) \]  
  (17)

- Overcomplete rep of Bernoulli dist.
  \[ p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \]  
  (18)
  \[ = \exp(\theta_0(1 - x) + \theta_1 x - A(\gamma)) \]  
  (19)

  where \( \theta = (\theta_0, \theta_1) \) and \( \phi(x) = (1 - x, x) \).
Exponential family models

- Minimal representation of Bernoulli distribution is

\[ p(x|\gamma) = \exp(\gamma x - A(\gamma)) \] (17)

- overcomplete rep of Bernoulli dist.

\[
p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \]
\[
= \exp(\theta_0(1-x) + \theta_1 x - A(\gamma)) \] (19)

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

\[ p(x|\gamma) = \exp(\gamma x - A(\gamma)) \]  

(17)

- Overcomplete rep of Bernoulli dist.

\[ p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \]  

(18)

\[ = \exp(\theta_0(1 - x) + \theta_1 x - A(\gamma)) \]  

(19)

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

\[ p(x | \gamma) = \exp(\gamma x - A(\gamma)) \] (17)

- overcomplete rep of Bernoulli dist.

\[ p(x | \theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \] (18)

\[ = \exp(\theta_0(1 - x) + \theta_1 x - A(\gamma)) \] (19)

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

\[ p(x|\gamma) = \exp(\gamma x - A(\gamma)) \] (17)

- Overcomplete rep of Bernoulli dist.

\[ p(x|\theta_0, \theta_1) = \exp(\langle \theta, \phi(x) \rangle) \] (18)

\[ = \exp(\theta_0(1 - x) + \theta_1 x - A(\gamma)) \] (19)

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

\[
p_\theta(x) = \exp \left\{ \sum_{v \in V} \theta_v x_v + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \right\}, \tag{20}
\]

with

\[
A(\theta) = \log \sum_{x \in \{0,1\}^m} \exp \left\{ \sum_{v \in V} \theta_v x_v + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \right\} \tag{21}
\]

- Note that this is in minimal form. Any change to parameters will result in different distribution
Note, in this case $\mathcal{I}$ is all singletons (unaries) and all pairs, so that
\[ \{ C_\alpha \}_\alpha = \left\{ \{ x_i \}_{i}, \{ x_i x_j \}_{(i,j) \in E} \right\}. \]

We can easily generalize this via a set system. I.e., consider $(V, \mathcal{V})$, where $\mathcal{V} = \{ V_1, V_2, \ldots, V_{|\mathcal{V}|} \}$ and where $\forall i, V_i \subseteq V$.

We can form sufficient statistic set via
\[ \{ C_\alpha \}_\alpha = \left\{ \{ x_V \}_{V \in \mathcal{V}} \right\}. \]

Higher order factors/interaction functions/potential functions/sufficient statistics.
Multivalued variables

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

$$1_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{else} \end{cases} \quad (22)$$

and

$$1_{st;jk}(x_s, x_t) = \begin{cases} 1 & \text{if } x_s = j \text{ and } x_t = k, \\ 0 & \text{else} \end{cases} \quad (23)$$

- Model becomes

$$p_\theta(x) = \exp \left\{ \sum_{v \in V} \sum_{i=0}^{r-1} \theta_{v;j} 1_{s;j}(x_v) + \sum_{(s,t) \in E} \sum_{j,k} \theta_{st;ij} 1_{st;jk}(x_s, x_t) - A(\theta) \right\} \quad (24)$$

- Is this overcomplete?
Multivalued variables

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

$$1_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{else} \end{cases}$$  \hspace{1cm} (22)$$

and

$$1_{st;jk}(x_s, x_t) = \begin{cases} 1 & \text{if } x_s = j \text{ and } x_t = k, \\ 0 & \text{else} \end{cases}$$  \hspace{1cm} (23)$$

- Model becomes

$$p_\theta(x) = \exp \left\{ \sum_{v \in V} \sum_{i=0}^{r-1} \theta_{v;i} 1_{s;j}(x_v) + \sum_{(s,t) \in E} \sum_{j,k} \theta_{st;ij} 1_{st;jk}(x_s, x_t) - A(\theta) \right\}$$  \hspace{1cm} (24)$$

- Is this overcomplete? Yes. Why?
Multivariate Gaussian

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

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

(25)
Multivariate Gaussian

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

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

(25)

- \(\langle \Theta, xx^T \rangle = \sum_{ij} \Theta_{ij}x_i x_j\) is Frobenius norm.
Multivariate Gaussian

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

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

(25)

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

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

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

(25)

- \(\langle \Theta, xx^T \rangle = \sum_{ij} \Theta_{ij} x_i x_j\) 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 Gaussian is parameterized via mean and covariance matrix. For canonical exponential form, we use mean and correlation matrix. I.e.

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

(25)

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

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

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

(25)

- \( \langle \Theta, xx^T \rangle = \sum_{ij} \Theta_{ij} x_i x_j \) 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 Gaussian is parameterized via mean and covariance matrix. For canonical exponential form, we use mean and correlation matrix. I.e.

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

(25)

- \(\langle \Theta, xx^T \rangle = \sum_{ij} \Theta_{ij}x_i x_j\) 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.
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).
Consider quantities $\mu_\alpha$ associated with statistic $\phi_\alpha$ defined as:

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

(26)
Mean Parameters, Convex Cores

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

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

(26)

- this defines a vector of “mean parameters” $(\mu_1, \mu_2, \ldots, \mu_d)$ with $d = |I|$.
Consider quantities $\mu_\alpha$ associated with statistic $\phi_\alpha$ defined as:

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x)p(x)\nu(dx) \quad (26)$$

This defines a vector of “mean parameters” $(\mu_1, \mu_2, \ldots, \mu_d)$ with $d = |I|$.

Define all the possible such vectors

$$\mathcal{M} \triangleq \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)], \forall \alpha \in I \right\} \quad (27)$$
Mean Parameters, Convex Cores

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

  $$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x)p(x)\nu(dx) \quad (26)$$

- This defines a vector of “mean parameters” $(\mu_1, \mu_2, \ldots, \mu_d)$ with $d = |I|$.
- Define all the possible such vectors

  $$\mathcal{M} \triangleq \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)], \forall \alpha \in I \right\} \quad (27)$$

- We don’t say $p$ was necessarily exponential family
Mean Parameters, Convex Cores

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

$$
\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x)p(x)\nu(dx) \quad (26)
$$

- this defines a vector of “mean parameters” $(\mu_1, \mu_2, \ldots, \mu_d)$ with $d = |I|$.

- Define all the possible such vectors

$$
\mathcal{M} \triangleq \{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)], \forall \alpha \in I \} \quad (27)
$$

- 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:

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x)p(x)\nu(dx) \quad (26)$$

- this defines a vector of “mean parameters” $(\mu_1, \mu_2, \ldots, \mu_d)$ with $d = |I|$.

- Define all the possible such vectors

$$\mathcal{M} \triangleq \left\{ \mu \in \mathbb{R}^d : \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)], \forall \alpha \in I \right\} \quad (27)$$

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

$$\mathcal{M} = \{ (\mu, C) \in \mathbb{R}^m \times S^m_+ | C - \mu\mu^\top \succeq 0 \}$$

(28)

where $S^m_+$ 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 constraint $\Sigma_{11} - \mu^2 \geq 0$. 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 constraint $\Sigma_{11} - \mu^2 \geq 0$. 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

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

$$= \text{conv} \left\{ \phi(x), x \in \mathcal{D}_X \right\}$$

where $\text{conv} \left\{ \cdot \right\}$ is the convex hull of the items in argument set.
Mean Parameters and Polytopes

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

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

$$= \text{conv} \{ \phi(x), x \in D_X \}$$

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

- So we have a convex polytope
Mean Parameters and Polytopes

Polytopes can be represented as a set of linear inequalities, i.e., there is a $|J| \times d$ matrix $A$ and $|J|$-element column vector $b$ with

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

(31)

with $A$ having rows $a_j$. 

\begin{figure}
\centering
\includegraphics[width=\textwidth]{polytope.png}
\caption{Polytope representation with inequalities $\langle a_j, \mu \rangle \geq b_j$ for all $j \in J$.}
\end{figure}
Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

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

we get

\[ \mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \quad \forall v \in V \]  
\[ \mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \quad \forall (s, t) \in E(G) \]
Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

\[ \phi(x) = \{ x_s, s \in V ; x_s x_t, (s, t) \in E(G) \} \in \mathbb{R}^{|V|+|E|} \] (32)

we get

\[ \mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \quad \forall v \in V \] (33)
\[ \mu_{s,t} = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \quad \forall (s, t) \in E(G) \] (34)

- 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 = conv \{ \phi(x), x \in \{0, 1\}^m \}. 
Mean Parameters and Polytopes

- Example: Ising mean parameters. Given sufficient statistics

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

we get

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

- 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

\[ \phi(x) = \{ x_s, s \in V; x_s x_t, (s, t) \in E(G) \} \in \mathbb{R}^{|V|+|E|} \] (32)

we get

\[ \mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \quad \forall v \in V \] (33)
\[ \mu_s,t = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \quad \forall (s, t) \in E(G) \] (34)

- Mean parameters lie in a polytope that represent the probabilities of a node being 1 or a pair of adjacent nodes being 1, 1 for each node and edge in the graph = \( \text{conv} \{ \phi(x), x \in \{0, 1\}^m \} \).

- Gives complete marginal since \( p_s(1) = 1 - p_s(0) \), \( p_{s,t}(1,0) = p_s(1) - p_{s,t}(1,1) \), \( p_{s,t}(0,1) = p_{t}(1) - p_{s,t}(1,1) \), etc.

- Recall: marginals are often the goal of inference.
Example: Ising mean parameters. Given sufficient statistics

\[\phi(x) = \{x_s, s \in V; x_s x_t, (s, t) \in E(G)\} \in \mathbb{R}^{\lvert V \rvert + \lvert E \rvert}\]  \hspace{1cm} (32)

we get

\[\mu_v = \mathbb{E}_p[X_v] = p(X_v = 1) \quad \forall v \in V \] \hspace{1cm} (33)

\[\mu_s,t = \mathbb{E}_p[X_s X_t] = p(X_s = 1, X_t = 1) \quad \forall (s, t) \in E(G) \] \hspace{1cm} (34)

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 = 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?
"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)

\[ \forall v \in V(G), j \in \{0 \ldots r - 1\}, \text{ define } \phi_{v,j}(x_v) \triangleq 1(x_v = j) \]  

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

\[ \phi_{st,jk}(x_s, x_t) \triangleq 1(x_s = j, x_t = k) = 1(x_s = j)1(x_t = k) \]

- So we now have \(|V|r + 2|E|r^2\) functions each with a corresponding parameter.
Mean Parameters and Marginal Polytopes

- Mean parameters are now true marginals, i.e., $\mu_v(j) = p(x_v = j)$ and $\mu_{st}(j, k) = p(x_s = j, x_t = k)$ since

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

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 $D = \{\bar{x}_E^{(i)}\}_{i=1}^M$ of size $M$, likelihood function

$$\ell(\theta, D) = \frac{1}{M} \sum_{i=1}^{M} \log p_{\theta}(\bar{x}^{(i)}) = \langle \theta, \hat{\mu} \rangle - A(\theta)$$ (40)

where

$$\hat{\mu} = \hat{\mathbb{E}}[\phi(X)] = \frac{1}{M} \sum_{i=1}^{M} \phi(\bar{x}^{(i)})$$ (41)
Learning is the dual of Inference

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

$$\ell(\theta, D) = \frac{1}{M} \sum_{i=1}^M \log p_\theta(\bar{x}^{(i)}) = \langle \theta, \hat{\mu} \rangle - A(\theta) \quad (40)$$

where

$$\hat{\mu} = \hat{E}[\phi(X)] = \frac{1}{M} \sum_{i=1}^M \phi(\bar{x}^{(i)}) \quad (41)$$

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

$$E_{\hat{\theta}}[\phi(X)] = \hat{\mu} \quad (42)$$

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 $D = \{\bar{x}_E^{(i)}\}_{i=1}^M$ of size $M$, likelihood function

$$\ell(\theta, D) = \frac{1}{M} \sum_{i=1}^M \log p_\theta(\bar{x}^{(i)}) = \langle \theta, \hat{\mu} \rangle - A(\theta) \quad (40)$$

where

$$\hat{\mu} = \hat{E}[\phi(X)] = \frac{1}{M} \sum_{i=1}^M \phi(\bar{x}^{(i)}) \quad (41)$$

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

$$E_{\hat{\theta}}[\phi(X)] = \hat{\mu} \quad (42)$$

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

- Also the maximum entropy problem.
Learning is the dual of Inference

- In other words, the solution to the maximum likelihood problem is one that satisfies the moment constraints and has the exponential model form. The exponential model form is exactly the equation that arises when we find the maximum entropy distribution over those distributions satisfying the moment constraints.
Learning is the dual of Inference

- In other words, the solution to the maximum likelihood problem is one that satisfies the moment constraints and has the exponential model form. The exponential model form is exactly the equation that arises when we find the maximum entropy distribution over those distributions satisfying the moment constraints.

- This shows that maximum entropy learning under a set of constraints (given by $\mathbb{E}_\theta[\phi(X)] = \hat{\mu}$) is the same as maximum likelihood learning of an exponential model form.
Learning is the dual of Inference

• In other words, the solution to the maximum likelihood problem is one that satisfies the moment constraints and has the exponential model form. The exponential model form is exactly the equation that arises when we find the maximum entropy distribution over those distributions satisfying the moment constraints.

• This shows that maximum entropy learning under a set of constraints (given by $E_\theta[\phi(X)] = \hat{\mu}$) 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 $\mu \in M$, this is the inference problem, getting the marginals.
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

$$A(\theta) = \log \int_{D_X} \langle \theta, \phi(x) \rangle \nu(dx)$$  \hspace{1cm} (43)

- 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 $\phi(X)$

$$\frac{\partial A}{\partial \theta^\alpha}(\theta) = \mathbb{E}_\theta[\phi^\alpha(X)] = \int \phi^\alpha(X)p_\theta(x)\nu(dx) = \mu^\alpha$$  \hspace{1cm} (44)

in general, derivative of log partition function is expected value of feature

- Also, we get

$$\frac{\partial^2 A}{\partial \theta^\alpha_1 \partial \theta^\alpha_2}(\theta) = \mathbb{E}_\theta[\phi^\alpha_1(X)\phi^\alpha_2(X)] - \mathbb{E}_\theta[\phi^\alpha_1(X)]\mathbb{E}_\theta[\phi^\alpha_2(X)]$$  \hspace{1cm} (45)

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 \( \nabla A : \Omega \rightarrow \mathcal{M}' \), where \( \mathcal{M}' \subseteq \mathcal{M} \), and where

\[
\mathcal{M} = \{ \mu \in \mathbb{R}^d | \exists p \text{ s.t. } E_p[\phi(X)] = \mu \}.
\]
So $\nabla A : \Omega \rightarrow \mathcal{M}'$, 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.
Log partition function

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

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

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

- For minimal exponential family models, this mapping is one-to-one, that is there is a unique pairing between $\mu$ 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 $\nabla A : \Omega \rightarrow \mathcal{M}'$, where $\mathcal{M}' \subseteq \mathcal{M}$, and where $\mathcal{M} = \{\mu \in \mathbb{R}^d | \exists \rho \text{ s.t. } \mathbb{E}_\rho[\phi(X)] = \mu\}$.

- 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.
In fact, we have

**Theorem 5.1**

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

**Theorem 5.2**

*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 $\mathbb{E}_\theta[\phi(X)] = \mu$.***
Moreover,

**Theorem 5.2**

*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 5.2**

*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).
Moreover,

**Theorem 5.2**

*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 same distribution (which in that case would not be an exponential family model with those moments).
Moreover,

**Theorem 5.2**

*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 same distribution (which in that case 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

- Maximum likelihood problem for exp. family

$$\theta^* \in \arg\max_{\theta} (\langle \theta, \hat{\mu} \rangle - A(\theta))$$  \hspace{1cm} (46)
Conjugate Duality

- Maximum likelihood problem for exp. family

\[ \theta^* \in \arg \max_{\theta} (\langle \theta, \hat{\mu} \rangle - A(\theta)) \]  

\[ \theta^* \in \arg \max_{\theta} (\langle \theta, \hat{\mu} \rangle - A(\theta)) \]  

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

\[ A^*(\mu) \triangleq \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) \]  

\[ A^*(\mu) \triangleq \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) \]
Conjugate Duality

- Maximum likelihood problem for exp. family

\[ \theta^* \in \arg \max_{\theta} (\langle \theta, \hat{\mu} \rangle - A(\theta)) \] (46)

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

\[ A^*(\mu) \triangleq \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) \] (47)

- So dual is optimal value of the ML problem, when \( \mu \in \mathcal{M} \)
Conjugate Duality

- Maximum likelihood problem for exp. family
  \[ \theta^* \in \arg\max_{\theta} (\langle \theta, \hat{\mu} \rangle - A(\theta)) \]  
  \hspace{1cm} (46)

- Convex conjugate dual of \( A(\theta) \) defined as:
  \[ A^*(\mu) \triangleq \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) \] 
  \hspace{1cm} (47)

- So dual is optimal value of the ML problem, when \( \mu \in M \)

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

  \[ \mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu \] 
  \hspace{1cm} (48)
Conjugate Duality

- Maximum likelihood problem for exp. family
  \[ \theta^* \in \arg \max_{\theta} (\langle \theta, \hat{\mu} \rangle - A(\theta)) \] (46)

- Convex conjugate dual of \( A(\theta) \) defined as:
  \[ A^*(\mu) \triangleq \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) \] (47)

- 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
  \[ \mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu \] (48)

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

Theorem 5.3

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

$$A^*(\mu) = \sup_{\theta \in \Omega} (\langle \theta, \mu \rangle - A(\theta)) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \in \mathcal{M} \end{cases}$$ (49)

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

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \}$$ (50)

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

$$\mu = \int_{D_X} \phi(x)p_{\theta}(x)\nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta)$$ (51)
Conjugate Duality

- Note that $A^*$ isn’t exactly entropy, but is only entropy sometimes.
- $A(\theta)$ in previous expression is the “inference” problem (dual of the dual) for a given $\theta$, whenever $\mu \notin \mathcal{M}$ we’ve got $\infty$ which can’t be sup, so need only consider $\mathcal{M}$.
- computing $A(\theta)$ in this way corresponds to the inference problem (finding mean parameters, or node and edge marginals). Key: we compute the log partition function simultaneously with solving inference, given the dual.
- Good news: problem is concave objective over a convex set. Should be easy. In simple examples, indeed, it is easy.
- Bad news: $\mathcal{M}$ is quite complicated to characterize, depends on the complexity of the graphical model.
- More bad news: $A^*$ not given explicitly in general and hard to compute.
Some good news: The above form gives us new avenues to do approximation.

Surprisingly, this is strongly related to belief propagation (i.e., the sum-product commutative semiring)!!
LBP and Bethe Approximation

- We’ll see that loopy belief-propagation (sum-product alg.) has much to do with an approximation to the aforementioned variational problems.
- Recall: we’re dealing only with pairwise interactions (natural for image processing, or convertible, as we’ve mentioned, or can define things via a factor graph).
- Exponential family model of form

\[
p_\theta(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{v \in V(G)} \theta_v(x_v) + \sum_{(s,t) \in E(G)} \theta_{st}(x_s, x_t) \right\}
\]

with simple new shorthand notation functions \( \theta_v \) and \( \theta_{st} \).

\[
\theta_v(x_v) \triangleq \sum_i \theta_{v,i} 1(x_v = i) \quad \text{and} \quad (52)
\]

\[
\theta_{s,t}(x_s, x_t) \triangleq \sum_{i,j} \theta_{st,ij} 1(x_s = i, x_t = j) \quad (53)
\]
We also have mean parameters that constitute the marginal polytope.

\[
\mu_v(x_v) \overset{\Delta}{=} \sum_{i \in D_{x_v}} \mu_{v,i} \mathbf{1}(x_v = i) \quad (54)
\]

\[
\mu_{st}(x_s, x_t) \overset{\Delta}{=} \sum_{(j,k) \in D_{x_{\{s,t\}}}} \mu_{st,jk} \mathbf{1}(x_s = j, x_t = k) \quad (55)
\]

\[
\text{And } \mathcal{M}(G) \text{ corresponds to the set of all singleton and pairwise marginals that can be jointly realized by some underlying probability distribution } p \in \mathcal{F}(G, R^{(f)}) \text{ that contains only pairwise interactions.}
\]

\[
\mathcal{M} \text{ can be represented as a convex hull of a set of points, or by a set of linear inequality constraints.}
\]
Local consistency polytope

- An “outer bound” of $M$ consists of a set that contains $M$, and if it is formed from a subset of the linear inequalities (subset of the rows of matrix module $(A, b)$), then it is a polyhedral outer bound. Let’s call this $L$.

- Another way to form outer bound: require only consistency, i.e., consider set $\{b_v, v \in V(G)\} \cup \{b_{s,t}, (s, t) \in E(G)\}$ that is non-negative and satisfies normalization

$$\sum_{x_v} b_v(x_v) = 1 \quad (57)$$

and pair-node marginal consistency constraints

$$\sum_{x_t'} b_{s,t}(x_s, x_t') = b_s(x_s) \quad (58)$$

$$\sum_{x_s'} b_{s,t}(x_s', x_t) = b_t(x_t) \quad (59)$$

$$\sum (60)$$
Local consistency polytope

- Define $\mathbb{L}(G)$ to be the (locally consistent) polytope that obeys these constraints.
- Clearly $\mathbb{M} \subseteq \mathbb{L}(G)$ since any member of $\mathbb{M}$ (true marginals) will be locally consistent.
- When $G$ is a tree, we say that local consistency implies global consistency, so for any tree $T$, we have $\mathbb{M}(T) = \mathbb{L}(T)$.
- When $G$ has cycles, however, $\mathbb{M}(G) \subset \mathbb{L}(G)$ strictly. We refer to members of $\mathbb{L}(G)$ as pseudo-marginals.
- Key issue is that members of $\mathbb{L}$ might not be possible marginals for any distribution.
Pseudo-marginals

\[ b_v(x_v) = [0.5, 0.5], \text{ and } b_{s,t}(x_s, x_t) = \begin{bmatrix} \beta_{st} & .5 - \beta_{st} \\ .5 - \beta_{st} & \beta_{st} \end{bmatrix} \quad (61) \]

- Consider on 3-cycle \( C_3 \), satisfies local consistency.
- But for this won’t give us a marginal. Below shows \( \mathbb{M}(C_3) \) for \( \mu_1(x_1 = 1) = \mu_2(x_2 = 1) = \mu_3(x_3 = 1) = 1/2 \) and the \( \mathbb{L}(C_3) \) outer bound.
Bethe Entropy Approximation

- Maybe it is hard to compute $A^*(\mu)$ but perhaps we can reasonably approximate it.

- In case when $-A^*(\mu)$ is the entropy, lets use an approximate entropy based on $\mathbb{L}$ being those distributions that factor w.r.t. a tree.

- When $G = T$ is a tree, we have

$$-A^*(\mu) = H(p_{\mu}) = \sum_{v \in V(T)} H(X_v) - \sum_{(s,t) \in E(T)} I(X_s; X_t) \quad (62)$$

$$= \sum_{v \in V(T)} H_v(\mu_v) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \quad (63)$$

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

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

Original variational representation of log partition function

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \quad (64) \]

Approximate variational representation of log partition function

\[ A_{\text{Bethe}}(\theta) = \sup_{b \in \mathbb{L}} \{ \langle \theta, b \rangle + H_{\text{Bethe}}(b) \} \quad (65) \]

\[ = \sup_{b \in \mathbb{L}} \left\{ \langle \theta, b \rangle + \sum_{v \in V(G)} H_v(b_v) - \sum_{(s,t) \in E(G)} l_{st}(b_{st}) \right\} \quad (66) \]

- Exact when \( G = T \) but we do this for any \( G \), still computable
- We get an approximate log partition function, and approximate (pseudo) marginals (in \( \mathbb{L} \)), but this is perhaps much easier to compute.
- We can optimize this directly using a Lagrangian formulation.
Bethe Variational Problem and LBP

- Lagrangian constraints for summing to unity at nodes

\[ C_{vv}(b) = 1 - \sum_{x_v} b_v(x_v) \]  

(67)

- Lagrangian constraints for local consistency

\[ C_{ts}(x_s; b) = b_s(x_s) - \sum_{x_t} b_{st}(x_s, x_t) \]  

(68)

- Yields following Lagrangian

\[
\mathcal{L}(b, \lambda; \theta) = \langle \theta, b \rangle + H_{\text{Bethe}}(b) + \sum_{v \in V} \lambda_{vv} C_{vv}(b) \\
+ \sum_{(s,t) \in E(G)} \left[ \sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s; b) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t; b) \right]
\]  

(69)
Fixed points: Variational Problem and LBP

Theorem 6.1

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

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

$$\nabla_b \mathcal{L}(b^*, \lambda^*; \theta) = 0 \text{ and } \nabla_\lambda \mathcal{L}(b^*, \lambda^*; \theta) = 0 \quad (71)$$

(b) For tree MRFs, Lagrangian equations have unique solution $(b^*, \lambda^*)$ where $b^*$ are exact node and edge marginals for the tree and the optimal value obtained is the true log partition function.

Remarkably, this means if we run loopy belief propagation, and we reach a point where we have converged, then we will have achieved a fixed-point of the above Lagrangian, and thus a (perhaps reasonable) local optimum of the underlying variational problem.
Fixed points: Variational Problem and LBP

- The resulting Lagrange multipliers $\lambda_{st}$ end up being exactly the messages that we have defined. I.e., we get

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

(72)

- Proof is not too difficult. Just take derivatives, set equal to zero, use Lagrangian constraints, do a bit of algebra, and amazingly, the BP messages suddenly pop out!!! (see page 86 in book).

- So we can now (at least) characterize any stable point of LBP.

- This does not mean that it will converge.

- For trees, we’ll get $A_{\text{Bethe}}(\theta) = A(\theta)$, results of previous lectures (parallel or MPP-based message passing).

- This does not mean $A_{\text{Bethe}}(\theta)$ will be a bound on $A(\theta)$ rather an approximation to it (mean-field methods which provide a lower bound on $A(\theta)$).

- For certain potential functions, we’ll get $A_{\text{Bethe}}(\theta) \leq A(\theta)$
Most of this material comes from the Wainwright and Jordan book.