CS 6347

Lecture 8 & 9

Lagrange Multipliers & Varitional Bounds

General Optimization

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$f_i(x) \leq 0, \qquad i = 1, \ldots, m$$
$$h_i(x) = 0, \qquad i = 1, \ldots, p$$

General Optimization

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

f_0 is not necessarily convex

subject to:

$$f_i(x) \leq 0, \qquad i = 1, \ldots, m$$
$$h_i(x) = 0, \qquad i = 1, \ldots, p$$

General Optimization

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$f_i(x) \leq 0, \quad i = 1, \ldots, m$$
$$h_i(x) = 0, \quad i = 1, \ldots, p$$

Constraints can be arbitrary functions

Lagrangian

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

- Incorporate constraints into a new objective function
- $\lambda \geq 0$ and ν are vectors of ***Lagrange multipliers***
- The Lagrange multipliers can be thought of as soft constraints

Duality

- Construct a dual function by minimizing the Lagrangian over the primal variables

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$$

- $g(\lambda, \nu) = -\infty$ whenever the Lagrangian is not bounded from below for a fixed λ and ν

The Primal Problem

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$\begin{aligned} f_i(x) &\leq 0, \qquad i = 1, \dots, m \\ h_i(x) &= 0, \qquad i = 1, \dots, p \end{aligned}$$

Equivalently,

$$\inf_x \sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu)$$

The Dual Problem

$$\sup_{\lambda \geq 0, \nu} g(\lambda, \nu)$$

Equivalently,

$$\sup_{\lambda \geq 0, \nu} \inf_{x} L(x, \lambda, \nu)$$

- The dual problem is always concave, even if the primal problem is not convex

Primal vs. Dual

$$\sup_{\lambda \geq 0, \nu} \inf_x L(x, \lambda, \nu) \leq \inf_x \sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu)$$

- Why?
  - $g(\lambda, \nu) \leq L(x, \lambda, \nu)$ for all x
  - $L(x', \lambda, \nu) \leq f_0(x')$ for any feasible x', $\lambda \geq 0$
    - x is feasible if it satisfies all of the constraints
  - Let x^* be the optimal solution to the primal problem and $\lambda \geq 0$

$$g(\lambda, \nu) \leq L(x^*, \lambda, \nu) \leq f_0(x^*)$$

Duality

- Under certain conditions, the two optimization problems are equivalent

$$\sup_{\lambda \geq 0, \nu} \inf_{x} L(x, \lambda, \nu) = \inf_{x} \sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu)$$

  - This is called strong duality

- If the inequality is strict, then we say that there is a duality gap

  - Size of gap measured by the difference between the two sides of the inequality

Slater's Condition

For any optimization problem of the form

$$\min_{x\in\mathbb{R}^n} f_0(x)$$

subject to:

$$f_i(x) \le 0, \qquad i = 1, \dots, m$$
$$Ax = b$$

where $f_0, \dots, f_m$ are convex functions, strong duality holds if there exists an x such that

$$f_i(x) < 0, \qquad i = 1, \dots, m$$
$$Ax = b$$

Some Examples

- Minimize $x^2 + y^2$ subject to $x + y \geq 2$
- Maximize $-x \log x - y \log y - z \log z$ subject to $x, y, z \geq 0$ and $x + y + z = 1$
- Minimize xy subject to $x + y \geq 1$

Approximate Marginal Inference

- Last week: 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 $q(x) \propto \prod_{i \in V} q_i(x_i)$
    - Similar idea to naïve Bayes
  - Relaxation based methods: replace hard pieces of the optimization with easier optimization problems
    - Similar to the MAP IP -> 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 τ 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 τ 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 τ with these marginals

Relaxation Approach

$$\max_{\tau \in \mathrm{M}} H(\tau) + \sum_{C} \sum_{x_C} \tau_C(x_C) \log \psi_c(x_c)$$

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

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

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

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^{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

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

- p_i is the marginal distribution of the i^{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

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

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
  - However, fixed points of loopy belief propagation correspond to saddle points of this objective over the local marginal polytope