Homework 3

Convex Optimization 10-725/36-725

Due Friday October 14 at 5:30pm
submitted to Christoph Dann in Gates 8013
(Remember to a submit separate writeup for each problem, with your name at the top)

Total: 75 points
v1.1

1 Duality in Linear Programs (17 pts) [Mariya]

(a, 3pts) Derive the dual of

$$\begin{aligned} \min_{x_1,x_2} \quad & -4x_1 + 2x_2 \\ \text{subject to} \quad & -x_1 + x_2 \geq 2 \\ & x_1 - x_2 \geq 1 \\ & x_1, x_2 \geq 0 \end{aligned}$$

What are the primal optimal value and the dual optimal value? What is the duality gap?

(b, 14pts) Both Ryan and the TAs want many students to attend their office hours. However, the TAs have noticed that students are less likely to go to their office hours if they attend Ryan‘s, so the TAs decide to sabotage Ryan's office hours. The TAs will block the paths between class in Wean and Ryan's office in Baker.

In this problem, we think of the CMU campus as a directed graph $G = (V, E, C)$. Here, vertices $v_i, v_j \in V$ correspond to the i^{th} and j^{th} landmark, e.g. the Wean café and the 1^{st} floor of Porter, the directed edge $(i, j) \in E$ is the directed path from v_i to v_j, and the capacity $c_{ij} \in C$ is the maximum number of convex optimization students that can pass through (i, j). Students start from v_s, our classroom in Wean, and move along the directed edges towards v_t, Ryan's office. We assume there are no edges that end in v_s or originate in v_t.

The TAs decide to block paths by building barricades. However, they want to do as little physical labor as possible, so they only want to block the tightest path (i.e. smallest total capacity) in a way that still prevents every student from reaching Ryan's office.

In other words, the TAs want to find a partition, or cut, $C = (S, T)$ of V, such that $v_s \in S$ and $v_t \in T$ and it has minimum capacity. The capacity of a cut is defined as:

$$c(S, T) = \sum_{(i,j)\in E} b_{ij} c_{ij}$$

where $b_{ij} = 1$ if $v_i \in S$ and $v_j \in T$, and $b_{ij} = 0$ otherwise.

The TA's min cut problem can be formulated as follows:

$$\begin{aligned} \min_{b\in\mathbb{R}^{|E|},x\in\mathbb{R}^{|V|}} \quad & \sum_{(i,j)\in E} b_{ij}c_{ij} \\ \text{subject to} \quad & x_s = 1, x_t = 0 \\ & b_{ij} \geq x_i - x_j \\ & b_{ij}, x_i, x_j \in \{0,1\} \\ & \text{for all } (i,j) \in E \end{aligned} \tag{1}$$

(i. 1pt) Explain what the variables x_i and x_j for all $(i,j) \in E$ mean and why the introduction of these variables is necessary (hint: what would happen if the x_i, x_j variables weren't introduced?).

(ii. 1pt) The problem in (1) is an integer linear program (ILP), because its variables take integer values. Because ILPs are mostly difficult to solve, they are often relaxed to LPs. Consider the following relaxation of the integer constraints in (1):

$$\begin{aligned} \min_{b\in\mathbb{R}^{|E|},x\in\mathbb{R}^{|V|}} \quad & \sum_{(i,j)\in E} b_{ij}c_{ij} \\ \text{subject to} \quad & b_{ij} \geq x_i - x_j \quad \text{for all } (i,j) \in E \\ & b \geq 0 \\ & x_s - x_t \geq 1 \end{aligned} \tag{2}$$

How does the optimal value of the original ILP, $f^\star_{ILP}$, compare to the optimal value of the relaxed LP, $f^\star_{LP}$?

(iii. 6pts) Next, derive the dual of (2). Use the following dual variables $f \in \mathbb{R}^{|E|}, y \in \mathbb{R}^{|E|}, w \in \mathbb{R}$ corresponding to the constraints in the order they appear in (2).

(iv. 2pts) What does each constraint of the dual you derived in (iii.) mean in the setting of our path-blocking problem? Hint: the dual of the relaxed min-cut problem is called max-flow.

(v. 1pt) Finally, how does the optimal value of the relaxed LP, $f^\star_{LP}$, compare to the optimal value of the dual, $f^\star_{dual}$?

(vi. 1pt) Interestingly, a well-known theorem (the max-flow min-cut theorem) tells us is that the original ILP and the max flow problem have equal optimal criterion values. What does this result imply about the tightness of the convex relaxation of the ILP?

(vii. 2pts) Consider the setting of our path-blocking problem in Figure 1. The capacities of all edges are shown in the figure, and the min cut has been drawn. Which paths will the TAs barricade? What is the value of the max flow in this problem?

2 Practice with KKT conditions and duality (17 points) [Justin]

(a) Take the LP:

$$\min_x \; c^T x \text{ such that } Ax = b \text{ and } x \geq 0 \tag{3}$$

(where the inequality is defined element-wise) and now consider the second, similar optimization problem

$$\min_x \; c^T x - \tau \sum_i \log(x_i) \text{ such that } Ax = b \tag{4}$$

The second term in the objective is sometimes called the log barrier function, and acts as a 'soft' inequality constraint, because it will tend to positive infinity as any of the x_i tend to zero from the right.

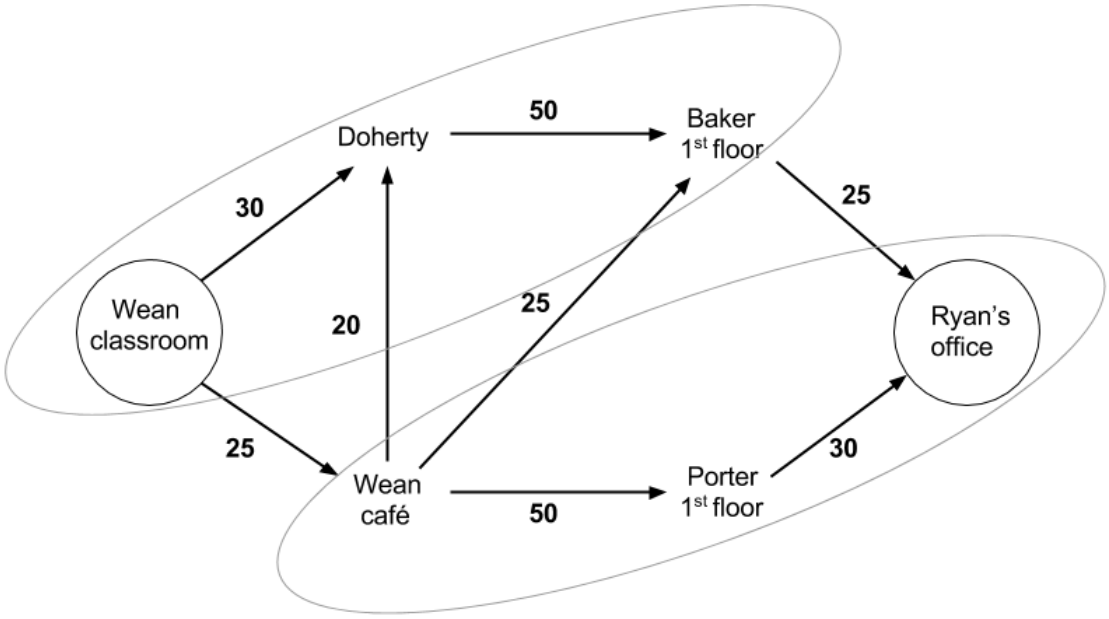

Figure 1: Min cut of the path-blocking problem

(i, 2pts) Derive the dual of the original LP.

(ii, 2pts) Then derive the KKT of original LP in (3).

(iii, 2pts) Then derive the KKT of the second problem with the log barrier problem in (4).

(iv, 2pts) Describe the differences in the two KKT conditions. (Hint: what can you observe about the second set of KKT conditions when τ is taken to be large?)

Throughout, assume that $\{x : x > 0, Ax = b\}$ and $\{y : A^T y > -c\}$ are non-empty. i.e. the primal LP and its dual are both strictly feasible.

(b, 9 pts) Take the least squares regression problem (for $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$):

$$\min_{\beta \in \mathbb{R}^p} \left(\|y - X\beta\|_2\right)^2 \tag{5}$$

Prove that an equivalent dual of this problem is

$$\min_{v \in \mathbb{R}^n} \|y - v\|_2^2 \text{ subject to } X^T v = 0 \tag{6}$$

(Hint: in deriving the dual, you may start by introducing the auxiliary variable $z = X\beta$.) What is the relationship between the primal and the dual solutions, implied by the KKT conditions? Explain why this relationship makes sense, given what you know about projections onto linear subspaces.

3 Convex conjugate and Moreau decomposition (18 pts) [Han]

The convex conjugate of a function $h : \mathbb{R}^n \mapsto \mathbb{R}$ is defined as follows:

$$h^*(x) = \sup_{y \in \text{dom}(h)} x^T y - h(y)$$

Also recall that the proximal operator for function $h(\cdot)$ with $t > 0$ is defined as:

$$\text{prox}_{th}(x) = \underset{z}{\text{argmin}} \frac{1}{2}||z - x||_2^2 + th(z)$$

(a, 3 pts) Show that $(th)^*(x) = th^*(x/t)$.

(b, 4 pts) Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a closed and convex function. Show that $y \in \partial f(x) \Leftrightarrow x \in \partial f^*(y)$. (*Hint: Recall that for closed and convex function f, we have $f^{**} = f$).*

(c, 4 pts) Show that

$$x = \text{prox}_{th}(x) + t\,\text{prox}_{\frac{1}{t}h^*}(x/t)$$

This is known as the *Moreau decomposition.* (*Hint: Feel free to use the results from (a) and (b) to prove this theorem. The subgradient optimality condition may be useful here.*)

(d, 3 pts) Let $h(y) = ||y||$ be a norm of y. Prove that its conjugate is $h^*(x) = \mathbb{I}_{\{z:||z||_* \leq 1\}}(x)$.

(e, 4 pts) Let $h(z) = ||z||_\infty$, where $||z||_\infty$ is defined as $||z||_\infty = \max_{i=1,\ldots,n} |z_i|$. Compute the proximal operator $\text{prox}_{th}(x)$ of $h(z) = ||z||_\infty$. Note that for this question you do not need to give an analytic solution for the prox operator. As long as you believe each part of your answer to be directly computable by a known algorithm, this is fine. (*Hint: You may find the results from (c) and (d) helpful.*)

4 Support vector machines and duality (23 points) [Christoph & Alnur]

In binary classification, we are, roughly speaking, interested in finding a hyperplane that separates two clouds of points living in, say, $\mathbb{R}^p$. The support vector machine (SVM), which we covered a little in class, is a pretty popular method for doing binary classification; to this day, it's (still) used in a number of fields outside of just machine learning and statistics.

One issue with the standard SVM, though, is that it doesn't work well in situations where we pay a higher "price" for misclassifications of one of the two point-clouds. For example, a bank will probably want to be quite certain that a customer won't default on their loan before deciding to give them one (here, the "price" that we pay is monetary). In this problem, you will develop a variant of the standard SVM that addresses these issues, called the *cost-sensitive* SVM. You will implement your own cost-sensitive SVM solver (in part (b) of this question), but as a starting point, we will first investigate the cost-sensitive SVM dual problem (in part (a) of this question).

Throughout, we assume that we are given n data samples, each one taking the form (x_i, y_i), where $x_i \in \mathbb{R}^p$ is a feature vector and $y_i \in \{-1, +1\}$ is a class. In order to make our notation more concise, we can transpose and stack the x_i vertically, collecting these feature vectors into the matrix $X \in \mathbb{R}^{n \times p}$; doing the same thing with the y_i lets us write $y \in \{-1, +1\}^n$. It will also be useful for us to define the following sets, containing the indices of the positive (i.e., those with $y_i = +1$) and negative (i.e., those with $y_i = -1$) samples, respectively:

$$\mathcal{S}_1 = \{i \in \{1, \ldots, n\} : y_i = +1\}, \quad \mathcal{S}_2 = \{i \in \{1, \ldots, n\} : y_i = -1\}.$$

Part (a)

One simple way to incorporate misclassification costs into the standard SVM formulation, is to pose the following (primal) cost-sensitive SVM optimization problem:

$$\begin{array}{ll} \underset{\beta \in \mathbb{R}^p,\, \beta_0 \in \mathbb{R},\, \xi \in \mathbb{R}^n}{\text{minimize}} & (1/2)\|\beta\|_2^2 + C_1 \sum_{i \in \mathcal{S}_1} \xi_i + C_2 \sum_{i \in \mathcal{S}_2} \xi_i \\ \text{subject to} & \xi_i \geq 0, \quad y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \quad i = 1, \ldots, n, \end{array} \tag{7}$$

where $\beta \in \mathbb{R}^p,\ \beta_0 \in \mathbb{R},\ \xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$ are our variables, and C_1, C_2 are positive costs, chosen by the implementer. (Just to remind you of some of the intuition here: when $C_1 = C_2$, problem (7) can be viewed as another way of writing a squared ℓ_2-norm penalized hinge loss minimization problem.)

(i, 2pts) Does strong duality hold for problem (7)? Why or why not? (Your answer to the latter question should be short.)

(ii, 3pts) Derive the Karush-Kuhn-Tucker (KKT) conditions for problem (7). Please use $\alpha \in \mathbb{R}^n$ for the dual variables (i.e., Lagrange multipliers) associated with the constraints "$y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i,\ i = 1, \ldots, n$", and $\mu \in \mathbb{R}^n$ for the dual variables associated with the constraints "$\xi_i \geq 0,\ i = 1, \ldots, n$".

(iii, 3pts) Show that the cost-sensitive SVM dual problem can be written as

$$\begin{array}{ll} \underset{\alpha \in \mathbb{R}^n}{\text{maximize}} & -(1/2)\alpha \tilde{X}\tilde{X}^T\alpha + 1^T\alpha \\ \text{subject to} & y^T\alpha = 0, \quad 0 \leq \alpha_{\mathcal{S}_1} \leq C_1 1, \quad 0 \leq \alpha_{\mathcal{S}_2} \leq C_2 1, \end{array} \tag{8}$$

where $\tilde{X} \in \mathbb{R}^{n \times p} = \text{diag}(y)X$, $\alpha_{\mathcal{S}}$ means selecting only the indices of α that are in the set $\mathcal{S}$, and the 1's here are vectors (of the appropriate and possibly different sizes) containing only ones.

(iv, 2pts) Give an expression for the optimal β in terms of the optimal α variables. Explain why, using just a couple sentences, the optimal β can be thought of as "cost-sensitive".

(v, 1pt) What kind of problem class are both (7) and (8)? You may choose none, one, or more than one of the following:

- linear program
- quadratic program
- second-order cone program
- semidefinite program
- cone program

Part (b)

Please submit your code as an appendix to this problem.

(i, 4pts) Implement the primal SVM in problem (7) using a standard QP solver, typically available as "quadprog" function (for example in Matlab, R or in `Mathprogbase.jl` in Julia). Load a small synthetic toy problem with inputs $X \in \mathbb{R}^{100 \times 2}$ and labels $y \in \{-1, 1\}^{100}$ from `toy.hdf5` (HDF5 file format) and solve the primal SVM with (1) $C_1 = C_2 = 1$, (2) $C_1 = 1, C_2 = 10$ and (3) $C_1 = 10, C_2 = 1$. For each pair of penalty parameters report the objective value of the optimal solution.

(ii, 2pts) For each parameter pair, show a scatter plot of the data and plot the decision border (where the predicted class label changes) as well as the boundaries of the margin (the area in which there is a nonzero penalty for predicting any label) on top. Also highlight the data points i that lie on the wrong side of the margin, that is, points with $\xi_i > 0$. How and why does the decision boundary change with different penalty parameters?

(iii, 2pts) Looking back at the KKT conditions derived in part (a, ii) and the form of the primal solution in part (a, iv), what can be said about the influence of the data points that lie strictly on the right side of the margin (points i with $y_i(x_i^\top \beta + \beta_0) > 1$)? How would the decision boundary change if we removed these data points from the dataset and recomputed the optimal solution? (Give a qualitative answer, no need to actually implement that)

(iv, 3pts) Implement now the dual SVM in problem (8) using again a standard QP solver and report the optimal objective value of the dual for the same penalty parameters as in (i). What can in general be said about the location of a data point $i \in \mathcal{S}_k$ with respect of the boundary of the margin if

- $\alpha_i = 0$;
- $\alpha_i \in (0, C_k)$;
- $\alpha_i = C_k$?

For each pair of penalty parameters, plot the signed distance to the decision boundary of each datapoint i obtained from the primal SVM $y_i(x_i^\top \beta + \beta_0)$ against dual variables α_i obtained from the dual SVM.

(v, 1pt) Cost-sensitive SVMs minimize the (regularized) cost-sensitive hinge-loss, a convex upper bound on the weighted classification error. Predict the class labels for each data point (of the same set that the SVM was trained on) and report the total weighted classification error. A datapoint incurs a loss of C_1 if the true label is $+1$ and -1 is predicted and C_2 if $+1$ is predicted for a data point with true label -1.