

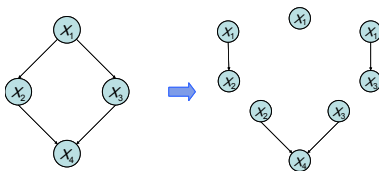
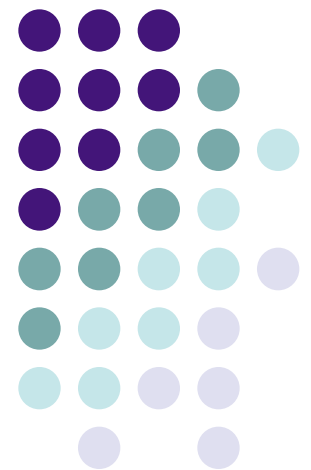


Probabilistic Graphical Models

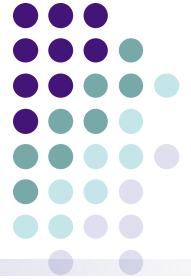
Variational Inference III: Variational Principle I

Junming Yin

Lecture 16, March 19, 2012

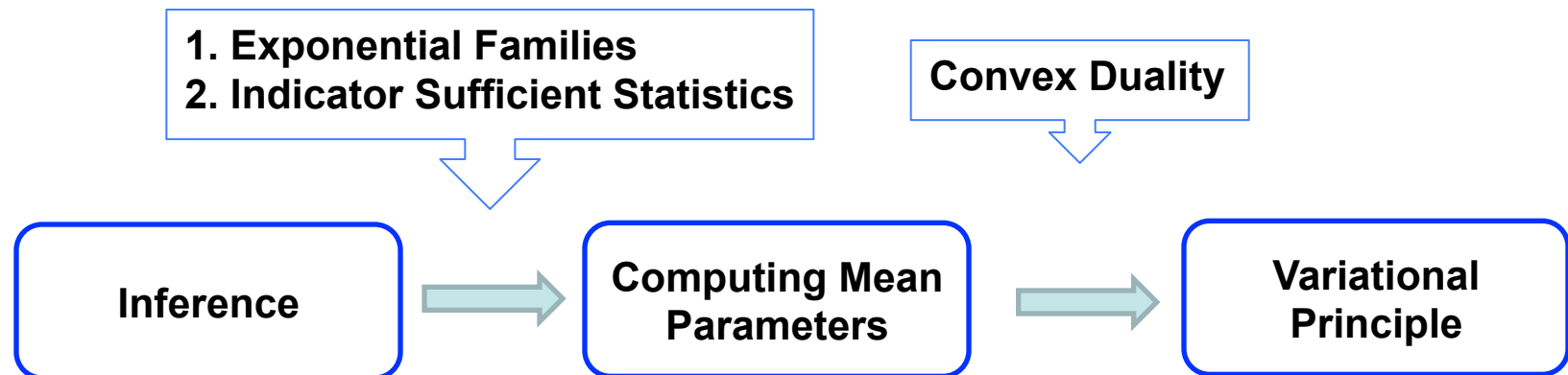


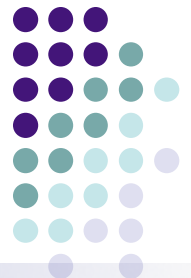
Reading:



What have we learned so far

- Free energy based approaches
 - Direct approximation of Gibbs free energy: Bethe free energy and loop BP
 - Restricting the family of approximation distribution: mean field method
- Convex duality based approaches





Computing Mean Parameter: Bernoulli

- A single Bernoulli random variable

$$\textcircled{X} \quad \theta$$

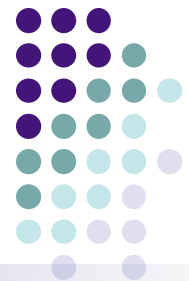
$$p(x; \theta) = \exp\{\theta x - A(\theta)\}, x \in \{0, 1\}, A(\theta) = \log(1 + e^\theta)$$

- Inference = Computing the mean parameter

$$\mu(\theta) = \mathbb{E}_\theta[X] = p(X = 1; \theta) = \frac{e^\theta}{1 + e^\theta}$$

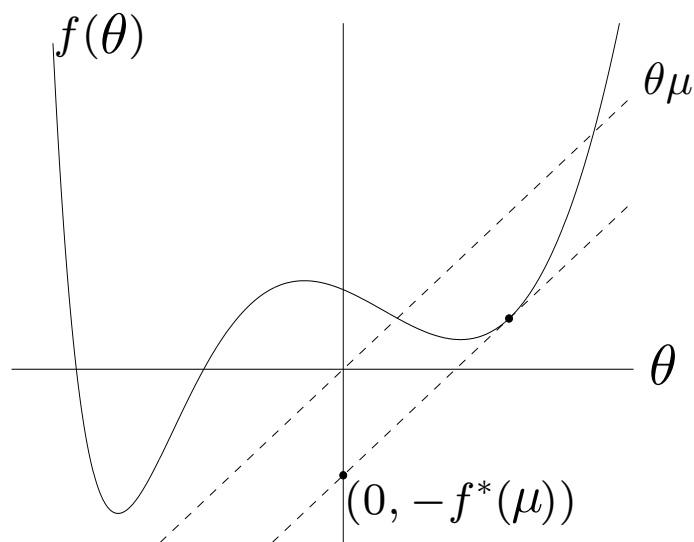
- Want to do it in a **variational** manner: cast the procedure of computing mean (summation) in an optimization-based formulation

Conjugate Dual Function

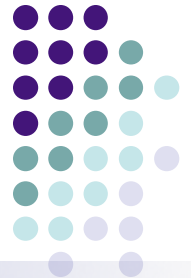


- Given any function $f(\theta)$, its conjugate dual function is:

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



- Conjugate dual is always a **convex** function: pointwise supremum of a class of linear functions



Dual of the Dual is the Original

- Under some technical condition on f (**convex** and lower semi-continuous), the dual of dual is itself:

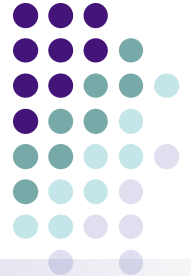
$$f = (f^*)^*$$

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

- For log partition function

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

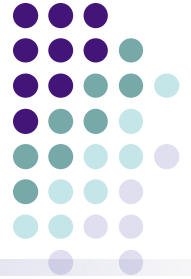
- The dual variable μ has a natural interpretation as mean parameters



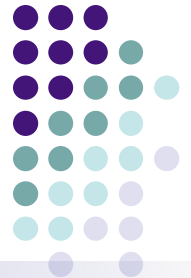
Computing Mean Parameter: Bernoulli

- The conjugate $A^*(\mu) := \sup_{\theta \in \mathbb{R}} \{ \mu\theta - \log[1 + \exp(\theta)] \}$
- Stationary condition $\mu = \frac{e^\theta}{1 + e^\theta} \quad (\mu = \nabla A(\theta))$
- If $\mu \in (0, 1)$, $\theta(\mu) = \log\left(\frac{\mu}{1 - \mu}\right)$, $A^*(\mu) = \mu \log(\mu) + (1 - \mu) \log(1 - \mu)$
- If $\mu \notin [0, 1]$, $A^*(\mu) = +\infty$
- We have $A^*(\mu) = \begin{cases} \mu \log \mu + (1 - \mu) \log(1 - \mu) & \text{if } \mu \in [0, 1] \\ +\infty & \text{otherwise.} \end{cases}$
- The variational form: $A(\theta) = \max_{\mu \in [0, 1]} \{ \mu \cdot \theta - A^*(\mu) \}$.
- The optimum is achieved at $\mu(\theta) = \frac{e^\theta}{1 + e^\theta}$. This is the mean!

Remark



- The last few identities are not coincidental but rely on a deep theory in general exponential family
 - The dual function is the negative **entropy** function
 - The mean parameter is restricted
 - Solving the optimization returns the mean parameter
- Next step: develop this framework for general exponential families/graphical models



Computation of Conjugate Dual

- Given an exponential family

$$p(x_1, \dots, x_m; \theta) = \exp \left\{ \sum_{i=1}^d \theta_i \phi_i(x) - A(\theta) \right\}$$

- The dual function

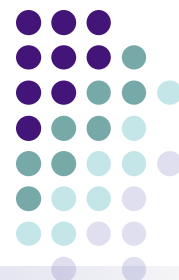
$$A^*(\mu) := \sup_{\theta \in \Omega} \{ \langle \mu, \theta \rangle - A(\theta) \}$$

- The stationary condition: $\mu - \nabla A(\theta) = 0$
- Derivatives of A yields mean parameters

$$\frac{\partial A}{\partial \theta_i}(\theta) = \mathbb{E}_\theta[\phi_i(X)] = \int \phi_i(x) p(x; \theta) dx$$

- The stationary condition becomes $\mu = \mathbb{E}_\theta[\phi(X)]$
- Question: for which $\mu \in \mathbb{R}^d$ does it have a solution $\theta(\mu)$?

Computation of Conjugate Dual



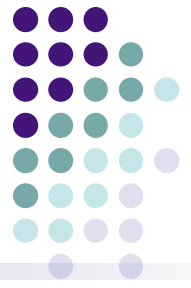
- Let's assume there is a solution $\theta(\mu)$ such that $\mu = \mathbb{E}_{\theta(\mu)}[\phi(X)]$
- The dual has the form

$$\begin{aligned} A^*(\mu) &= \langle \theta(\mu), \mu \rangle - A(\theta(\mu)) \\ &= \mathbb{E}_{\theta(\mu)} [\langle \theta(\mu), \phi(X) \rangle - A(\theta(\mu))] \\ &= \mathbb{E}_{\theta(\mu)} [\log p(X; \theta(\mu))] \end{aligned}$$

- The entropy is defined as

$$H(p(x)) = - \int p(x) \log p(x) dx$$

- So the dual is $A^*(\mu) = -H(p(x; \theta(\mu)))$ when there is a solution $\theta(\mu)$
- Question: for which $\mu \in \mathbb{R}^d$ does it have a solution $\theta(\mu)$?



Marginal Polytope

- For any distribution $p(x)$ and a set of sufficient statistics define a vector of **mean parameters**

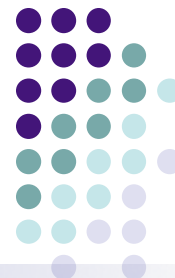
$$\mu_i = \mathbb{E}_p[\phi_i(X)] = \int \phi_i(x) p(x) dx$$

- $p(x)$ is **not** necessarily an exponential family
- The set of all realizable mean parameters

$$\mathcal{M} := \{\mu \in \mathbb{R}^d \mid \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\}.$$

- It is a **convex** set
- For discrete exponential families, this is called **marginal polytope**

Convex Polytope



- Convex hull representation

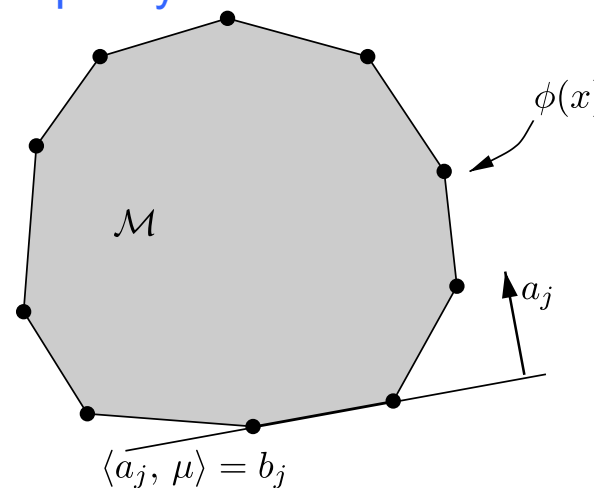
$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d \mid \sum_{x \in \mathcal{X}^m} \phi(x)p(x) = \mu, \text{ for some } p(x) \geq 0, \sum_{x \in \mathcal{X}^m} p(x) = 1 \right\}$$
$$\triangleq \text{conv} \left\{ \phi(x), x \in \mathcal{X}^m \right\}$$

- Half-plane representation

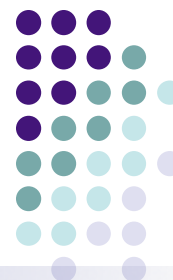
- **Minkowski-Weyl Theorem:** any non-empty convex polytope can be characterized by a **finite** collection of linear inequality constraints

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d \mid a_j^\top \mu \geq b_j, \forall j \in \mathcal{J} \right\},$$

where $|\mathcal{J}|$ is finite.



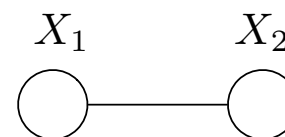
Example: Ising Model



- Sufficient statistics: $\phi(x) := (x_s, s \in V; x_s x_t, (s, t) \in E) \in \mathbb{R}^{|V|+|E|}$.

- Mean parameters: $\mu_s = \mathbb{E}_p[X_s] = \mathbb{P}[X_s = 1]$ for all $s \in V$, and
 $\mu_{st} = \mathbb{E}_p[X_s X_t] = \mathbb{P}[(X_s, X_t) = (1, 1)]$ for all $(s, t) \in E$.

- Two-node Ising model

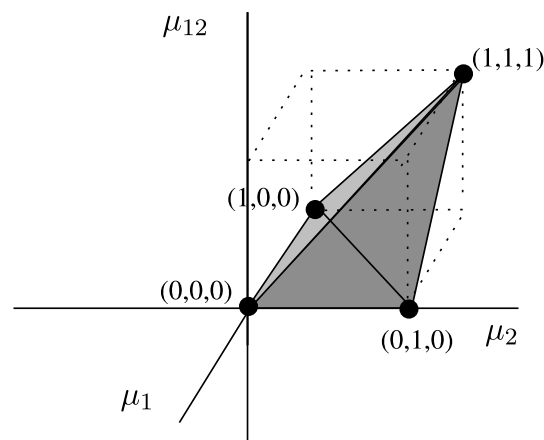


- Convex hull representation

$$\text{conv}\{(0,0,0), (1,0,0), (0,1,0), (1,1,1)\}$$

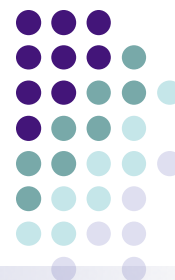
- Half-plane representation

$$\begin{aligned} \mu_1 &\geq \mu_{12} \\ \mu_2 &\geq \mu_{12} \\ \mu_{12} &\geq 0 \\ 1 + \mu_{12} &\geq \mu_1 + \mu_2 \end{aligned}$$



- Exercise: three-node Ising model

Example: Discrete MRF



- Sufficient statistics:
$$\begin{aligned} &\mathbb{I}_j(x_s) && \text{for } s = 1, \dots, n, \quad j \in \mathcal{X}_s \\ &\mathbb{I}_{jk}(x_s, x_t) && \text{for } (s, t) \in E, \quad (j, k) \in \mathcal{X}_s \times \mathcal{X}_t \end{aligned}$$

- Mean parameters are marginal probabilities:

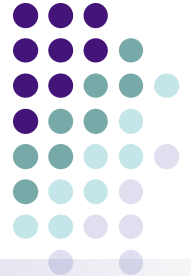
$$\mu_{s;j} = \mathbb{E}_p[\mathbb{I}_j(X_s)] = \mathbb{P}[X_s = j] \quad \forall j \in \mathcal{X}_s,$$

$$\mu_{st;jk} = \mathbb{E}_p[\mathbb{I}_{st;jk}(X_s, X_t)] = \mathbb{P}[X_s = j, X_t = k] \quad \forall (j, k) \in \mathcal{X}_s \times \mathcal{X}_t.$$

- Marginal Polytope

$$\mathcal{M}(G) = \{\mu \in \mathbb{R}^d \mid \exists p \text{ with marginals } \mu_{s;j}, \mu_{st;jk}\}$$

- For tree graphical models, the number of half-planes (**facet complexity**) grows only *linearly* in the graph size
- For general graphs, it is extremely difficult to characterize the marginal polytope



Variational Principle (Theorem 3.4)

- The dual function takes the form

$$A^*(\mu) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}}. \end{cases}$$

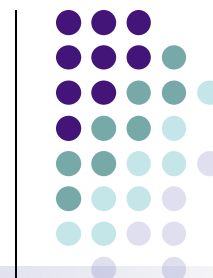
- $\theta(\mu)$ satisfies $\mu = \mathbb{E}_{\theta(\mu)}[\phi(X)]$
- The log partition function has the variational form

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

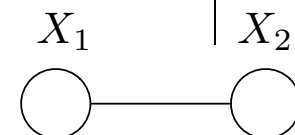
- For all $\theta \in \Omega$, the above optimization problem is attained uniquely at $\mu(\theta) \in \mathcal{M}^\circ$ that satisfies

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

Example: Two-node Ising Model



- The distribution $p(x; \theta) \propto \exp\{\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_{12}\}$



- The marginal polytope is characterized by

$$\begin{array}{rcl} \mu_1 & \geq & \mu_{12} \\ \mu_2 & \geq & \mu_{12} \\ \mu_{12} & \geq & 0 \\ 1 + \mu_{12} & \geq & \mu_1 + \mu_2 \end{array}$$

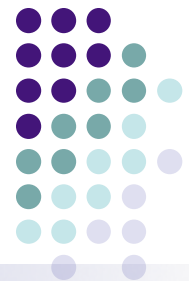
- The dual has an explicit form

$$\begin{aligned} A^*(\mu) = & \mu_{12} \log \mu_{12} + (\mu_1 - \mu_{12}) \log(\mu_1 - \mu_{12}) + (\mu_2 - \mu_{12}) \log(\mu_2 - \mu_{12}) \\ & + (1 + \mu_{12} - \mu_1 - \mu_2) \log(1 + \mu_{12} - \mu_1 - \mu_2) \end{aligned}$$

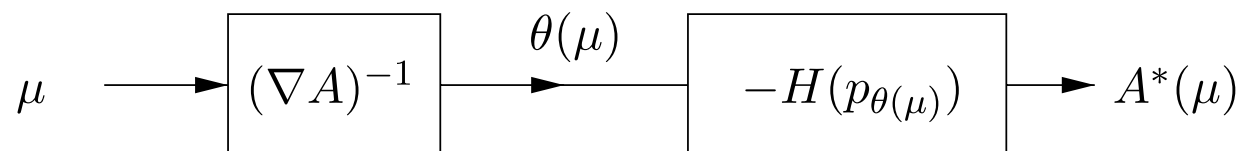
- The variational problem $A(\theta) = \max_{\{\mu_1, \mu_2, \mu_{12}\} \in \mathcal{M}} \{\theta_1 \mu_1 + \theta_2 \mu_2 + \theta_{12} \mu_{12} - A^*(\mu)\}$
- The optimum is attained at

$$\mu_1(\theta) = \frac{\exp\{\theta_1\} + \exp\{\theta_1 + \theta_2 + \theta_{12}\}}{1 + \exp\{\theta_1\} + \exp\{\theta_2\} + \exp\{\theta_1 + \theta_2 + \theta_{12}\}}$$

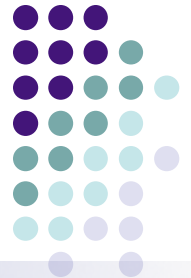
Challenges



- In general graphical models, the marginal polytope can be very difficult to characterize explicitly
- The dual function is implicitly defined:



- Inverse mapping is nontrivial
- Evaluating the entropy requires high-dimensional integration (summation)



Variational Inference

- Variational formulation

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

- General idea of variational inference for graphical models:
 - Approximate the function to be optimized, i.e., the entropy term (Bethe-Kikuchi, sum-product)
 - Restrict the set over which the optimization takes place to a **subset**, i.e., the marginal polytope (mean field methods)