Logistics
Review
Mean Field
Scratch
Summary

EE12A – Advanced Inference in Graphical Models
Fall 2011

Prof. Jeff Bilmes
University of Washington, Seattle
Department of Electrical Engineering
Fall Quarter, 2011
http://j.ee.washington.edu/~bilmes/classes/ee512a_fall_2011/

Lecture 16 - Nov 30th, 2011
Outstanding Reading

 Finish chapter 4 in the Wainwright and Jordan book.

 Another useful reading: An Introduction to Variational Methods for Graphical Models, by Jordan et. al. 1999 (available everywhere via web search).

 More project status reports (3rd status reports) are due tonight Wednesday evening (night before thanksgiving break) at 11:00pm via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/17463). Remember, no more than one page!
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): pseudo marg, Bethe
- LXX (11/11): Veterans Day, no class
- L13 (11/16): Bethe, loop series
- LXX (11/25): Thanksgiving, no class
- L16 (11/30): mean field
- L17 (12/2):
- L18 (12/7):
- L19: (12/9):
- L20: ???:
Finish chapter 4 from Wainwright & Jordan book.
Conjugate Duality

Theorem 2.1 (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:

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

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

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

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

$$\mu = \int_{\mathcal{D}_X} \phi(x) p_{\theta}(x) \nu(dx) = E_{\theta}[\phi(X)] = \nabla A(\theta)$$
Variational Problem

Original variational representation of log partition function

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

- Set \( \mathcal{M} \leftarrow \mathbb{L} \) and \( -A^*(\mu) \leftarrow H_{\text{Bethe}}(\tau) \) to get Bethe variational approximation, LBP fixed point.
Variational Problem

Original variational representation of log partition function

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \] (4)

- Set \( \mathcal{M} \leftarrow \mathbb{L} \) and \( -A^*(\mu) \leftarrow H_{\text{Bethe}}(\tau) \) to get Bethe variational approximation, LBP fixed point.
- Set \( \mathcal{M} \leftarrow \mathbb{L}_t(G) \) (hypertree marginal polytope), \( -A^*(\mu) \leftarrow H_{\text{app}}(\tau) \) where \( H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \) (via Möbius) to get Kikuchi variational approximation, message passing on hypertrees.
Variational Problem

Original variational representation of log partition function

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \] (4)

- Set \( \mathcal{M} \leftarrow \mathbb{L} \) and \(-A^*(\mu) \leftarrow H_{\text{Bethe}}(\tau)\) to get Bethe variational approximation, LBP fixed point.
- Set \( \mathcal{M} \leftarrow \mathbb{L}_t(G) \) (hypertree marginal polytope), \(-A^*(\mu) \leftarrow H_{\text{app}}(\tau)\) where \( H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \) (via Möbius) to get Kikuchi variational approximation, message passing on hypertrees.
- Partition \( \tau \) into \((\tau, \tilde{\tau})\), and set \( \mathcal{M} \leftarrow \mathcal{L}(\phi, \Phi) \) and set \(-A^*(\mu) \leftarrow H_{\text{ep}}(\tau, \tilde{\tau})\) to get expectation propagation.
Expectation Propagation: basic idea

- Came from a method called “assumed density filtering” (ADF).
- Doing full inference involves exponential computation.
- We do a bit of inference, involving reasonable computation, and getting us a new distribution that is a bit more complex but not too much more complex.
- Before going further, we “project” this new distribution back down to a class of simple distributions.
- We then repeat the above step with a bit more of inference, different than what we did above.
- We keep repeating: do a bit of inference, and project, until all inference has been done.
- The difference between ADF and EP is that, with ADF at this stage we’re done. With EP we can keep repeating the process of inference, projection.
- EP can be seen as a generalization of BP.
- Interestingly, EP is instance of our variational framework, Equation 2.
Term Decoupling

- Partition the sufficient statistics into two parts:
  - Tractable component
    \[
    \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T})
    \]  
    (5)
  - Intractable component
    \[
    \Phi \triangleq (\Phi^1, \Phi^2, \ldots, \Phi^{d_I})
    \]  
    (6)

- \(\phi_i\) are typically univariate, while \(\Phi^i\) are multivariate (\(b\)-dimensional).
- Consider exponential families associated with subcollection \((\phi, \Phi)\).
Tractable component

\[ \phi \triangleq (\phi_1, \phi_2, \ldots, \phi_{d_T}) \] (7)

So \( \phi : \mathcal{X}^m \rightarrow \mathbb{R}^{d_T} \) with vector of parameters \( \theta \in \mathbb{R}^{d_T} \).

Could instantiate model based only on this subcomponent, called the base model.


Intractable component

\[ \Phi \triangleq (\Phi_1, \Phi_2, \ldots, \Phi_{d_I}) \tag{8} \]

- Each \( \Phi_i : \mathcal{X}^m \to \mathbb{R}^b \).
- \( \Phi : \mathcal{X}^m \to \mathbb{R}^{b \times d_I} \).
- Parameters \( \tilde{\theta} \in \mathbb{R}^{b \times d_I} \).
Associated Distributions

- The associated exponential family

\[ p(x; \theta, \tilde{\theta}) \propto f_0(x) \exp (\langle \theta, \phi(x) \rangle) \exp (\langle \tilde{\theta}, \Phi(x) \rangle) \quad (9) \]

\[ = f_0(x) \exp (\langle \theta, \phi(x) \rangle) \prod_{i=1}^{d_I} \exp (\langle \tilde{\theta}^i, \Phi^i(x) \rangle) \quad (10) \]

- Base model is tractable

\[ p(x; \theta, \tilde{0}) \propto f_0(x) \exp (\langle \theta, \phi(x) \rangle) \quad (11) \]

- \( \Phi^i \)-augmented model

\[ p(x; \theta, \tilde{\theta}^i) \propto f_0(x) \exp (\langle \theta, \phi(x) \rangle) \exp (\langle \tilde{\theta}^i, \Phi^i(x) \rangle) \quad (12) \]
The basic premises in the tractable-intractable partitioning between $\phi$ and $\Phi$ are:

- It is possible to compute marginals exactly in polynomial time for distributions of the base form (any member of the $\phi$-exponential family).
- For each $i = 1, \ldots, d_I$, exact polynomial-time computation is still possible for any $\Phi^i$-augmented form (any member of the $(\phi, \Phi^i)$-exponential family).
- Intractable to perform exact computations with the full $(\phi, \Phi)$-exponential family.
Example: Mixture models

- Let $\varphi(y; \mu, \Lambda)$ be Normal with mean $\mu$ covariance $\Lambda$.
- Two component Gaussian mixture model

\[ p(y|X = x) = (1 - \alpha)\varphi(y; 0, \sigma_0^2 I) + \alpha\varphi(y; x, \sigma_1^2 I) \]  

(13)

- Assume $n$ i.i.d. samples $y^1, \ldots, y^n$ from mixture density, and goal is to produce posterior $p(x|y^1, \ldots, y^n)$, similar to Bayes-rule inverting a Naive-Bayes model.
- Using Bayes rule

\[
p(x|y^1, \ldots, y^n) \propto \exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right) \prod_{i=1}^{n} p(y^i|X = x) \]

(14)

\[
= \exp \left( -\frac{1}{2} x^\top \sigma^{-1} x \right) \exp \left\{ \sum_{i=1}^{n} \log p(y^i|X = x) \right\} \]

(15)
Example: Mixture models

- We equate $\exp\left(-\frac{1}{2}x^\top \sigma^{-1} x\right)$ with $f_0(x) \exp\langle \theta, \phi(x) \rangle$, with $d_T = m$.
- Such a distribution is Gaussian, and getting marginals is “cheap” $O(m^3)$.
- Also, $\exp \left\{ \sum_{i=1}^{n} \log p(y^i|X = x) \right\}$ cor. to $\prod_{i=1}^{d_I} \exp \left( \langle \tilde{\theta}^i, \Phi^i(x) \rangle \right)$, with $b = 1$.
- Base distribution $p(x; \theta, \vec{0}) \propto \exp \left(-\frac{1}{2}x^\top \sigma^{-1} x\right)$ which is a Gaussian and easy.
- If we multiply in one intractable term, still not so bad (quite easy in fact).
- I.e., $\Phi^i$-augmented distribution is proportional to

$$\exp \left(-\frac{1}{2}x^\top \sigma^{-1} x\right) \left[(1 - \alpha)\varphi(y^i; 0, \sigma_0^2 I) + \alpha\varphi(y^i; x, \sigma_1^2 I)\right] \quad (16)$$

- Computing marginals on this is not too bad (mixture of 2 components)
- If we multiply in all $\Phi^i$, becomes intractable ($2^n$ components)
Polytope and Base case

- We can partition the mean parameters \((\mu, \tilde{\mu}) \in \mathbb{R}^{d_T + d_I \times b}\)
- Marginal polytope associated with these means

\[
\mathcal{M}(\phi, \Phi) = \{ (\mu, \tilde{\mu}) | (\mu, \tilde{\mu}) = \mathbb{E}_p[(\phi(X), \Phi(X))] \text{ for some } p \} \quad (17)
\]

along with negative dual of cumulant, or entropy \(H(\mu, \tilde{\mu}) = -A^*(\mu, \tilde{\mu})\).
- We also have polytope associated with only base distribution

\[
\mathcal{M}(\phi) = \left\{ \mu \in \mathbb{R}^{d_T} | \mu = \mathbb{E}_p(\phi(X)) \right\} \quad (18)
\]

- Recall thm: any mean in the interior is realizable via an exponential family model, and associated entropy \(H(\mu)\) is tractable.
Augmented Base case

For each $i = 1 \ldots d_I$ we have a $\Phi^i$-augmented exp. model and polytope

$$M(\phi, \Phi^i) = \left\{ (\mu, \tilde{\mu}^i) \in \mathbb{R}^{d_T+b} \mid (\mu, \tilde{\mu}^i) = \mathbb{E}_p[(\phi(X), \Phi^i(X))] \text{ for some } p \right\}$$

(19)

Thus, any such mean parameters has instance for associated exponential family, and also $H(\mu, \tilde{\mu}^i)$ is easy to compute.

Goal, variational approximation: Need outer bounds on $M(\phi, \Phi)$ and expression for entropy (as is now normal).

Turns out we can do this, and an iterative algorithm to find fixed points of associated Lagrangian, that correspond to EP.
New outer bound

- For any mean parms \((\tau, \tilde{\tau})\) define “projection operation”

\[
\Pi^i(\tau, \tilde{\tau}) \rightarrow (\tau, \tilde{\tau}^i)
\]  (20)

- Define outer bound on true means \(M(\phi, \Phi)\) (which is still convex)

\[
\mathcal{L}(\phi, \Phi) = \left\{ (\tau, \tilde{\tau}) \mid \tau \in M(\phi), \Pi^i(\tau, \tilde{\tau}) \in M(\phi, \Phi^i), \forall i \right\}
\]  (21)

- Note, based on a set of projections onto \(M(\phi, \Phi^i)\). Clearly outer bound.

- If \(\Phi^i\) are edges of a graph, then standard \(L\) outer bound we saw before with Bethe approximation.
Members in new outer bound \( H(i,j) - H(i) - H(j) \)

For any mean params \((\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)\)

- There is a member of the \(\phi\)-exponential family which mean parameters \(\tau\) with entropy \(H(\tau)\)
- For \(i = 1 \ldots d_I\), there is a member of the \((\phi, \Phi^i)\)-exponential family with mean parameters \((\tau, \tilde{\tau}^i)\) with entropy \(H(\tau, \tilde{\tau}^i)\)
- Consider new entropy approximation

\[
H(\tau, \tilde{\tau}) \approx H_{ep}(\tau, \tilde{\tau}) \triangleq H(\tau) + \sum_{l=1}^{d_I} \left[ H(\tau, \tilde{\tau}^l) - H(\tau) \right] \quad (22)
\]

- With outer bound and entropy expression, we get new variational form

\[
\max_{(\tau, \tilde{\tau}) \in \mathcal{L}(\phi, \Phi)} \left\{ \langle \tau, \theta \rangle + \langle \tilde{\tau}, \tilde{\theta} \rangle + H_{ep}(\tau, \tilde{\tau}) \right\} \quad (23)
\]

- This covers the EP algorithms.
- When we take \(\phi\) to be unaries and \(\Phi\) to be edges, we get exactly Bethe approximation.
Moment Matching

- We may define a Lagrangian version of the objective

\[
L(\tau; \lambda) = \langle \tau, \theta \rangle + \sum_{i=1}^{d_I} \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + F(\tau; (\eta^i, \tilde{\tau}^i)) + \sum_{i=1}^{d_I} \langle \lambda^i, \tau - \eta^i \rangle
\]

(24)

where

\[
F(\tau; (\eta^i, \tilde{\tau}^i)) = H(\tau) + \sum_{i=1}^{d_I} \left[ H(\eta^i, \tilde{\tau}^i) - H(\eta^i) \right]
\]

(25)

- Resulting exponential family may be written as:

\[
q(x; \theta, \lambda) \propto f_0(x) \exp \left\{ \langle \theta + \sum_{i=1}^{d_I} \lambda^i, \phi(x) \rangle \right\}
\]

(26)
Moment Matching → Expectation Propagation Updates

1. At iteration $n = 0$, initialize the Lagrange multiplier vectors $(\lambda^1, \ldots, \lambda^{d_I})$.

2. At each iteration $n = 1, 2, \ldots$ choose some index $i(n) \in \{1, \ldots, d_I\}$.

3. For the following distribution

$$q^i(x; \theta, \tilde{\theta}^i, \lambda) \propto f_0(x) \exp \left( \left\langle \theta + \sum_{\ell \neq i} \lambda^\ell, \phi(x) \right\rangle + \left\langle \tilde{\theta}^i, \Phi^i(x) \right\rangle \right) \tag{27}$$

and then compute the mean parameters

$$\eta^i(n) = \int q^i(n)(x) \phi(x) \nu(dx) = \mathbb{E}_{q^i(n)}[\phi(X)] \tag{28}$$

4. Form base distribution $q$ using Equation 26 and adjust $\lambda^{i(n)}$ to satisfy the moment-matching condition (a max-ent problem)

$$\mathbb{E}_q[\phi(X)] = \eta^i(n) \tag{29}$$
Moment Matching $\rightarrow$ Expectation Propagation Updates

1. When base distribution is unaries and $\Phi^i$ is the edges of a graph, we get Bethe approximation, and standard sum-product LBP.

2. When base distribution is a tree, we get tree-structured EP.

3. Lost of flexibility here, depending on what the base distribution is (e.g., could be a $k$-tree or any other structure).

4. Can also be done for Gaussian mixture models.

5. Many more details in text and also see Tom Minka’s papers.
So far, we have been using an outer bound on $\mathcal{M}$. 
So far, we have been using an outer bound on $\mathcal{M}$.

In mean-field methods, we use an “inner bound”, a subset of $\mathcal{M}$ constructed so as to make the optimization of $A(\theta)$ easier.
Mean Field

- So far, we have been using an outer bound on $\mathcal{M}$.
- In mean-field methods, we use an “inner bound”, a subset of $\mathcal{M}$ constructed so as to make the optimization of $A(\theta)$ easier.
- Since subset, we get immediate bound on $A(\theta)$. 
So far, we have been using an outer bound on $\mathcal{M}$.

In mean-field methods, we use an “inner bound”, a subset of $\mathcal{M}$ constructed so as to make the optimization of $A(\theta)$ easier.

Since subset, we get immediate bound on $A(\theta)$.

Key: we based the inner bound on a “tractable family” like a 1-tree or even a 0-tree (all independent) so that the variational problem can be computed efficiently.
So far, we have been using an outer bound on $\mathcal{M}$.

In mean-field methods, we use an “inner bound”, a subset of $\mathcal{M}$ constructed so as to make the optimization of $A(\theta)$ easier.

Since subset, we get immediate bound on $A(\theta)$.

Key: we based the inner bound on a “tractable family” like a 1-tree or even a 0-tree (all independent) so that the variational problem can be computed efficiently.

Convexity is often lost still, however.
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a subgraph $F$ that is a spanning subgraph.

  all nodes \ subset of \ edges
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a subgraph $F$ that is a spanning subgraph.
- Simplest example: $F = (V, \emptyset)$.
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a subgraph $F$ that is a spanning subgraph.
- Simplest example: $F = (V, \emptyset)$ all independence model.
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a subgraph $F$ that is a spanning subgraph.
- Simplest example: $F = (V, \emptyset)$ all independence model.
- Tree example: $F = (V, E_T)$ where edges $E_T \subset E$ constitute a spanning tree.
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a subgraph $F$ that is a spanning subgraph.
- Simplest example: $F = (V, \emptyset)$ all independence model.
- Tree example: $F = (V, E_T)$ where edges $E_T \subset E$ constitute a spanning tree.
- Exponential family, sufficient statistics $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ associated with this family $\mathcal{I}(F) \subseteq \mathcal{I}$. These are the statistics that respect the Markov properties of subgraph $F$. 

\[ \Omega(\mathcal{I}(F)) \]
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a subgraph $F$ that is a spanning subgraph.
- Simplest example: $F = (V, \emptyset)$ all independence model.
- Tree example: $F = (V, E_T)$ where edges $E_T \subset E$ constitute a spanning tree.
- Exponential family, sufficient statistics $\phi = (\phi_\alpha, \alpha \in I)$ associated with this family $I(F) \subseteq I$. These are the statistics that respect the Markov properties of subgraph $F$.
- $\Omega$ gets smaller too. The parameters that respect $F$ are of the form:

$$
\Omega(F) \triangleq \{ \theta \in \Omega | \theta_\alpha = 0 \ \forall \alpha \in I \setminus I(F) \} 
$$

$$
\downarrow
$$

$$
\mathcal{N}(F) \in \mathbb{R}^{|I|}
$$
Tractable Families

- We have graph $G = (V, E)$ which is intractable and we find a subgraph $F$ that is a spanning subgraph.
- Simplest example: $F = (V, \emptyset)$ all independence model.
- Tree example: $F = (V, E_T)$ where edges $E_T \subset E$ constitute a spanning tree.

- Exponential family, sufficient statistics $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ associated with this family $\mathcal{I}(F) \subseteq \mathcal{I}$. These are the statistics that respect the Markov properties of subgraph $F$.
- $\Omega$ gets smaller too. The parameters that respect $F$ are of the form:

\[
\Omega(F) \triangleq \{ \theta \in \Omega | \theta_\alpha = 0 \ \forall \alpha \in \mathcal{I} \setminus \mathcal{I}(F) \} \tag{30}
\]

Notice, all parameters associated with sufficient statistic not in $\mathcal{I}(F)$ are set to zero, so those statistics are essentially “turned off.”
Tractable Subgraphs: All Independent Example

- **Ex:** MRF with potential functions for nodes and edges.
Tractable Subgraphs: All Independent Example

- Ex: MRF with potential functions for nodes and edges.
- For each $(s, t) \in E(G)$, we have $\theta_{(s,t)}$. 

$$F_0 = (V, \emptyset)$$ which yields

$$\Omega(F_0) = \{\theta \in \Omega | \theta((s,t)) = 0 \forall (s,t) \in E(G)\}$$

(31)
Tractable Subgraphs: All Independent Example

- Ex: MRF with potential functions for nodes and edges.
- For each \((s, t) \in E(G)\), we have \(\theta_{(s,t)}\).
- \(F_0 = (V, \emptyset)\) which yields

\[
\Omega(F_0) = \{ \theta \in \Omega | \theta_{(s,t)} = 0 \ \forall (s, t) \in E(G) \}
\] (31)
Tractable Subgraphs: All Independent Example

- Ex: MRF with potential functions for nodes and edges.
- For each \((s, t) \in E(G)\), we have \(\theta_{(s,t)}\).
- \(F_0 = (V, \emptyset)\) which yields

\[
\Omega(F_0) = \{ \theta \in \Omega | \theta_{(s,t)} = 0 \ \forall (s, t) \in E(G) \} \tag{31}
\]

- This is the all independence model, giving family of distributions

\[
p_\theta(x) = \prod_{s \in V} p(x_s; \theta_s) \tag{32}
\]
Tractable Subgraphs: Tree Example

- Ex: MRF with potential functions for nodes and edges.
Tractable Subgraphs: Tree Example

- Ex: MRF with potential functions for nodes and edges.
- For each \((s, t) \in E(G)\), we have \(\theta_{(s,t)}\).
Tractable Subgraphs: Tree Example

- Ex: MRF with potential functions for nodes and edges.
- For each \((s, t) \in E(G)\), we have \(\theta_{(s,t)}\).
- \(F_T = (V, T)\) where \(T \subset E\) are edges that constitute a spanning tree of \(G\), giving

\[
\Omega(F_0) = \{ \theta \in \Omega | \theta_{(s,t)} = 0 \ \forall (s, t) \notin T \} \tag{33}
\]
Tractable Subgraphs: Tree Example

- Ex: MRF with potential functions for nodes and edges.
- For each \((s, t) \in E(G)\), we have \(\theta_{(s,t)}\).
- \(F_T = (V, T)\) where \(T \subset E\) are edges that constitute a spanning tree of \(G\), giving

\[
\Omega(F_0) = \{ \theta \in \Omega | \theta_{(s,t)} = 0 \ \forall (s, t) \notin T \} \tag{33}
\]

- This gives a tree-dependent family

\[
p_{\theta}(x) = \prod_{s \in V} p(x_s; \theta_s) \prod_{(s,t) \in T} \frac{p(x_s, x_t; \theta_{st})}{p(x_s; \theta_s)p(x_t; \theta_t)} \tag{34}
\]
Before, we had $\mathcal{M}(G; \phi)$, all possible mean parameters associated with $G$ and associated set of sufficient statistics $\phi$. 
Before, we had $\mathcal{M}(G; \phi)$, all possible mean parameters associated with $G$ and associated set of sufficient statistics $\phi$.

For a given subgraph $F$, we only consider those mean parameters possible under such models. I.e.,

$$\mathcal{M}_F(G; \phi) = \left\{ \mu \in \mathbb{R}^d | \mu = \mathbb{E}_\theta[\phi(x)] \text{ for some } \theta \in \Omega(F) \right\}$$  (35)
Before, we had $\mathcal{M}(G; \phi)$, all possible mean parameters associated with $G$ and associated set of sufficient statistics $\phi$.

For a given subgraph $F$, we only consider those mean parameters possible under such models. I.e.,

$$
\mathcal{M}_F(G; \phi) = \left\{ \mu \in \mathbb{R}^d | \mu = \mathbb{E}_\theta[\phi(x)] \text{ for some } \theta \in \Omega(F) \right\}
$$

Therefore, we have that

$$
\mathcal{M}^\circ_F(G; \phi) \subseteq \mathcal{M}^\circ(G; \phi)
$$

and so $\mathcal{M}^\circ_F(G; \phi)$ is an inner approximation of the set of realizable mean parameters.
Before, we had $\mathcal{M}(G; \phi)$, all possible mean parameters associated with $G$ and associated set of sufficient statistics $\phi$.

For a given subgraph $F$, we only consider those mean parameters possible under such models. I.e.,

$$
\mathcal{M}_F(G; \phi) = \left\{ \mu \in \mathbb{R}^d | \mu = \mathbb{E}_\theta[\phi(x)] \text{ for some } \theta \in \Omega(F) \right\}
$$

Therefore, we have that

$$
\mathcal{M}_F^\circ(G; \phi) \subseteq \mathcal{M}^\circ(G; \phi)
$$

and so $\mathcal{M}_F^\circ(G; \phi)$ is an inner approximation of the set of realizable mean parameters.

Shorthand notation: $M_F^\circ(G) = M_F^\circ(G; \phi)$ and $M^\circ(G) = M^\circ(G; \phi)$
Mean field variational lower bound

Mean field methods generate lower bounds on their estimated $A(\theta)$ and approximate mean parameters $\mu = \mathbb{E}_\theta[\phi(X)]$.

Proposition 3.1 (mean field lower bound)

Any mean parameter $\mu \in \mathcal{M}^\circ$ yields a lower bound on the cumulant function:

$$A(\theta) \geq \langle \theta, \mu \rangle - A^*(\mu)$$

Moreover, equality holds if and only if $\theta$ and $\mu$ are dually coupled (i.e., $\mu = \mathbb{E}_\theta[\phi(X)]$).
Mean field variational lower bound

Proof.

On the one hand, obvious due to $A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$
Proof.

- On the one hand, obvious due to $A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \}$

- More traditional proof, let $q$ be any distribution that satisfies moment matching $E_q[\phi(X)] = \mu$, then:

\[ A(\theta) = \log \int_{\mathcal{X}^m} q(x) \frac{\exp \langle \theta, \phi(x) \rangle}{q(x)} \nu(dx) \]

\[ \geq \int_{\mathcal{X}^m} q(x) [\langle \theta, \phi(x) \rangle - \log q(x)] \nu(dx) \]

\[ = \langle \theta, \mu \rangle - H(q) \]
Mean field variational lower bound

Proof.

- On the one hand, obvious due to \( A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \)

- More traditional proof, let \( q \) be any distribution that satisfies moment matching \( \mathbb{E}_q[\phi(X)] = \mu \), then:

\[
A(\theta) = \log \int_{X^m} q(x) \frac{\exp \langle \theta, \phi(x) \rangle}{q(x)} \nu(dx)
\]

\[
\geq \int_{X^m} q(x) [\langle \theta, \phi(x) \rangle - \log q(x)] \nu(dx)
\]

\[
= \langle \theta, \mu \rangle - H(q)
\]  

- If we optimize \( q \) over all \( \mathcal{M}(G) \), then we’ll get equality.
Mean field variational lower bound

Proof.

- On the one hand, obvious due to $A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \}$

- More traditional proof, let $q$ be any distribution that satisfies moment matching $\mathbb{E}_q[\phi(X)] = \mu$, then:

$$A(\theta) = \log \int_{\mathcal{X}^m} q(x) \frac{\exp \langle \theta, \phi(x) \rangle}{q(x)} \nu(dx)$$  \hspace{1cm} (38)

$$\geq \int_{\mathcal{X}^m} q(x) [\langle \theta, \phi(x) \rangle - \log q(x)] \nu(dx)$$  \hspace{1cm} (39)

$$= \langle \theta, \mu \rangle - H(q)$$  \hspace{1cm} (40)

- If we optimize $q$ over all $\mathcal{M}(G)$, then we’ll get equality.

- If we optimize $q$ over a subset of $\mathcal{M}(G)$ (e.g., such as $\mathcal{M}_F(G)$), then we’ll get inequality.
Tractable Dual

- Normally dual $A^*$ is intractable, but key idea is that if $\mu \in \mathcal{M}_F(G)$ it will be possible to compute easily.
Normalize dual $A^*$ is intractable, but key idea is that if $\mu \in \mathcal{M}_F(G)$ it will be possible to compute easily.

Thus, goal of mean field (from variational approximation perspective) is to do

$$\max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A^*_F(\mu) \}$$

where $A^*_F(\mu)$ corresponds to dual function restricted to inner bound set $\mathcal{F}(G)$, i.e., when we expand $A^*_F(\mu)$ we can take advantage of $\mu \in \mathcal{M}_F(G)$ so that $A^*_F(\mu)$ is simpler.
Given two distributions \( p, q \), KL-Divergence of \( p \) w.r.t. \( q \) is defined as

\[
D(q || p) = \int_{\mathcal{X}} q(x) \left[ \log \frac{q(x)}{p(x)} \right] \nu(dx) \tag{42}
\]
Given two distributions $p, q$, KL-Divergence of $p$ w.r.t. $q$ is defined as

$$D(q||p) = \int_{\mathcal{X}} q(x) \left[ \log \frac{q(x)}{p(x)} \right] \nu(dx)$$  \hspace{1cm} (42)

In summation form, we have

$$D(q||p) = \sum_{x \in \mathcal{X}} q(x) \left[ \log \frac{q(x)}{p(x)} \right]$$  \hspace{1cm} (43)
Given two distributions $p, q$, KL-Divergence of $p$ w.r.t. $q$ is defined as

$$D(q||p) = \int_{\mathcal{X}^m} q(x) \left[ \log \frac{q(x)}{p(x)} \right] \nu(dx)$$

(42)

In summation form, we have

$$D(q||p) = \sum_{x \in \mathcal{X}^m} q(x) \left[ \log \frac{q(x)}{p(x)} \right]$$

(43)

For exponential models this takes on some interesting forms, and moreover, we can see the variational approximation above as a KL-divergence minimization problem.
Mean field, KL-Divergence, Exponential Model Families

- Consider $\theta^1, \theta^2 \in \Omega$
Mean field, KL-Divergence, Exponential Model Families

- Consider $\theta^1, \theta^2 \in \Omega$
- Let $D(\theta^1||\theta^2)$ have obvious meaning (KL-divergence between two corresponding distributions), and let $\mu^i = \mathbb{E}_{\theta^i}[\phi(X)]$, 

\[ D(\theta^1||\theta^2) = A(\theta^2) - A(\theta^1) - \langle \mu^1, \theta^2 - \theta^1 \rangle \]
Mean field, KL-Divergence, Exponential Model Families

- Consider $\theta^1, \theta^2 \in \Omega$
- Let $D(\theta^1 || \theta^2)$ have obvious meaning (KL-divergence between two corresponding distributions), and let $\mu^i = \mathbb{E}_{\theta^i} [\phi(X)]$,
- Then we have a Bregman divergence form:

$$D(\theta^1 || \theta^2) = \mathbb{E}_{\theta^1} \left[ \frac{p_{\theta^1}(x)}{p_{\theta^2}(x)} \log \frac{p_{\theta^1}(x)}{p_{\theta^2}(x)} \right]$$

(44)

$$= A(\theta^2) - A(\theta^1) - \langle \mu^1, \theta^2 - \theta^1 \rangle$$

(45)

$$= A(\theta^2) - \left[ A(\theta^1) + \langle \nabla A(\theta^1), \theta^2 - \theta^1 \rangle \right]$$

(46)

$$= A(\theta^2) - B_{\theta^1}(\theta^2)$$
Purely dual form of KL divergence can be formed as well, i.e.,

$$\mathcal{D}(\theta^1 \| \theta^2) = \mathcal{D}(\mu^1 \| \mu^2) = A^*(\mu^1) - A^*(\mu^2) - \langle \theta^2, \mu^1 - \mu^2 \rangle \tag{47}$$
Purely dual form of KL divergence can be formed as well, i.e.,

\[
D(\theta^1 \| \theta^2) = D(\mu^1 \| \mu^2) = A^*(\mu^1) - A^*(\mu^2) - \langle \theta^2, \mu^1 - \mu^2 \rangle \tag{47}
\]

- Dual Bregman form
Mean field, KL-Divergence, Exponential Model Families

- Mixed/hybrid form of KL in terms of dual
Mixed/hybrid form of KL in terms of dual

We can also write the KL as:

\[
D(\theta^1 || \theta^2) = D(\mu^1 || \theta^2) = A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle
\]  

(48)

which comes from dual expression \( A^*(\mu^1) = \langle \theta^1, \mu^1 \rangle - A(\theta^1) \) for dually coupled parameters \( \mu^1 = \mathbb{E}_{\theta^1}[\phi(X)] \).
Mixed/hybrid form of KL in terms of dual

We can also write the KL as:

\[
D(\theta^1 || \theta^2) = D(\mu^1 || \theta^2) = A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle
\] (48)

which comes from dual expression \( A^*(\mu^1) = \langle \theta^1, \mu^1 \rangle - A(\theta^1) \) for dually coupled parameters \( \mu^1 = \mathbb{E}_{\theta^1}[\phi(X)] \).

In particular, this equation (variational expression for the cumulant):

\[
A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \}
\] (2)
Mean field, KL-Divergence, Exponential Model Families

- Mixed/hybrid form of KL in terms of dual
- We can also write the KL as:

\[ D(\theta^1||\theta^2) = D(\mu^1||\theta^2) = A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle \]  

(48)

which comes from dual expression \( A^*(\mu^1) = \langle \theta^1, \mu^1 \rangle - A(\theta^1) \) for dually coupled parameters \( \mu^1 = \mathbb{E}_{\theta^1}[\phi(X)] \).

- In particular, this equation (variational expression for the cumulant):

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

(2)

... can be written as:

\[ \inf_{\mu \in \mathcal{M}} \{ A(\theta) + A^*(\mu) - \langle \theta, \mu \rangle \} = \inf_{\mu \in \mathcal{M}} D(\mu||\theta) = 0 \]  

(49)
Since

\[
\inf_{\mu \in \mathcal{M}} \{ A(\theta) + A^*(\mu) - \langle \theta, \mu \rangle \} = \inf_{\mu \in \mathcal{M}} D(\mu || \theta) = 0
\]
Since

\[ \inf_{\mu \in \mathcal{M}} \{ A(\theta) + A^*(\mu) - \langle \theta, \mu \rangle \} = \inf_{\mu \in \mathcal{M}} D(\mu||\theta) = 0 \]  

Thus, solving the mean-field variational problem of:

\[ \max_{\mu \in \mathcal{M}_F(G)} \{ \langle \mu, \theta \rangle - A^*_F(\mu) \} \]  

is identical to minimizing KL Divergence \( D(\mu||\theta) \) subject to constraint \( \mu \in \mathcal{M}_F(G) \).
Mean field, KL-Divergence, Exponential Model Families

Since

$$\inf_{\mu \in \mathcal{M}} \left\{ A(\theta) + A^*(\mu) - \langle \theta, \mu \rangle \right\} = \inf_{\mu \in \mathcal{M}} D(\mu||\theta) = 0$$  \hspace{1cm} (50)

Thus, solving the mean-field variational problem of:

$$\max_{\mu \in \mathcal{M}_F(G)} \left\{ \langle \mu, \theta \rangle - A^*_F(\mu) \right\}$$  \hspace{1cm} (41)

is identical to minimizing KL Divergence $D(\mu||\theta)$ subject to constraint $\mu \in \mathcal{M}_F(G)$.

I.e., mean field can be seen as finding the best approximation, in terms of this particular KL-divergence, to $p_\theta$, over a family of “nice” distributions $\mathcal{M}_F(G)$. 

Prof. Jeff Bilmes
EE12A/Fall 2011/Graphical Models – Lecture 16 - Nov 30th, 2011
A classic example of mean-field (goes back to statistical physics)
Naive Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)
- Mean parameters for Ising: \( \mu_s = \mathbb{E}[X_s] = p(X_s = 1) \),
  \( \mu_{st} = \mathbb{E}[X_sX_t] = p(X_s = 1, X_t = 1) \), thus \( \mu \in \mathbb{R}^{\left|V\right|+\left|E\right|} \).
Naive Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)
- Mean parameters for Ising: 
  \[ \mu_s = \mathbb{E}[X_s] = p(X_s = 1), \]
  \[ \mu_{st} = \mathbb{E}[X_s X_t] = p(X_s = 1, X_t = 1), \text{ thus } \mu \in \mathbb{R}^{V+|E|}. \]

- Let \( F_0 = (V, \emptyset) \) be our mean field approximation family. Thus,

\[
\mathcal{M}_{F_0}(G) = \left\{ \mu \in \mathbb{R}^{V+|E|} | 0 \leq \mu_s \leq 1 \ \forall s \in V, \text{ and } \mu_{st} = \mu_s \mu_t \ \forall \right\}
\]
Naive Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)

- Mean parameters for Ising: \( \mu_s = \mathbb{E}[X_s] = p(X_s = 1) \),
  \( \mu_{st} = \mathbb{E}[X_sX_t] = p(X_s = 1, X_t = 1) \), thus \( \mu \in \mathbb{R}^{|V|+|E|} \).

- Let \( F_0 = (V, \emptyset) \) be our mean field approximation family. Thus,

\[
\mathcal{M}_{F_0}(G) = \left\{ \mu \in \mathbb{R}^{|V|+|E|} | 0 \leq \mu_s \leq 1 \ \forall s \in V, \text{ and } \mu_{st} = \mu_s \mu_t \ \forall \right\}
\]

- Key is that for \( \mu \in \mathcal{M}_{F_0}(G) \), dual is not hard to calculate, that is

\[
-A^*_F(\mu) = \sum_{s \in V} H_s(\mu_s)
\]  \hspace{1cm} (51)

which are sum of unary entropy terms, very cheap.
Naive Mean field for Ising Model

- A classic example of mean-field (goes back to statistical physics)
- Mean parameters for Ising: $\mu_s = \mathbb{E}[X_s] = p(X_s = 1)$, 
  $\mu_{st} = \mathbb{E}[X_s X_t] = p(X_s = 1, X_t = 1)$, thus $\mu \in \mathbb{R}^{\vert V \vert + \vert E \vert}$.
- Let $F_0 = (V, \emptyset)$ be our mean field approximation family. Thus,

  $$\mathcal{M}_{F_0}(G) = \left\{ \mu \in \mathbb{R}^{\vert V \vert + \vert E \vert} \mid 0 \leq \mu_s \leq 1 \; \forall s \in V, \; \text{and} \; \mu_{st} = \mu_s \mu_t \; \forall \right\}$$

- Key is that for $\mu \in \mathcal{M}_{F_0}(G)$, dual is not hard to calculate, that is

  $$-A^*_F(\mu) = \sum_{s \in V} H_s(\mu_s) \tag{51}$$

  which are sum of unary entropy terms, very cheap.
- Moreover, polytope for $M_{F_0}(G)$ is also very simple, namely the hypercube $[0, 1]^m$. 
Naive Mean field for Ising Model

- We get variational lower bound problem

\[
A(\theta) \geq \max_{(\mu_1, \ldots, \mu_m) \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\}
\]

(52)
Naive Mean field for Ising Model

- We get variational lower bound problem

\[ A(\theta) \geq \max_{(\mu_1, \ldots, \mu_m) \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\} \]  \hspace{1cm} (52)

- Once again, we have a non-convex problem.
Naive Mean field for Ising Model

- We get variational lower bound problem

\[
A(\theta) \geq \max_{(\mu_1, \ldots, \mu_m) \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\}
\] (52)

- Once again, we have a non-convex problem.

- One way to optimize is to do coordinate ascent (given otherwise fixed vector, optimize one value at a time).
Naive Mean field for Ising Model

- We get variational lower bound problem

\[
A(\theta) \geq \max_{(\mu_1, \ldots, \mu_m) \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\}
\]  \hspace{1cm} (52)

- Once again, we have a non-convex problem.
- One way to optimize is to do coordinate ascent (given otherwise fixed vector, optimize one value at a time).
- If each coordinate optimization is optimal, we’ll get a stationary point.
Naive Mean field for Ising Model

- coordinate ascent: choose some \( s \) and optimize \( \mu_s \) fixing all \( \mu_t \) for \( t \neq s \).
Naive Mean field for Ising Model

- **coordinate ascent**: choose some $s$ and optimize $\mu_s$ fixing all $\mu_t$ for $t \neq s$.

- **Taking derivatives w.r.t. $\mu_s$**, we get the following update rule for element $\mu_s$

\[
\mu_s \leftarrow \theta \left( \theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t \right)
\]  

where $\sigma(z) = [1 + \exp(-z)]^{-1}$ is the sigmoid (logistic) function.
Naive Mean field for Ising Model

- coordinate ascent: choose some $s$ and optimize $\mu_s$ fixing all $\mu_t$ for $t \neq s$.

- Taking derivatives w.r.t. $\mu_s$, we get the following update rule for element $\mu_s$

$$\mu_s \leftarrow \theta \left( \theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t \right)$$  \hspace{1cm} (53)

where $\sigma(z) = \left[ 1 + \exp(-z) \right]^{-1}$ is the sigmoid (logistic) function.

- This is the standard mean-field update that is quite well known, but derived from coordinate ascent optimization of a variational perspective of the problem.
coordinate ascent: choose some \( s \) and optimize \( \mu_s \) fixing all \( \mu_t \) for \( t \neq s \).

Taking derivatives w.r.t. \( \mu_s \), we get the following update rule for element \( \mu_s \)

\[
\mu_s \leftarrow \sigma \left( \theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t \right) \tag{53}
\]

where \( \sigma(z) = \frac{1}{1 + \exp(-z)} \) is the sigmoid (logistic) function.

This is the standard mean-field update that is quite well known, but derived from coordinate ascent optimization of a variational perspective of the problem.

The variational approach indeed seems quite general and powerful.
Lack of Convexity example

- Consider simple two variable example \((X_1, X_2), X_i \in \{-1, +1\}\).
Lack of Convexity example

- Consider simple two variable example \((X_1, X_2), \; X_i \in \{-1, +1\}\).
- Exponential family form

\[
p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2)
\]

having mean parameters \(\mu_i = \mathbb{E}[X_i]\) and \(\mu_{12} = \mathbb{E}[X_1X_2]\).
Lack of Convexity example

- Consider simple two variable example \((X_1, X_2), X_i \in \{-1, +1\}\).
- Exponential family form
  \[
  p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2)
  \]
  having mean parameters \(\mu_i = \mathbb{E}[X_i]\) and \(\mu_{12} = \mathbb{E}[X_1 X_2]\).
- Impose constraint \(\mu_{12} = \mu_1 \mu_2\), we get mean field objective
  \[
  f(\mu_1, \mu_2; \theta) = \theta_{12} \mu_1 \mu_2 + \theta_1 \mu_1 + \theta_2 \mu_2 + H(\mu_1) + H(\mu_2)
  \]
  where \(H(\mu_i) = -\frac{1}{2} (1 + \mu_i) \log \frac{1}{2} (1 + \mu_i) - \frac{1}{2} (1 - \mu_i) \log \frac{1}{2} (1 - \mu_i)\)
Lack of Convexity example

- Consider simple two variable example \((X_1, X_2), X_i \in \{-1, +1\}\).
- Exponential family form
\[
p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2)
\] (54)

having mean parameters \(\mu_i = \mathbb{E}[X_i]\) and \(\mu_{12} = \mathbb{E}[X_1 X_2]\).
- Impose constraint \(\mu_{12} = \mu_1 \mu_2\), we get mean field objective
\[
f(\mu_1, \mu_2; \theta) = \theta_{12} \mu_1 \mu_2 + \theta_1 \mu_1 + \theta_2 \mu_2 + H(\mu_1) + H(\mu_2)
\] (55)

where \(H(\mu_i) = -\frac{1}{2}(1 + \mu_i) \log \frac{1}{2}(1 + \mu_i) - \frac{1}{2}(1 - \mu_i) \log \frac{1}{2}(1 - \mu_i)\)
- Consider sub-models of the form:
\[
(\theta_1, \theta_2, \theta_{12}) = \left(0, 0, \frac{1}{4} \log \frac{q}{1 - q}\right) \triangleq \theta(q)
\] (56)

where \(q \in (0, 1)\) is a parameter such that, for any \(q\) we have \(\mathbb{E}[X_i] = 0\).
Also, \(q = p(X_1 = X_2)\).
Lack of Convexity example

- Consider simple two variable example \((X_1, X_2), X_i \in \{-1, +1\}\).
- Exponential family form

\[
p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2)
\]

having mean parameters \(\mu_i = \mathbb{E}[X_i]\) and \(\mu_{12} = \mathbb{E}[X_1 X_2]\).
- Impose constraint \(\mu_{12} = \mu_1 \mu_2\), we get mean field objective

\[
f(\mu_1, \mu_2; \theta) = \theta_{12} \mu_1 \mu_2 + \theta_1 \mu_1 + \theta_2 \mu_2 + H(\mu_1) + H(\mu_2)
\]

where \(H(\mu_i) = -\frac{1}{2}(1 + \mu_i) \log \frac{1}{2}(1 + \mu_i) - \frac{1}{2}(1 - \mu_i) \log \frac{1}{2}(1 - \mu_i)\)
- Consider sub-models of the form:

\[
(\theta_1, \theta_2, \theta_{12}) = \left(0, 0, \frac{1}{4} \log \frac{q}{1 - q}\right) \triangleq \theta(q)
\]

where \(q \in (0, 1)\) is a parameter such that, for any \(q\) we have \(\mathbb{E}[X_i] = 0\). Also, \(q = p(X_1 = X_2)\).
- Is mean field objective convex for all \(q\)?
Lack of Convexity example

- For $q = 0.5$, objective $f(\mu_1, \mu_2; \theta(0.5))$ has global maximum at $(\mu_1, \mu_2) = (0, 0)$ so mean field is exact and convex. This corresponds to $p(X_1 = X_2) = 0$. 
Lack of Convexity example

- For $q = 0.5$, objective $f(\mu_1, \mu_2; \theta(0.5))$ has **global** maximum at $(\mu_1, \mu_2) = (0, 0)$ so mean field is exact and convex. This corresponds to $p(X_1 = X_2) = 0$.
- When $q$ gets small, $f$ becomes non-convex, e.g., has multiple modes in figure.
Lack of Convexity example

- For $q = 0.5$, objective $f(\mu_1, \mu_2; \theta(0.5))$ has global maximum at $(\mu_1, \mu_2) = (0, 0)$ so mean field is exact and convex. This corresponds to $p(X_1 = X_2) = 0$.

- When $q$ gets small, $f$ becomes non-convex, e.g., has multiple modes in figure.
\[ H(\tau) + \sum_{\ell=1}^{d_1} \left[ H(\tau, \bar{z}_\ell) - H(\tau) \right] \]

\[ \sum_{\tau} H(\tau) \]

\[ H(\tau, \bar{z}_\ell) = \sum_{\ell \notin \{S_1\}} H(\tau) + H(\tau_{S_1}) \]

\[ \ell = (S_1) \]

\[ H(\tau, \bar{z}_\ell) - H(\ell T) \]

\[ = H(\tau_{S_1}) - H(\tau) - H(\tau_T) \]

\[ = -I(\tau; \tau_T) \]
this material comes from the Wainwright and Jordan book.