CSci 8980: Advanced Topics in Graphical Models

Variational Inference

Instructor: Arindam Banerjee

October 17, 2007

Directed Graphical Models

- Graph $G = (V, E)$

Directed Graphical Models

- Graph $G = (V, E)$
- Each vertex is a random variable

Directed Graphical Models

- Graph $G = (V, E)$
- Each vertex is a random variable
- $\pi(s)$ denote the set of all parents of $s \in V$

Directed Graphical Models

- Graph $G = (V, E)$
- Each vertex is a random variable
- $\pi(s)$ denote the set of all parents of $s \in V$
- The joint distribution

$$p(\mathbf{x}) = \prod_{s \in V} p(x_s | x_{\pi(s)})$$

Undirected Graphical Models

- Distribution factorizes over cliques of the graph

Undirected Graphical Models

- Distribution factorizes over cliques of the graph
- Let $\psi_C : \mathcal{X}^n \mapsto \mathbb{R}_+$ be a function over clique C

Undirected Graphical Models

- Distribution factorizes over cliques of the graph
- Let $\psi_C : \mathcal{X}^n \mapsto \mathbb{R}_+$ be a function over clique C
- The joint distribution

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(x_C)$$

Undirected Graphical Models

- Distribution factorizes over cliques of the graph
- Let $\psi_C : \mathcal{X}^n \mapsto \mathbb{R}_+$ be a function over clique C
- The joint distribution

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(x_C)$$

- Z ensures the distribution is normalized

Undirected Graphical Models

- Distribution factorizes over cliques of the graph
- Let $\psi_C : \mathcal{X}^n \mapsto \mathbb{R}_+$ be a function over clique C
- The joint distribution

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(x_C)$$

- Z ensures the distribution is normalized
- Known as a Markov random field

Basics (Review)

- For any $h : \mathcal{X}^n \mapsto \mathbb{R}_+$, define measure ν as $d\nu = h(x)dx$

Basics (Review)

- For any $h : \mathcal{X}^n \mapsto \mathbb{R}_+$, define measure ν as $d\nu = h(x)dx$
- Let $t = \{\phi_\alpha | \alpha \in \mathcal{I}\}$ be a set of sufficient statistics

Basics (Review)

- For any $h : \mathcal{X}^n \mapsto \mathbb{R}_+$, define measure ν as $d\nu = h(x)dx$
- Let $t = \{\phi_\alpha | \alpha \in \mathcal{I}\}$ be a set of sufficient statistics
- Let $\theta = \{\theta_\alpha | \alpha \in \mathcal{I}\}$ be the natural parameters

Basics (Review)

- For any $h : \mathcal{X}^n \mapsto \mathbb{R}_+$, define measure ν as $d\nu = h(x)dx$
- Let $t = \{\phi_\alpha | \alpha \in \mathcal{I}\}$ be a set of sufficient statistics
- Let $\theta = \{\theta_\alpha | \alpha \in \mathcal{I}\}$ be the natural parameters
- The family of density functions w.r.t. $d\nu$

$$p(x; \theta) = \exp(\langle \theta, t(x) \rangle - \psi(\theta))$$

where

$$\psi(\theta) = \log \int_x \exp(\langle \theta, t(x) \rangle) \nu(dx)$$

Graphical Models as Exponential Families

- Graphical models are described as products of functions

Graphical Models as Exponential Families

- Graphical models are described as products of functions
- Products are additive in the exponent

Graphical Models as Exponential Families

- Graphical models are described as products of functions
- Products are additive in the exponent
- Ising Model:

Graphical Models as Exponential Families

- Graphical models are described as products of functions
- Products are additive in the exponent
- Ising Model:
  - Each vertex is a Bernoulli random variable

Graphical Models as Exponential Families

- Graphical models are described as products of functions
- Products are additive in the exponent
- Ising Model:
  - Each vertex is a Bernoulli random variable
  - Components x_s, x_t interact only if there is an edge

Graphical Models as Exponential Families

- Graphical models are described as products of functions
- Products are additive in the exponent
- Ising Model:
  - Each vertex is a Bernoulli random variable
  - Components x_s, x_t interact only if there is an edge
  - The joint distribution

$$p(x;\theta) = \exp\left(\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t - \psi(\theta)\right)$$

Graphical Models as Exponential Families

- Graphical models are described as products of functions
- Products are additive in the exponent
- Ising Model:
  - Each vertex is a Bernoulli random variable
  - Components x_s, x_t interact only if there is an edge
  - The joint distribution

$$p(x;\theta) = \exp\left(\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - \psi(\theta)\right)$$

  - Dimensionality of the model is $d = n + |E|$

Graphical Models as Exponential Families

- Graphical models are described as products of functions
- Products are additive in the exponent
- Ising Model:
  - Each vertex is a Bernoulli random variable
  - Components x_s, x_t interact only if there is an edge
  - The joint distribution

$$p(x;\theta) = \exp\left(\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t - \psi(\theta)\right)$$

  - Dimensionality of the model is $d = n + |E|$
  - It is a regular exponential family, with $\Theta = \mathbb{R}^d$

Graphical Models as Exponential Families (Contd.)

- Latent Dirichlet Allocation: For a single document

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

$$\propto \exp\left(\sum_{i=1}^{k} (\alpha_i - 1) \log \theta_i + \sum_{n=1}^{N} \sum_{i=1}^{k} \mathbb{I}_i(z_n) \log \theta_i + \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} \mathbb{I}_i[z_n]\mathbb{I}\right.$$

Graphical Models as Exponential Families (Contd.)

- Latent Dirichlet Allocation: For a single document

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

$$\propto \exp\left(\sum_{i=1}^{k} (\alpha_i - 1) \log \theta_i + \sum_{n=1}^{N} \sum_{i=1}^{k} \mathbb{I}_i(z_n) \log \theta_i + \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} \mathbb{I}_i[z_n] \mathbb{I}\right.$$

- The sufficient statistics consists of:

$$\{\log \theta_i, [i]_1^k\} \quad \{\mathbb{I}_i[z_n] \log \theta_i, [i]_1^k, [n]_1^N\} \quad \{\mathbb{I}_i[z_n] \mathbb{I}_j[w_n], [i]_1^k, [n]_1^N, [j]_1^V\}$$

Properties of the Cumulant ψ

- ψ is the cumulant or log-partition function

Properties of the Cumulant ψ

- ψ is the cumulant or log-partition function
- $\psi(\theta)$ is C^∞ on Θ

Properties of the Cumulant ψ

- ψ is the cumulant or log-partition function
- $\psi(\theta)$ is C^∞ on Θ
- Its derivatives gives the moments of θ

$$\frac{\partial \psi(\theta)}{\partial \theta_\alpha} = E_\theta[t_\alpha(x)]$$

$$\frac{\partial^2 \psi(\theta)}{\partial \theta_\alpha \partial \theta(\beta)} = E_\theta[t_\alpha(x) t_\beta(x)] - E_\theta[t_\alpha(x)] E_\theta[t_\beta(x)]$$

Properties of the Cumulant ψ

- ψ is the cumulant or log-partition function
- $\psi(\theta)$ is C^∞ on Θ
- Its derivatives gives the moments of θ

$$\frac{\partial \psi(\theta)}{\partial \theta_\alpha} = E_\theta[t_\alpha(x)]$$

$$\frac{\partial^2 \psi(\theta)}{\partial \theta_\alpha \partial \theta(\beta)} = E_\theta[t_\alpha(x) t_\beta(x)] - E_\theta[t_\alpha(x)] E_\theta[t_\beta(x)]$$

- ψ is a convex function, strictly convex if $t(x)$ is minimal

Properties of the Cumulant ψ (Contd.)

- The set of mean parameters

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d | \exists p(.) s.t. \int t(x) p(x) \nu(dx) = \mu \right\}$$

Properties of the Cumulant ψ (Contd.)

- The set of mean parameters

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d | \exists p(.) s.t. \int t(x)p(x)\nu(dx) = \mu \right\}$$

- Consider the mapping $\Lambda : \Theta \mapsto \mathcal{M}$ as

$$\Lambda(\theta) = E_\theta[t(x)] = \int_x t(x)p(x;\theta)\nu(dx)$$

Properties of the Cumulant ψ (Contd.)

- The set of mean parameters

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d | \exists p(.) s.t. \int t(x) p(x) \nu(dx) = \mu \right\}$$

- Consider the mapping $\Lambda : \Theta \mapsto \mathcal{M}$ as

$$\Lambda(\theta) = E_\theta[t(x)] = \int_x t(x) p(x;\theta) \nu(dx)$$

- If t is minimal, Λ is one-to-one

Properties of the Cumulant ψ (Contd.)

- The set of mean parameters

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d | \exists p(.) s.t. \int t(x) p(x) \nu(dx) = \mu \right\}$$

- Consider the mapping $\Lambda : \Theta \mapsto \mathcal{M}$ as

$$\Lambda(\theta) = E_\theta[t(x)] = \int_x t(x) p(x;\theta) \nu(dx)$$

- If t is minimal, Λ is one-to-one
- Further, Λ is onto the (relative) interior of $\mathcal{M}$

Fenchel-Legendre Conjugacy

- The conjugate dual function

$$\psi^*(\mu) = \sup_{\theta \in \Theta} \{\langle \mu, \theta \rangle - \psi(\theta)\}$$

Fenchel-Legendre Conjugacy

- The conjugate dual function

$$\psi^*(\mu) = \sup_{\theta \in \Theta} \{\langle \mu, \theta \rangle - \psi(\theta)\}$$

- The (Bolzmann-Shannon) entropy of $p(x; \theta)$ w.r.t. ν is

$$H(p(x;\theta)) = -\int_x p(x;\theta) \log p(x;\theta) \nu(dx) = -E_\theta[\log p(x;\theta)]$$

Fenchel-Legendre Conjugacy

- The conjugate dual function

$$\psi^*(\mu) = \sup_{\theta \in \Theta} \{ \langle \mu, \theta \rangle - \psi(\theta) \}$$

- The (Bolzmann-Shannon) entropy of $p(x;\theta)$ w.r.t. ν is

$$H(p(x;\theta)) = -\int_x p(x;\theta) \log p(x;\theta) \nu(dx) = -E_\theta[\log p(x;\theta)]$$

- If $\mu \in \operatorname{ri} \mathcal{M}$, then

$$\psi^*(\mu) = -H(p(x;\theta(\mu)))$$

Fenchel-Legendre Conjugacy

- The conjugate dual function

$$\psi^*(\mu) = \sup_{\theta \in \Theta} \{\langle \mu, \theta \rangle - \psi(\theta)\}$$

- The (Bolzmann-Shannon) entropy of $p(x;\theta)$ w.r.t. ν is

$$H(p(x;\theta)) = -\int_x p(x;\theta) \log p(x;\theta) \nu(dx) = -E_\theta[\log p(x;\theta)]$$

- If $\mu \in \operatorname{ri} \mathcal{M}$, then

$$\psi^*(\mu) = -H(p(x;\theta(\mu)))$$

- In terms of the dual, ψ has a variational representation

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

Main Issues

- Key problems:

Main Issues

- Key problems:
  - Computation of the cumulant function $\psi(\theta)$

Main Issues

- Key problems:
  - Computation of the cumulant function $\psi(\theta)$
  - Computation of the mean parameter $\mu = E_\theta[t(x)]$

Main Issues

- Key problems:
  - Computation of the cumulant function $\psi(\theta)$
  - Computation of the mean parameter $\mu = E_\theta[t(x)]$
- The key equation for both problems

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

Main Issues

- Key problems:
  - Computation of the cumulant function $\psi(\theta)$
  - Computation of the mean parameter $\mu = E_\theta[t(x)]$
- The key equation for both problems

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

- For all $\theta \in \Theta$, the supremum is attained by $\mu \in \operatorname{ri} \mathcal{M}$

$$\mu = E_\theta[t(x)] = \int_x t(x) p(x; \theta) \nu(dx)$$

Main Issues

- Key problems:
  - Computation of the cumulant function $\psi(\theta)$
  - Computation of the mean parameter $\mu = E_\theta[t(x)]$
- The key equation for both problems

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

- For all $\theta \in \Theta$, the supremum is attained by $\mu \in \operatorname{ri} \mathcal{M}$

$$\mu = E_\theta[t(x)] = \int_x t(x) p(x; \theta) \nu(dx)$$

- Two primary challenges

Main Issues

- Key problems:
  - Computation of the cumulant function $\psi(\theta)$
  - Computation of the mean parameter $\mu = E_\theta[t(x)]$
- The key equation for both problems

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

- For all $\theta \in \Theta$, the supremum is attained by $\mu \in \operatorname{ri} \mathcal{M}$

$$\mu = E_\theta[t(x)] = \int_x t(x) p(x; \theta) \nu(dx)$$

- Two primary challenges
  - Set $\mathcal{M}$ is difficult to characterize

Main Issues

- Key problems:
  - Computation of the cumulant function $\psi(\theta)$
  - Computation of the mean parameter $\mu = E_\theta[t(x)]$
- The key equation for both problems

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

- For all $\theta \in \Theta$, the supremum is attained by $\mu \in \operatorname{ri} \mathcal{M}$

$$\mu = E_\theta[t(x)] = \int_x t(x) p(x; \theta) \nu(dx)$$

- Two primary challenges
  - Set $\mathcal{M}$ is difficult to characterize
  - Function ψ^* lacks an explicit definition

Mean Parameters

- $\mathcal{M}$ has the following properties

Mean Parameters

- $\mathcal{M}$ has the following properties
  - $\mathcal{M}$ is full-dimensional if t is minimal

Mean Parameters

- $\mathcal{M}$ has the following properties
  - $\mathcal{M}$ is full-dimensional if t is minimal
  - $\mathcal{M}$ is bounded iff $\Theta = \mathbb{R}^d$ and ψ is Lipschitz

Mean Parameters

- $\mathcal{M}$ has the following properties
  - $\mathcal{M}$ is full-dimensional if t is minimal
  - $\mathcal{M}$ is bounded iff $\Theta = \mathbb{R}^d$ and ψ is Lipschitz
- Example: Mutinomial random vector $x \in \mathcal{X}^n$

Mean Parameters

- $\mathcal{M}$ has the following properties
  - $\mathcal{M}$ is full-dimensional if t is minimal
  - $\mathcal{M}$ is bounded iff $\Theta = \mathbb{R}^d$ and ψ is Lipschitz
- Example: Mutinomial random vector $x \in \mathcal{X}^n$
  - The set $\mathcal{M}$ is a polytope

$$\mathcal{M} = \{\mu \in \mathbb{R}^d | \langle a_j, \mu \rangle \leq b_j, \forall j \in \mathcal{J}\}$$

Mean Parameters

- $\mathcal{M}$ has the following properties
  - $\mathcal{M}$ is full-dimensional if t is minimal
  - $\mathcal{M}$ is bounded iff $\Theta = \mathbb{R}^d$ and ψ is Lipschitz
- Example: Mutinomial random vector $x \in \mathcal{X}^n$
  - The set $\mathcal{M}$ is a polytope

$$\mathcal{M} = \{\mu \in \mathbb{R}^d | \langle a_j, \mu \rangle \leq b_j, \forall j \in \mathcal{J}\}$$

  - Index set $\mathcal{J}$ is finite, but can be large

Mean Parameters

- $\mathcal{M}$ has the following properties
  - $\mathcal{M}$ is full-dimensional if t is minimal
  - $\mathcal{M}$ is bounded iff $\Theta = \mathbb{R}^d$ and ψ is Lipschitz
- Example: Mutinomial random vector $x \in \mathcal{X}^n$
  - The set $\mathcal{M}$ is a polytope

$$\mathcal{M} = \{\mu \in \mathbb{R}^d | \langle a_j, \mu \rangle \leq b_j, \forall j \in \mathcal{J}\}$$

  - Index set $\mathcal{J}$ is finite, but can be large
- Facets of the polytope can grow very fast with n

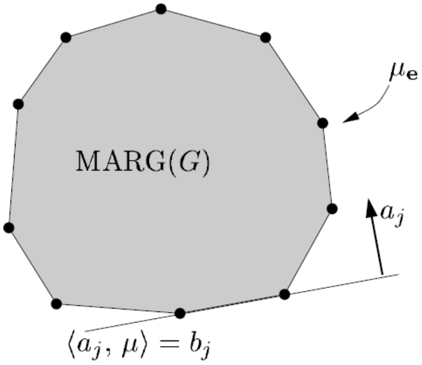
Mean Parameters

- $\mathcal{M}$ has the following properties
  - $\mathcal{M}$ is full-dimensional if t is minimal
  - $\mathcal{M}$ is bounded iff $\Theta = \mathbb{R}^d$ and ψ is Lipschitz
- Example: Mutinomial random vector $x \in \mathcal{X}^n$
  - The set $\mathcal{M}$ is a polytope

$$\mathcal{M} = \{\mu \in \mathbb{R}^d | \langle a_j, \mu \rangle \leq b_j, \forall j \in \mathcal{J}\}$$

  - Index set $\mathcal{J}$ is finite, but can be large
- Facets of the polytope can grow very fast with n
- A complete graph with $n = 7$ has more than 2×10^8 facets

Mean Parameters (Contd.)



Dual Function

- ψ^* is the negative entropy

Dual Function

- ψ^* is the negative entropy
- Typically, does not have an explicit closed form

Dual Function

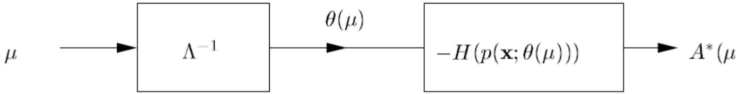
- ψ^* is the negative entropy
- Typically, does not have an explicit closed form
- In general, can be specified as a composition of two functions

Dual Function

- ψ^* is the negative entropy
- Typically, does not have an explicit closed form
- In general, can be specified as a composition of two functions
  - Compute an inverse image $\theta(\mu)$ using $\Lambda^{-1}(\mu)$

Dual Function

- ψ^* is the negative entropy
- Typically, does not have an explicit closed form
- In general, can be specified as a composition of two functions
  - Compute an inverse image $\theta(\mu)$ using $\Lambda^{-1}(\mu)$
  - Compute the negative entropy of $p(x; \theta(\mu))$



Tractable Families

- Based on the key equation

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

Tractable Families

- Based on the key equation

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

- Mean field focuses on tractable distributions

Tractable Families

- Based on the key equation

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

- Mean field focuses on tractable distributions
- Let $H \subseteq G$ on which exact calculations are feasible

Tractable Families

- Based on the key equation

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

- Mean field focuses on tractable distributions
- Let $H \subseteq G$ on which exact calculations are feasible
- $\mathcal{I}(H)$ be the indices of cliques in H

Tractable Families

- Based on the key equation

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

- Mean field focuses on tractable distributions
- Let $H \subseteq G$ on which exact calculations are feasible
- $\mathcal{I}(H)$ be the indices of cliques in H
- Natural parameters for distributions corresponding to H

$$\mathcal{E}(H) = \{\theta \in \Theta | \theta_\alpha = 0, \forall \alpha \in \mathcal{I} \setminus \mathcal{I}(H)\}$$

Tractable Families (Contd.)

- Simple tractable subgraph is $H = (V, \emptyset)$

Tractable Families (Contd.)

- Simple tractable subgraph is $H = (V, \emptyset)$
- Natural parameters belong to the subspace

$$\mathcal{E}(H) = \{\theta \in \Theta | \theta_{st} = 0, \ \forall (s, t) \in E\}$$

Tractable Families (Contd.)

- Simple tractable subgraph is $H = (V, \emptyset)$
- Natural parameters belong to the subspace

$$\mathcal{E}(H) = \{\theta \in \Theta | \theta_{st} = 0, \ \forall (s,t) \in E\}$$

- Corresponding distribution $p(x;\theta) = \prod_{s \in V} p(x_s;\theta_s)$

Tractable Families (Contd.)

- Simple tractable subgraph is $H = (V, \emptyset)$
- Natural parameters belong to the subspace

$$\mathcal{E}(H) = \{\theta \in \Theta | \theta_{st} = 0, \ \forall (s,t) \in E\}$$

- Corresponding distribution $p(x;\theta) = \prod_{s \in V} p(x_s; \theta_s)$
- Structured approximation using spanning tree $T = (V, E(T))$

Tractable Families (Contd.)

- Simple tractable subgraph is $H = (V, \emptyset)$
- Natural parameters belong to the subspace

$$\mathcal{E}(H) = \{\theta \in \Theta | \theta_{st} = 0, \ \forall (s,t) \in E\}$$

- Corresponding distribution $p(x;\theta) = \prod_{s \in V} p(x_s;\theta_s)$
- Structured approximation using spanning tree $T = (V, E(T))$
- Natural parameters belong to the subspace

$$\mathcal{E}(T) = \{\theta \in \Theta | \theta_{st} = 0, \ \forall (s,t) \in E(T)\}$$

Tractable Families (Contd.)

- Simple tractable subgraph is $H = (V, \emptyset)$
- Natural parameters belong to the subspace

$$\mathcal{E}(H) = \{\theta \in \Theta | \theta_{st} = 0, \ \forall (s,t) \in E\}$$

- Corresponding distribution $p(x;\theta) = \prod_{s \in V} p(x_s; \theta_s)$
- Structured approximation using spanning tree $T = (V, E(T))$
- Natural parameters belong to the subspace

$$\mathcal{E}(T) = \{\theta \in \Theta | \theta_{st} = 0, \ \forall (s,t) \in E(T)\}$$

- For a subgraph H, the set of realizable mean parameters

$$\mathcal{M}_{tract}(G; H) = \{\mu \in \mathbb{R}^d | \mu = E_\theta[t(x)], \theta \in \mathcal{E}(H)\}$$

Tractable Families (Contd.)

- Simple tractable subgraph is $H = (V, \emptyset)$
- Natural parameters belong to the subspace

$$\mathcal{E}(H) = \{\theta \in \Theta | \theta_{st} = 0, \ \forall (s,t) \in E\}$$

- Corresponding distribution $p(x;\theta) = \prod_{s \in V} p(x_s;\theta_s)$
- Structured approximation using spanning tree $T = (V, E(T))$
- Natural parameters belong to the subspace

$$\mathcal{E}(T) = \{\theta \in \Theta | \theta_{st} = 0, \ \forall (s,t) \in E(T)\}$$

- For a subgraph H, the set of realizable mean parameters

$$\mathcal{M}_{tract}(G;H) = \{\mu \in \mathbb{R}^d | \mu = E_\theta[t(x)], \theta \in \mathcal{E}(H)\}$$

- The inclusion $\mathcal{M}_{tract}(G;H) \subseteq \mathcal{M}(G)$ always holds

Lower Bounds

- For any $\mu \in \operatorname{ri} \mathcal{M}$, $\psi(\theta) \geq \langle \theta, \mu \rangle - \psi^*(\mu)$

Lower Bounds

- For any $\mu \in \operatorname{ri} \mathcal{M}$, $\psi(\theta) \geq \langle \theta, \mu \rangle - \psi^*(\mu)$
- Alternative proof using Jensen's inequality

$$\begin{aligned}
\psi(\theta) &= \log \int_x p(x;\theta) \frac{\exp(\langle \theta, t(x) \rangle)}{p(x;\theta)} \nu(dx) \\
&\geq \int_x p(x;\theta) \left[\langle \theta, t(x) \rangle - \log p(x;\theta(\mu)) \right] \nu(dx) \\
&= \langle \theta, \mu \rangle - \psi^*(\mu)
\end{aligned}$$

Lower Bounds

- For any $\mu \in \text{ri}\,\mathcal{M}$, $\psi(\theta) \geq \langle \theta, \mu \rangle - \psi^*(\mu)$
- Alternative proof using Jensen's inequality

$$\begin{aligned} \psi(\theta) &= \log \int_x p(x;\theta) \frac{\exp(\langle \theta, t(x) \rangle)}{p(x;\theta)} \nu(dx) \\ &\geq \int_x p(x;\theta) \left[\langle \theta, t(x) \rangle - \log p(x;\theta(\mu)) \right] \nu(dx) \\ &= \langle \theta, \mu \rangle - \psi^*(\mu) \end{aligned}$$

- In general, ψ^* does not have closed form

Lower Bounds

- For any $\mu \in \mathrm{ri}\,\mathcal{M}$, $\psi(\theta) \geq \langle \theta, \mu \rangle - \psi^*(\mu)$
- Alternative proof using Jensen's inequality

$$\begin{aligned} \psi(\theta) &= \log \int_x p(x;\theta) \frac{\exp(\langle \theta, t(x) \rangle)}{p(x;\theta)} \nu(dx) \\ &\geq \int_x p(x;\theta) \left[\langle \theta, t(x) \rangle - \log p(x;\theta(\mu)) \right] \nu(dx) \\ &= \langle \theta, \mu \rangle - \psi^*(\mu) \end{aligned}$$

- In general, ψ^* does not have closed form
- Since ψ^*_H has an explicit form, solve approximation

$$\sup_{\mu \in \mathcal{M}_{tract}} \{ \langle \mu, \theta \rangle - \psi^*_H(\mu) \}$$

Naive Mean Field

- Chooses a fully factorized distribution to approximate the original distribution

Naive Mean Field

- Chooses a fully factorized distribution to approximate the original distribution
- We will study Ising model as an example

Naive Mean Field

- Chooses a fully factorized distribution to approximate the original distribution
- We will study Ising model as an example
- Approximate G by fully disconnected graph H_0 with no edges

Naive Mean Field

- Chooses a fully factorized distribution to approximate the original distribution
- We will study Ising model as an example
- Approximate G by fully disconnected graph H_0 with no edges
- Then, the mean parameter set

$$\mathcal{M}_{tract} = \{(\mu_s, \mu_{st}) | 0 \leq \mu_s \leq 1, \mu_{st} = \mu_s \mu_t\}$$

Naive Mean Field

- Chooses a fully factorized distribution to approximate the original distribution
- We will study Ising model as an example
- Approximate G by fully disconnected graph H_0 with no edges
- Then, the mean parameter set

$$\mathcal{M}_{tract} = \{(\mu_s, \mu_{st}) | 0 \leq \mu_s \leq 1, \mu_{st} = \mu_s \mu_t\}$$

- The negative entropy of the product distribution is

$$\psi^*_{H_0}(\mu) = \sum_{s \in V} [\mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s)]$$

Naive Mean Field (Contd.)

- The naive mean field problem takes the form

$$\max_{\mu \in \mathcal{M}_{tract}} \{\langle \mu, \theta \rangle - \psi^*_{H_0}(\mu)\}$$

Naive Mean Field (Contd.)

- The naive mean field problem takes the form

$$\max_{\mu \in \mathcal{M}_{tract}} \left\{ \langle \mu, \theta \rangle - \psi^*_{H_0}(\mu) \right\}$$

- Using $\mu_{st} = \mu_s \mu_t$, we get the reduced problem

$$\max_{\{\mu_s\} \in [0,1]^n} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t - \sum_{s \in V} [\mu_s \log \mu_s + (1 - \mu_s) \log$$

Naive Mean Field (Contd.)

- The naive mean field problem takes the form

$$\max_{\mu \in \mathcal{M}_{tract}} \{\langle \mu, \theta \rangle - \psi^*_{H_0}(\mu)\}$$

- Using $\mu_{st} = \mu_s \mu_t$, we get the reduced problem

$$\max_{\{\mu_s\} \in [0,1]^n} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t - \sum_{s \in V} [\mu_s \log \mu_s + (1 - \mu_s) \log$$

- It is concave in μ_s with other co-ordinates held fixed

Naive Mean Field (Contd.)

- The naive mean field problem takes the form

$$\max_{\mu \in \mathcal{M}_{tract}} \{\langle \mu, \theta \rangle - \psi^*_{H_0}(\mu)\}$$

- Using $\mu_{st} = \mu_s \mu_t$, we get the reduced problem

$$\max_{\{\mu_s\} \in [0,1]^n} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t - \sum_{s \in V} [\mu_s \log \mu_s + (1 - \mu_s) \log$$

- It is concave in μ_s with other co-ordinates held fixed
- Taking gradient and setting it to zero yields

$$\mu_s \leftarrow \frac{1}{1 + \exp(-(\theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t))}$$

Structured Mean Field

- Considers tractable distributions with additional structure

Structured Mean Field

- Considers tractable distributions with additional structure
- For subgraph H, lets $\mathcal{I}(H)$ be the index set associated with H

Structured Mean Field

- Considers tractable distributions with additional structure
- For subgraph H, lets $\mathcal{I}(H)$ be the index set associated with H
- With $\mu(H) = \{\mu_\alpha | \alpha \in \mathcal{H}\}$, we have

Structured Mean Field

- Considers tractable distributions with additional structure
- For subgraph H, lets $\mathcal{I}(H)$ be the index set associated with H
- With $\mu(H) = \{\mu_\alpha | \alpha \in \mathcal{H}\}$, we have
  - The subvector $\mu(H)$ can be an arbitrary member of $\mathcal{M}(H)$

Structured Mean Field

- Considers tractable distributions with additional structure
- For subgraph H, lets $\mathcal{I}(H)$ be the index set associated with H
- With $\mu(H) = \{\mu_\alpha | \alpha \in \mathcal{H}\}$, we have
  - The subvector $\mu(H)$ can be an arbitrary member of $\mathcal{M}(H)$
  - Dual ψ_H^* depends only on $\mu(H)$, not on $\mu_\beta, \beta \in \mathcal{I}(G) \setminus \mathcal{I}(H)$

Structured Mean Field

- Considers tractable distributions with additional structure
- For subgraph H, lets $\mathcal{I}(H)$ be the index set associated with H
- With $\mu(H) = \{\mu_\alpha | \alpha \in \mathcal{H}\}$, we have
  - The subvector $\mu(H)$ can be an arbitrary member of $\mathcal{M}(H)$
  - Dual ψ_H^* depends only on $\mu(H)$, not on $\mu_\beta, \beta \in \mathcal{I}(G) \setminus \mathcal{I}(H)$
- But such μ_β do appear in the $\langle \mu, \beta \rangle$ term

Structured Mean Field

- Considers tractable distributions with additional structure
- For subgraph H, lets $\mathcal{I}(H)$ be the index set associated with H
- With $\mu(H) = \{\mu_\alpha | \alpha \in \mathcal{H}\}$, we have
  - The subvector $\mu(H)$ can be an arbitrary member of $\mathcal{M}(H)$
  - Dual ψ_H^* depends only on $\mu(H)$, not on $\mu_\beta, \beta \in \mathcal{I}(G) \setminus \mathcal{I}(H)$
- But such μ_β do appear in the $\langle \mu, \beta \rangle$ term
- Each $\mu_\beta = g_\beta(\mu(H))$, i.e., depends on $\mu(H)$ non-linearly

Structured Mean Field

- Considers tractable distributions with additional structure
- For subgraph H, lets $\mathcal{I}(H)$ be the index set associated with H
- With $\mu(H) = \{\mu_\alpha | \alpha \in \mathcal{H}\}$, we have
  - The subvector $\mu(H)$ can be an arbitrary member of $\mathcal{M}(H)$
  - Dual ψ_H^* depends only on $\mu(H)$, not on $\mu_\beta, \beta \in \mathcal{I}(G) \setminus \mathcal{I}(H)$
- But such μ_β do appear in the $\langle \mu, \beta \rangle$ term
- Each $\mu_\beta = g_\beta(\mu(H))$, i.e., depends on $\mu(H)$ non-linearly
- The approximate optimization problem can be written as

$$\sup_{\mu(H) \in \mathcal{M}(H)} \left\{ \sum_{\alpha \in \mathcal{I}(H)} \theta_\alpha \mu_\alpha + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha g_\alpha(\mu(H)) - \psi_H^*(\mu(H)) \right\}$$

Structured Mean Field

- Considers tractable distributions with additional structure
- For subgraph H, lets $\mathcal{I}(H)$ be the index set associated with H
- With $\mu(H) = \{\mu_\alpha | \alpha \in \mathcal{H}\}$, we have
  - The subvector $\mu(H)$ can be an arbitrary member of $\mathcal{M}(H)$
  - Dual ψ_H^* depends only on $\mu(H)$, not on $\mu_\beta, \beta \in \mathcal{I}(G) \setminus \mathcal{I}(H)$
- But such μ_β do appear in the $\langle \mu, \beta \rangle$ term
- Each $\mu_\beta = g_\beta(\mu(H))$, i.e., depends on $\mu(H)$ non-linearly
- The approximate optimization problem can be written as

$$\sup_{\mu(H) \in \mathcal{M}(H)} \left\{ \sum_{\alpha \in \mathcal{I}(H)} \theta_\alpha \mu_\alpha + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha g_\alpha(\mu(H)) - \psi_H^*(\mu(H)) \right\}$$

- For Ising model, with $H_0 = (V, \emptyset)$, $g_{st}(\mu(H_0)) = \mu_s \mu_t$

Structured Mean Field (Contd.)

- Let $F(\mu(H))$ denote the cost function

Structured Mean Field (Contd.)

- Let $F(\mu(H))$ denote the cost function
- Taking derivative w.r.t. $\mu_\beta, \beta \in \mathcal{I}(H)$ yields

$$\frac{\partial F(\mu(H))}{\partial \mu_\beta} = \theta_\beta + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha \frac{\partial g_\alpha(\mu(H))}{\partial \mu_\beta} - \frac{\partial \psi_H^*(\mu(H))}{\partial \mu_\beta}$$

Structured Mean Field (Contd.)

- Let $F(\mu(H))$ denote the cost function
- Taking derivative w.r.t. $\mu_\beta, \beta \in \mathcal{I}(H)$ yields

$$\frac{\partial F(\mu(H))}{\partial \mu_\beta} = \theta_\beta + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha \frac{\partial g_\alpha(\mu(H))}{\partial \mu_\beta} - \frac{\partial \psi_H^*(\mu(H))}{\partial \mu_\beta}$$

- $\gamma_\beta(H) = \frac{\partial \psi_H^*(\mu(H))}{\partial \mu_\beta}$ is the inverse moment mapping

Structured Mean Field (Contd.)

- Let $F(\mu(H))$ denote the cost function
- Taking derivative w.r.t. $\mu_\beta, \beta \in \mathcal{I}(H)$ yields

$$\frac{\partial F(\mu(H))}{\partial \mu_\beta} = \theta_\beta + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha \frac{\partial g_\alpha(\mu(H))}{\partial \mu_\beta} - \frac{\partial \psi_H^*(\mu(H))}{\partial \mu_\beta}$$

- $\gamma_\beta(H) = \frac{\partial \psi_H^*(\mu(H))}{\partial \mu_\beta}$ is the inverse moment mapping
- Setting the gradient to zero yields the update

$$\gamma_\beta(H) \leftarrow \theta_\beta + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha \frac{\partial g_\alpha(\mu(H))}{\partial \mu_\beta}$$

Structured Mean Field (Contd.)

- Let $F(\mu(H))$ denote the cost function
- Taking derivative w.r.t. $\mu_\beta, \beta \in \mathcal{I}(H)$ yields

$$\frac{\partial F(\mu(H))}{\partial \mu_\beta} = \theta_\beta + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha \frac{\partial g_\alpha(\mu(H))}{\partial \mu_\beta} - \frac{\partial \psi_H^*(\mu(H))}{\partial \mu_\beta}$$

- $\gamma_\beta(H) = \frac{\partial \psi_H^*(\mu(H))}{\partial \mu_\beta}$ is the inverse moment mapping
- Setting the gradient to zero yields the update

$$\gamma_\beta(H) \leftarrow \theta_\beta + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha \frac{\partial g_\alpha(\mu(H))}{\partial \mu_\beta}$$

- For Ising model, $\frac{\partial g_{st}}{\partial \mu_s} = \mu_t$ and so on

Structured Mean Field (Contd.)

- Let $F(\mu(H))$ denote the cost function
- Taking derivative w.r.t. $\mu_\beta, \beta \in \mathcal{I}(H)$ yields

$$\frac{\partial F(\mu(H))}{\partial \mu_\beta} = \theta_\beta + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha \frac{\partial g_\alpha(\mu(H))}{\partial \mu_\beta} - \frac{\partial \psi_H^*(\mu(H))}{\partial \mu_\beta}$$

- $\gamma_\beta(H) = \frac{\partial \psi_H^*(\mu(H))}{\partial \mu_\beta}$ is the inverse moment mapping
- Setting the gradient to zero yields the update

$$\gamma_\beta(H) \leftarrow \theta_\beta + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha \frac{\partial g_\alpha(\mu(H))}{\partial \mu_\beta}$$

- For Ising model, $\frac{\partial g_{st}}{\partial \mu_s} = \mu_t$ and so on
- We get the exact updates as naive mean field

Structured Mean Field (Contd.)

- Let $F(\mu(H))$ denote the cost function
- Taking derivative w.r.t. $\mu_\beta, \beta \in \mathcal{I}(H)$ yields

$$\frac{\partial F(\mu(H))}{\partial \mu_\beta} = \theta_\beta + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha \frac{\partial g_\alpha(\mu(H))}{\partial \mu_\beta} - \frac{\partial \psi_H^*(\mu(H))}{\partial \mu_\beta}$$

- $\gamma_\beta(H) = \frac{\partial \psi_H^*(\mu(H))}{\partial \mu_\beta}$ is the inverse moment mapping
- Setting the gradient to zero yields the update

$$\gamma_\beta(H) \leftarrow \theta_\beta + \sum_{\alpha \in \mathcal{I}^c(H)} \theta_\alpha \frac{\partial g_\alpha(\mu(H))}{\partial \mu_\beta}$$

- For Ising model, $\frac{\partial g_{st}}{\partial \mu_s} = \mu_t$ and so on
- We get the exact updates as naive mean field
- In general, H can be more involved

Non-convexity of Mean Field

- The original problem is concave

Non-convexity of Mean Field

- The original problem is concave
  - The constraint set $\mathcal{M}(H)$ is convex

Non-convexity of Mean Field

- The original problem is concave
  - The constraint set $\mathcal{M}(H)$ is convex
  - The objective contains entropy and linear terms in μ_α

Non-convexity of Mean Field

- The original problem is concave
  - The constraint set $\mathcal{M}(H)$ is convex
  - The objective contains entropy and linear terms in μ_α
- The (structured) mean field contains non-linear terms

Non-convexity of Mean Field

- The original problem is concave
  - The constraint set $\mathcal{M}(H)$ is convex
  - The objective contains entropy and linear terms in μ_α
- The (structured) mean field contains non-linear terms
  - $\sum_{\alpha \in \mathcal{I}(H)} \theta_\alpha g_\alpha(\mu)$ involves non-linear function g_α

Non-convexity of Mean Field

- The original problem is concave
  - The constraint set $\mathcal{M}(H)$ is convex
  - The objective contains entropy and linear terms in μ_α
- The (structured) mean field contains non-linear terms
  - $\sum_{\alpha \in \mathcal{I}(H)} \theta_\alpha g_\alpha(\mu)$ involves non-linear function g_α
  - For Ising model, g_α is of the form $\mu_s \mu_t$

Non-convexity of Mean Field

- The original problem is concave
  - The constraint set $\mathcal{M}(H)$ is convex
  - The objective contains entropy and linear terms in μ_α
- The (structured) mean field contains non-linear terms
  - $\sum_{\alpha \in \mathcal{I}(H)} \theta_\alpha g_\alpha(\mu)$ involves non-linear function g_α
  - For Ising model, g_α is of the form $\mu_s \mu_t$
  - A quadratic form need not be concave