

#### **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



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

• 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:



• 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 semicontinuous), the dual of dual is itself:

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

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

• For log partition function

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

• The dual variable  $\mu$  has a natural interpretation as mean parameters



#### 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

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

- 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  $\mu = \mathbb{E}_{\theta(u)}[\phi(X)]$
- The dual has the form

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

• 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

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

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

$$\mathcal{M} = \Big\{ \mu \in \mathbb{R}^d | a_j^\top \mu \ge b_j, \ \forall j \in \mathcal{J} \Big\},\$$

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



case: namely, a pair of variables  $(X_1, X_2)$ , and the graph consisting of case: namely, a pair of variables  $(X_1, X_2)$ , and the graph consisting of the single edge joining them. In this case, the set  $\mathcal{M}$  is a polytope in **Example 3h8 (Isingd Vieun Parameters**) case; otherseting of the convex hull of three dimensions (two nodes plus one edge). It is the convex hull of the sumcle ft stati **Example 3.8** (Isingective and the pairwise functions  $(x_s, s \in V)$  and the pairwise functions ple 3.1, veheorsufficient not statistics vere the ling of (0, del) and vehe singlet due Extample, s shadthyodefingtions ( $x_{e}x_{t+1}$  $(s,t) \in E$ ). The 3.30) presentation (3.29) for this for this The associated Eleventary probability theory in the associated Eleventary probability theory in the space representation case. Elementary probability theory is the space representation of the space y and a little calculation shows that Sufficient state the anti-parameters in the three mean parameters ( $\mu_{12} = \mu_{12} = \mu_{12}$  $\overline{\mathbf{A}} = \mathbf{A} \cdot \mathbf{A$ the single edge joining them. In this case, the set al polytope\_in. Force of the state concersistation ( the vectors (1,2,3,4) the vectors (1,2,3,4)wise marginals convex hull of  $\{\beta(x), x\}$  in the second strain of the second sites of the second strain strain of the second str • Marcharter 2001 24 5 Consequentive athest mar gin singleton prize i over singleton (2,0), it enthaver wise realize hals over Lett as herspatistic graph fredges a (cest) prehen bet of the capacity of the  $\begin{array}{c} \mu_1 & \text{combleatorics literature} \\ \textbf{this set is known as the consex Elementary (s) obtained in the probability literature (s) where a lister (12) of the polytope [69, 187]. \\ \mu_2 & \text{ar the 2 cut polytope [69, 187].} \\ \textbf{Inthe non-adjustate polytope [69, 187].} \\ \textbf{Inthe non-adjustate polytope [69, 187].} \end{array}$ carrelation polytope rune Liquis zon (3130). sfrothine set Mitsall Adt. Can ben waitzed singleton stride Plainvise 2 manginal prababy Exercise: three-node ising mode  $\{0,1\}^{\mu_1\mu_2}$  the polyhedral combinatorics literature, this set is known as the correlation polytope, or the cut polytope. 69 187] Fig. 3.6 Hustration of  $\mathcal{M}$  for the special case of an Ising model with two variables  $(X_1, X_2) \in \{0, 1\}^2$ . The four real parameters  $\mu_1 = \mathbb{E}[X_1], \mu_2 = \mathbb{E}[X_2]$  and  $\mu_{12} = \mathbb{E}[X_2]$  and  $\mu_{12} = \mathbb{E}[X_2]$  and  $\mu_{13} = \mathbb{E}[X_2], \mu_1 = \mathbb{E}[X_1], \mu_2 = \mathbb{E}[X_2]$  and  $\mu_{13} = \mathbb{E}[X_2], \mu_1 = \mathbb{E}[X_1], \mu_2 = \mathbb{E}[X_2], \mu_1 = \mathbb{E}[X_1], \mu_2 = \mathbb{E}[X_2]$  and  $\mu_{13} = \mathbb{E}[X_2], \mu_1 = \mathbb{E}[X_1], \mu_2 = \mathbb{E}[X_2], \mu_3 = \mathbb{E}[X_2], \mu_4 = \mathbb{E}[X_1], \mu_4 = \mathbb{E}[X_2], \mu_4 = \mathbb{E}[X_2], \mu_4 = \mathbb{E}[X_1], \mu_4 = \mathbb{E}$ 

#### **Example: Discrete MRF**

- Sufficient statistics:  $\begin{array}{cc} \mathbb{I}_{j}(x_{s}) & \text{for } s = 1, \dots n, \quad j \in \mathcal{X}_{s} \\ \mathbb{I}_{jk}(x_{s}, x_{t}) & \text{for}(s, t) \in E, \quad (j, k) \in \mathcal{X}_{s} \times \mathcal{X}_{t} \end{array}$
- Mean parameters are marginal probabilities:

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

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

- Marginal Polytope
- *M*(*G*) = {µ ∈ ℝ<sup>d</sup> | ∃p with marginals µ<sub>s;j</sub>, µ<sub>st;jk</sub>}
  F *P* at models, the number of half-planes (facet is only *linearly* in the graph size *phs*, 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  $\mu = \mathbb{E}_{\theta(u)}[\phi(X)]$ 

• 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  $\mu(\theta) \in \mathcal{M}^o$  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
- The dual has an explicit form  $A^*(\mu) = \mu_{12} \log \mu_{12} + (\mu_1 - \mu_{12}) \log(\mu_1 - \mu_{12}) + (\mu_2 - \mu_{12}) \log(\mu_2 - \mu_{12}) + (1 + \mu_{12} - \mu_1 - \mu_2) \log(1 + \mu_{12} - \mu_1 - \mu_2)$
- 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

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

 $X_2$ 

 $X_1$ 

 $\mu_2$ 

 $\mu_1 \geq \mu_{12}$ 

 $\mu_{12} \geq 0$ 

 $\geq \mu_{12}$ 

#### Challenges



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

$$\mu \longrightarrow (\nabla A)^{-1} \longrightarrow -H(p_{\theta(\mu)}) \longrightarrow A^*(\mu)$$

- 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)