Homework 3

Convex Optimization 10-725

Due Friday, October 12 at 11:59pm

Submit your work as a single PDF on Gradescope, including the source code.
Make sure to prepare your solution to each problem on a separate page.
On Gradescope, please select source code along with the corresponding problem.
Please choose either Q1 or Q2 (Score = $\max(\text{Q1}, \text{Q2}) + \text{Q3} + \text{Q4}$).

Total: 75 points

1 Duality in Linear Programs (20 pts) [Akash]

(a, 2pts each) Derive the duals of the following LPs

i, 2pts) $\max_x 2x_1 + x_2$ subject to $x_1 - x_2 \leq 4, x_1 - x_2 \leq 2, x_1 \geq 0, x_2 \geq 0$

ii, 2pts) $\max_x 2x_1 + x_2$ subject to $-x_1 - x_2 \leq -4, x_1 + x_2 \leq 2, x_1 \geq 0, x_2 \geq 0$

iii, 2pts) $\max_x 2x_1 + x_2$ subject to $-x_1 + x_2 \leq -4, x_1 - x_2 \leq 2, x_1 \geq 0, x_2 \geq 0$

What can you say about the primal and dual feasibility and optimal values in all these settings?

(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. 2pts) 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. 2pts) 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?

2 Practice with KKT conditions and duality (20 points) [Po-Wei]

(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.

(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) The Kanotorovich inequality (*BV Additional Exercise 4.14).*

(i, 6pts) Suppose $a \in \mathbb{R}^n$ with $a_1 \geq a_2 \geq a_3 > ... \geq a_n > 0$, and $b \in \mathbb{R}^n$ with $b_k = 1/a_k$. Derive the KKT conditions for the convex optimization:

$$\begin{aligned} \min \quad & -\log(a^T x) - \log(b^T x) \\ \text{subject to} \quad & x \in \mathbb{R}^n_+, \qquad \mathbf{1}^T x = 1 \end{aligned}$$

Where $\mathbb{R}^n_+$ is the positive reals. Show that $x = (1/2, 0, ..., 0, 1/2)$ is the optimal solution.

(ii, 6pts) Suppose $A \in \mathbb{S}^n{}_{++}$ (set of symmetric positive definite matrices) with eigenvalues λ_k sorted in decreasing order. Apply the result of part (b.i), with $a_k = \lambda_k$, to prove the Kantorovich inequality:

$$2(u^T A u)^{1/2}(u^T A^{-1} u)^{1/2} \leq \sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}$$

for all u with $\|u\|_2 = 1$.

3 Screening rules for support vector machines (28 points) [Ryan]

As we've seen, the KKT conditions can be an extremely useful tool. In machine learning, a series of papers have emerged that use the KKT conditions to derive what are called *screening rules*, originally developed in the context of ℓ_1 regularization problems.[1] These are analytic (closed-form) rules that we can apply to any given data set $(x_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}$, $i = 1, \ldots, n$, to determine *a priori* that certain dimensions of the feature space $\mathbb{R}^p$ would not contribute to (say) the lasso or logistic lasso solution, and thus these could be "safely" eliminated before solving (say) the lasso or logistic lasso optimization problem. The rules are usually based on manipulation of the KKT conditions, and typically, properties the solution to the optimization at hand at a "nearby" tuning parameter value.

Screening rules have also been developed for support vector machine (SVMs). In this problem, we'll follow one of the early analyses[2] and look at a constrained version of SVMs: given $y_i \in \{-1, 1\}$ and $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, we solve

$$\min_w \ \frac{1}{2}\|w\|_2^2 \ \text{ subject to } \ \sum_{i=1}^n [1 - y_i f_w(x_i)]_+ \leq s, \tag{5}$$

where $f_w(x) = w^T x$. We write $w^*(s)$ for the unique solution in (5). For convenience, we abbreviate $f_{w^*(s)}(x)$ by $f^*(x|s)$. We also abbreviate $J(w) = (1/2)\|w\|_2^2$ and $H(w) = \sum_{i=1}^n [1 - y_i f_w(x_i)]_+$.

[1] It all started with El Ghaoui et al. (2010): http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-126.pdf.

[2] See http://jmlr.csail.mit.edu/proceedings/papers/v28/ogawa13b.pdf; you may read this paper if it helps, but you must write out arguments to all parts of this problem in your own words.

(a, 5pts) Prove that problem (5) and

$$\min_w \ \frac{1}{2}\|w\|_2^2 \ \text{ subject to } \ \sum_{i \notin \mathcal{R}} [1 - y_i f_w(x_i)]_+ \leq s,$$

have the same solution, where $\mathcal{R} = \{i : y_i f^*(x_i|s) > 1\}$. In other words, show that the instances $i \in \mathcal{R}$ do not affect the solution of (5), and can hence be safely discarded. Hint: look at the KKT conditions on all n points, and on only on the points in $\mathcal{R}$.

(b) Fix any $s_a > s_b$.

(i, 3pts) Show that $s \in [s_b, s_a] \implies J(w^*(s)) \leq J(w^*(s_b))$.

(ii, 4pts) Show that $s \in [s_b, s_a] \implies w^*(s_a)^T(w^*(s) - w^*(s_a)) \geq 0$. Hint: consider the KKT conditions for (5), consider subgradients of $H(w)$, and primal feasibility of $w^*(s)$ for (5) when the tuning parameter is s_a.

(iii, 3pts) Show that we may safely discard a point i from the optimization in problem (5) with tuning parameter $s \in [s_b, s_a]$ if $g_{[s_b,s_a]}(i) > 1$ where:

$$g_{[s_b,s_a]}(i) = \min_{w \in \Theta_{[s_b,s_a]}} \ y_i f_w(x_i), \tag{6}$$

and $\Theta_{[s_b,s_a]} = \{w : J(w) \leq J(w^*(s_b)) \wedge w^*(s_a)^T(w - w^*(s_a)) \geq 0\}$ (and we use the shorthand $u \wedge v = \min\{u, v\}$).

(c) We now reduce the screening rule of $g_{[s_b,s_a]}(i) > 1$ to an analytical formula below. Let $\gamma_b = J(w_b^*)$ and $\gamma_a = J(w_a^*)$.

(i, 3pts) Write out the Lagrangian for the problem (6) with Lagrange multipliers of μ and ν for constraints $J(w) \leq \gamma_b$ and $w^*(s_a)^T(w - w^*(s_a)) \geq 0$ respectively.

(ii, 3pts) Write out the KKT conditions for the problem (6). Use these conditions to get an expression for solution to this problem (call it) z in terms of the optimal dual variables μ, ν, and furthermore, an expression for $y_i z^T x_i$ in terms of μ, ν and $f^*(x_i|s_a)$.

(iii, 4pts) Show that:

$$g_{[s_b,s_a]}(i) = \begin{cases} -\sqrt{2\gamma_b}\|x_i\| & \text{if } -\frac{y_i f^*(x_i|s_a)}{\|x_i\|} \geq \frac{\sqrt{2}\gamma_a}{\sqrt{\gamma_b}} \\ y_i f^*(x_i|s_a) - \sqrt{\frac{\gamma_b - \gamma_a}{\gamma_a}\left(2\gamma_a\|x_i\|_2^2 - f^*(x_i|s_a)^2\right)} & \text{otherwise.} \end{cases}$$

Hint: use the sign of $-\frac{y_i f^*(x_i|s_a)}{\|x_i\|} - \frac{\sqrt{2}\gamma_a}{\sqrt{\gamma_b}}$ to guide whether $\nu = 0$.

(iv, 3pts) Further simplify the screening rule of $g_{[s_b,s_a]}(i) > 1$ to:

$$y_i f^*(x_i|s_a) > 1 \ \text{ and } \ y_i f^*(x_i|s_a) - \sqrt{\frac{\gamma_b - \gamma_a}{\gamma_a}\left(2\gamma_a\|x_i\|_2^2 - f^*(x_i|s_a)^2\right)} > 1.$$

4 Support vector machines and duality (27 points) [Wenbo]

In binary classification, we are interested in finding a hyperplane that separates two clouds of points living in, say, $\mathbb{R}^p$. The support vector machine (SVM), which we talked about 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 arises with the standard SVM, though, when the data points are not linearly separable in $\mathbb{R}^p$, i.e., we cannot find a hyperplane which separates the two classes of points. In such cases, it is often useful to map the data points to a different space (potentially of higher dimension than $\mathbb{R}^p$) where the points become separable. Such maps are called nonlinear feature maps.

In this problem, you will develop a SVM with the RBF kernel to address the nonlinearly separable problem of the standard SVM. You will implement your own RBF-SVM in part (b) of this question, but as a starting point, we will first investigate the 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 $y \in \{-1, +1\}^n$.

Part (a)

The primal problem of SVM with slack variables is

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

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 is a positive margin coefficient chosen by the implementer. (Just to remind you of some of the intuition here: 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 SVM dual problem can be written as

$$\begin{array}{ll} \underset{\alpha\in\mathbb{R}^n}{\text{maximize}} & -(1/2)\alpha^T\tilde{X}\tilde{X}^T\alpha + 1^T\alpha \\ \text{subject to} & \alpha^T y = 0, \\ & 0 \leq \alpha \leq C1, \end{array} \tag{8}$$

where $\tilde{X} \in \mathbb{R}^{n\times p} = \text{diag}(y)X$, α is the dual variable, 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 and explain how.

(v, 2pt) 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

Now we are going to take a glimpse of the "magic" of kernels. Let's first see what is a kernel. Given a feature map $\phi : \mathbb{R}^d \to \mathcal{K}$, where $\mathcal{K}$ is a Hilbert space (i.e., a vector space with inner product), the kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the corresponding inner product function

$$K(x_i, x_j) \coloneqq \langle \phi(x_i), \phi(x_j) \rangle. \tag{9}$$

Here the feature map, as we mentioned earlier, is used to "embed" the original data into a higher dimensional space such that they become separable. Recall the objective of the dual SVM, and it can be rewritten as

$$\begin{aligned} & -\tfrac{1}{2}\alpha^T \tilde{X}\tilde{X}^T \alpha + 1^T \alpha && (10)\\ \Leftrightarrow & -\tfrac{1}{2}\alpha^T Y X X^T Y \alpha + 1^T \alpha && (11)\\ \Leftrightarrow & -\tfrac{1}{2}\alpha^T Y G Y \alpha + 1^T \alpha, && (12)\\ & && (13) \end{aligned}$$

where $Y = \mathrm{diag}(y)$, and $G = XX^T$ is the so called Gram matrix, $G_{ij} = \langle x_i, x_j \rangle$. One nice property of the Gram matrix of a kernel K is that

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = G_{ij}. \tag{14}$$

Hence, the kernel builds a bridge between the feature maps and the original dual problem.

(vi, 3pts) Show that the Gram matrix of a kernel K is positive semidefinite. Let the dimension of the feature space after the feature map be p'. If $p << p'$, which one is more efficient to solve, the primal or the dual? Why?

Now we are going to probe into the infinite dimensional space. We have seen so far how to build a kernel from a given feature map, but can we do the opposite? Suppose a map K is a kernel, can we find the corresponding feature map ϕ such that $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{K}}$? Fortunately, thanks to the Mercer's theorem, we know that we are able to construct the feature map by finding the eigenfunctions of the integral operator with the kernel.

There is no need to go into such difficulty of finding the feature maps, however, since we have the kernel-feature map equivalence (14). We only need to compute the value of the kernel function, avoiding the complexity of computing the inner product of high dimensional feature maps.

Given this intuition, we consider the radial basis function (RBF) kernel

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \exp\left(-\gamma \|x_i - x_j\|^2\right), \tag{15}$$

where γ controls the bandwidth of the kernel. For RBF kernel, the corresponding feature maps have infinite dimensional feature spaces. The RBF kernel is a reasonable measure of x_i and x_j's similarity, and is close to 1 when x_i and x_j are close, and near 0 when they are far apart. In the following problems, you are going to use the RBF kernel in SVM.

Part (b)

Please submit your code as an appendix to this problem.

(i, 4pts) Implement the dual SVM in problem (8) with the RBF kernel using a standard QP solver (typically available as "quadprog" function in Matlab, R, or in `Mathprogbase.jl` in Julia; you may also refer `CVXOPT` in Python, `GORUBI`, or `MOSEK`). Load a small synthetic toy problem with inputs $X \in \mathbb{R}^{863\times 2}$ and labels $y \in \{-1, 1\}^{863}$ from `data.txt` and solve the dual SVM with $\gamma = \{10, 50, 100, 500\}$ and $C = \{0.01, 0.1, 0.5, 1\}$. Report the optimal objective values of the dual.

(ii, 2pts) For each of the parameter pairs, show a scatter plot of the data and plot the decision border (where the predicted class label changes) on top. How and why does the decision boundary change with different pair of parameters?

(iii, 2pts) For each of the parameter pairs, identify the support vectors (i.e., data points with nonzero α_is; in implementation select $\alpha > 1e^{-5}$) in the plots, and report the number of support vectors. What can in general be said about the location of a data point i with respect of the boundary of the margin if

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

(iv, 2pts) Looking back at the KKT conditions derived in part (a, ii), what can be said about the influence of the data points that lie strictly on the correct side of the margin? 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.)

(v, 2pt) SVM minimizes the ℓ_2-regularized hinge-loss, a convex upper bound on the classification error. For each of the above parameter pairs (C, γ), predict the class labels for each data point (of the same set that the SVM was trained on). Report the classification error for each class and the total classification error.

Bonus Implement the screening rules for SVM derived in problem 3.