

CS 6347

Lecture 8

Variational Bounds

# Approximate Marginal Inference

- Last time: approximate MAP inference
  - Reparamaterizations
  - Linear programming over the local marginal polytope
- Approximate marginal inference (e.g.,  $p(y_i|x)$ )
  - Sampling methods (MCMC, etc.)
  - Variational methods (loopy belief propagation, TRW, etc.)

# KL Divergence

- In order to perform approximate marginal inference, we will try to find distributions that approximate the true distribution
  - Ideally, the marginals of the approximating distribution should be easy to compute
- For this, we need a notion of closeness of distributions

# KL Divergence

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

- Called the Kullback-Leibler divergence
- $D(p||q) \geq 0$  with equality if and only if  $p = q$
- Not symmetric,  $D(p||q) \neq D(q||p)$

# Jensen's Inequality

- Let  $f(x)$  be a convex function and  $a_i \geq 0$  such that  $\sum_i a_i = 1$

$$\sum_i a_i f(x_i) \geq f\left(\sum_i a_i x_i\right)$$

- Useful inequality when dealing with convex/concave functions
- When does equality hold?

# KL Divergence

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

- Suppose that we want to approximate the distribution  $p$  with some other distribution  $q$  in some family of distributions  $Q$
- Could minimize KL divergence in one of two ways
  - $\arg \min_{q \in Q} D(p||q)$
  - $\arg \min_{q \in Q} D(q||p)$

# KL Divergence

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

- Suppose that we want to approximate the distribution  $p$  with some other distribution  $q$  in some family of distributions  $Q$
- Could minimize KL divergence in one of two ways
  - $\arg \min_{q \in Q} D(p||q)$  **Called the M-projection**
  - $\arg \min_{q \in Q} D(q||p)$  **Called the I-projection**

# KL Divergence

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

- Suppose that we want to approximate the distribution  $p$  with some other distribution  $q$  in some family of distributions  $Q$
- Could minimize KL divergence in one of two ways
  - $\arg \min_{q \in Q} D(p||q)$  As hard as the original inference problem
  - $\arg \min_{q \in Q} D(q||p)$  Potentially easier...



# Variational Inference

- Let's let  $p(x) = \frac{1}{Z} \prod_c \psi_c(x_c)$  be the distribution that we want to approximate with distribution  $q$

$$\begin{aligned} D(q||p) &= \sum_x q(x) \log \frac{q(x)}{p(x)} \\ &= \sum_x q(x) \log q(x) - \sum_x q(x) \log p(x) \\ &= -H(q) - \sum_x q(x) \log p(x) \\ &= -H(q) + \log Z - \sum_x \sum_c q(x) \log \psi_c(x_c) \\ &= -H(q) + \log Z - \sum_c \sum_{x_c} q_c(x_c) \log \psi_c(x_c) \end{aligned}$$

# Variational Inference

- Let's let  $p(x) = \frac{1}{Z} \prod_c \psi_c(x_c)$  be the distribution that we want to approximate with distribution  $q$

$$\begin{aligned} D(q||p) &= \sum_x q(x) \log \frac{q(x)}{p(x)} \\ &= \sum_x q(x) \log q(x) - \sum_x q(x) \log p(x) \\ &= -H(q) - \sum_x q(x) \log p(x) \\ &= -H(q) + \log Z - \sum_x \sum_c q(x) \log \psi_c(x_c) \\ &= -H(q) + \log Z - \sum_c \sum_{x_c} q_c(x_c) \log \psi_c(x_c) \end{aligned}$$

Where have we seen this before?

# MAP Integer Program

$$\max_{\tau} \sum_{i \in V} \sum_{x_i} \tau_i(x_i) \log \phi_i(x_i) + \sum_{(i,j) \in E} \sum_{x_i, x_j} \tau_{ij}(x_i, x_j) \log \psi_{ij}(x_i, x_j)$$

such that

$$\sum_{x_i} \tau_i(x_i) = 1$$

For all  $i \in V$

$$\sum_{x_j} \tau_{ij}(x_i, x_j) = \tau_i(x_i)$$

For all  $(i, j) \in E, x_i$

$$\tau_i(x_i) \in \{0, 1\}$$

For all  $i \in V, x_i$

$$\tau_{ij}(x_i, x_j) \in \{0, 1\}$$

For all  $(i, j) \in E, x_i, x_j$

# Variational Inference

- Let's let  $p(x) = \frac{1}{Z} \prod_c \psi_c(x_c)$  be the distribution that we want to approximate with distribution  $q$

$$D(q||p) = -H(q) + \log Z - \sum_C \sum_{x_C} q_C(x_C) \log \psi_C(x_C)$$

- Using the observation that the KL divergence is non-negative

$$\log Z \geq H(q) + \sum_C \sum_{x_C} q_C(x_C) \log \psi_C(x_C)$$

# Variational Inference

- Let's let  $p(x) = \frac{1}{Z} \prod_c \psi_c(x_c)$  be the distribution that we want to approximate with distribution  $q$

$$D(q||p) = -H(q) + \log Z - \sum_C \sum_{x_C} q_C(x_C) \log \psi_C(x_C)$$

- Using the observation that the KL divergence is non-negative

$$\log Z \geq H(q) + \sum_C \sum_{x_C} q_C(x_C) \log \psi_C(x_C)$$

- This lower bound holds for **any**  $q$

# Variational Inference

- Let's let  $p(x) = \frac{1}{Z} \prod_c \psi_c(x_c)$  be the distribution that we want to approximate with distribution  $q$

$$D(q||p) = -H(q) + \log Z - \sum_c \sum_{x_c} q_c(x_c) \log \psi_c(x_c)$$

- Using the observation that the KL divergence is non-negative

$$\log Z \geq H(q) + \sum_c \sum_{x_c} q_c(x_c) \log \psi_c(x_c)$$

Maximizing this over  $q$  gives  
equality

# Variational Inference

$$\log Z \geq H(q) + \sum_C \sum_{x_C} q_C(x_C) \log \psi_C(x_C)$$

- The right hand side is a concave function of  $q$
- Despite that, this optimization problem is **hard!** (surprised?)
  - Exponentially many distributions,  $q(x)$   
We need a more compact way to express them
  - Computing the entropy is non-trivial

# Variational Inference

$$\log Z \geq H(q) + \sum_C \sum_{x_C} q_C(x_C) \log \psi_C(x_C)$$

- Two kinds of methods that are used to deal with these difficulties
  - Mean-field methods: assume that the approximating distribution factorizes as  $q(x) \propto \prod_{i \in V} q_i(x_i)$
  - Relaxation based methods: replace hard pieces of the optimization with easier optimization problems
    - Similar to the MAP IP  $\rightarrow$  MAP LP relaxation



# Relaxation Approach

$$\log Z \geq H(q) + \sum_C \sum_{x_C} q_C(x_C) \log \psi_C(x_C)$$

- To handle the representation problem, we can use the same LP relaxation trick that we did before
- For each  $\tau$  in the marginal polytope, we can rewrite the RHS as

$$\log Z \geq H(\tau) + \sum_C \sum_{x_C} \tau_C(x_C) \log \psi_C(x_C)$$

# Relaxation Approach

$$\log Z \geq H(q) + \sum_C \sum_{x_C} q_C(x_C) \log \psi_C(x_C)$$

- To handle the representation problem, we can use the same LP relaxation trick that we did before
- For each  $\tau$  in the marginal polytope, we can rewrite the RHS as

$$\log Z \geq H(\tau) + \sum_C \sum_{x_C} \tau_C(x_C) \log \psi_C(x_C)$$

Maximum entropy over all  $\tau$  with  
these marginals

# Relaxation Approach

$$\max_{\tau \in M} H(\tau) + \sum_C \sum_{x_C} \tau_C(x_C) \log \psi_C(x_C)$$

- Marginal polytope,  $M$ , is intractable to optimize over
- Use the local polytope,  $T$

$$\sum_{x_{C \setminus i}} \tau_C(x_C) = \tau_i(x_i) \text{ for all } C, i \in V$$

$$\sum_{x_i} \tau_i(x_i) = 1 \text{ for all } i \in V$$

# Relaxation Approach

$$\max_{\tau \in \mathbf{T}} H(\tau) + \sum_C \sum_{x_C} \tau_C(x_C) \log \psi_C(x_C)$$

- Even with the polytope relaxation, the optimization problem still remains challenging as computing the entropy remains nontrivial
  - We will need to approximate the entropy as well
  - For which distributions is it easy to compute the entropy?

# Tree Reparameterization

- On a tree, the joint distribution factorizes in a special way

$$p(x_1, \dots, x_n) = \frac{1}{Z'} \prod_{i \in V} p_i(x_i) \prod_{(i,j) \in E} \frac{p_{ij}(x_i, x_j)}{p_i(x_i)p_j(x_j)}$$

- $p_i$  is the marginal distribution of the  $i^{\text{th}}$  variable and  $p_{ij}$  is the max-marginal distribution for the edge  $(i, j) \in E$
- This applies to “clique trees” as well (i.e., when the factor graph is a tree)

# Tree Reparameterization

- On a tree, the joint distribution factorizes in a special way

$$p(x_1, \dots, x_n) = \frac{1}{Z'} \prod_{i \in V} p_i(x_i) \prod_C \frac{p_C(x_C)}{\prod_{i \in C} p_i(x_i)}$$

- $p_i$  is the marginal distribution of the  $i^{\text{th}}$  variable and  $p_{ij}$  is the max-marginal distribution for the edge  $(i, j) \in E$
- This applies to “clique trees” as well (i.e., when the factor graph is a tree)

# Entropy of a Tree

- Given this factorization, we can easily compute the entropy of a tree structured distribution

$$H_{Tree} = - \sum_{i \in V} \sum_{x_i} p_i(x_i) \log p_i(x_i) - \sum_C \sum_{x_C} p_C(x_C) \log \frac{p_C(x_C)}{\prod_{i \in C} p_i(x_i)}$$

- This only depends on the marginals
- Use this as an approximation for general distributions!

# Bethe Free Energy

- Combining these two approximations gives us the so-called Bethe free energy approximation

$$\max_{\tau \in \mathbf{T}} H_B(\tau) + \sum_C \sum_{x_C} \tau_C(x_C) \log \psi_C(x_C)$$

where

$$H_B(\tau) = - \sum_{i \in V} \sum_{x_i} \tau_i(x_i) \log \tau_i(x_i) - \sum_C \sum_{x_C} \tau_C(x_C) \log \frac{\tau_C(x_C)}{\prod_{i \in C} \tau_i(x_i)}$$



# Bethe Free Energy

$$\max_{\tau \in \mathbf{T}} H_B(\tau) + \sum_C \sum_{x_C} \tau_C(x_C) \log \psi_C(x_C)$$

- This is not a concave optimization problem for general graphs
  - It is still difficult to maximize
  - Fixed points of loopy belief propagation correspond to saddle points of this objective over the local marginal polytope