EECS 127/227AT **Fall 2020**
Solutions for Discussion 9

1 Sphere Enclosure

For $i = 1, \ldots, m$, let B_i be a ball in $\mathbb{R}^n$ with center x_i, and radius $\rho_i \geq 0$. We wish to find a ball B of minimum radius that contains all the B_i for $i = 1, \ldots, m$. Cast this problem as an SOCP.

Solution:

Let $c \in \mathbb{R}^n$ and $r \geq 0$ denote the center and radius of the enclosing ball B, respectively. We express the given balls B_i as

$$B_i = \{x : x = x_i + \delta_i,\ \|\delta_i\|_2 \leq \rho_i\}, \quad i = 1, \ldots, m.$$

We have that $B_i \subseteq B$ if and only if

$$\max_{x \in B_i} \|x - c\|_2 \leq r.$$

Note that

$$\max_{x \in B_i} \|x - c\|_2 = \max_{\|\delta_i\|_2 \leq \rho_i} \|x_i - c + \delta_i\|_2 = \|x_i - c\|_2 + \rho_i.$$

The last step follows by choosing δ_i in the direction of $x_i - c$.
The problem is then cast as the following SOCP:

$$\begin{aligned} &\min_{c,r} r \\ \text{subject to: } &\|x_i - c\|_2 + \rho_i \leq r, i = 1, \ldots, m. \end{aligned}$$

2 Dual Norms and SOCP

Consider the problem

$$p^* := \min_{x \in \mathbb{R}^n} \ \|Ax - y\|_1 + \mu \|x\|_2, \tag{1}$$

where $A \in \mathbb{R}^{m \times n}$, $y \in \mathbb{R}^m$, and $\mu > 0$.

a) Express this (primal) problem in standard SOCP form.

Solution:
Introducing slack variables $z \in \mathbb{R}^m$, $t \in \mathbb{R}$, we can write the problem as

$$\begin{aligned} \min_{x,z,t} \ & z^\top \mathbb{1} + \mu t \\ \text{s.t. } & |(Ax)_i - y_i| \le z_i, \quad i = 1, \dots, m, \\ & \|x\|_2 \le t. \end{aligned} \tag{2}$$

This is an SOCP. To see this, note first that the objective is a linear function of the variables (x, z, t). The constraint $\|x\|_2 \le t$ is an SOC constraint because x is an affine function (in fact it turns out to be a linear function) of the variables (x, z, t) and t is also an affine function (in fact it turns out to be a linear function) of the variables (x, z, t). Each of the constraints $|(Ax)_i - y_i| \le z_i$ for $1 \le i \le m$ is an SOC constraint because for a scalar the absolute value is the same as the ℓ_2 norm, $(Ax)_i - y_i$ is an affine function of the variables (x, z, t), and z_i is an affine function of the variables (x, z, t).

b) Find the conic dual of the primal SOCP and express it as an SOCP.

Hint: Recall that for every vector z, the following dual norm equalities hold:

$$\|z\|_2 = \max_{u \,:\, \|u\|_2 \le 1} u^\top z, \quad \|z\|_1 = \max_{u \,:\, \|u\|_\infty \le 1} u^\top z.$$

Solution:
Using the hint, we can rewrite the objective function of the original problem in (1) as

$$\|Ax - y\|_1 + \mu \|x\|_2 = \max_{u : \|u\|_\infty \le 1} u^\top (Ax - y) + \mu \max_{v : \|v\|_2 \le 1} v^\top x.$$

We can then express the original (primal) problem as

$$p^* = \min_x \max_{u,v : \|u\|_\infty \le 1, \|v\|_2 \le 1} u^\top (Ax - y) + \mu v^\top x.$$

To form the dual, we reverse the order of min and max:

$$\begin{aligned} d^* &:= \max_{u,v : \|u\|_\infty \le 1, \|v\|_2 \le 1} \min_x \ u^\top (Ax - y) + \mu v^\top x \\ &= \max_{u,v : \|u\|_\infty \le 1, \|v\|_2 \le 1} g(u, v), \end{aligned}$$

where g is defined as

$$\begin{aligned} g(u, v) &:= \min_x \ u^\top (Ax - y) + \mu v^\top x \\ &= \min_x \ (u^\top A + \mu v^\top) x - u^\top y \\ &= \begin{cases} -u^\top y & \text{if } A^\top u + \mu v = 0, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

We can thus rewrite the dual problem as

$$\begin{aligned} d^* = \max_{u,v} \; & -u^\top y \\ \text{s.t. } & A^\top u + \mu v = 0, \\ & \|u\|_\infty \leq 1, \;\; \|v\|_2 \leq 1. \end{aligned} \tag{3}$$

This is already an SOCP (as a maximization problem). To see this, note that the objective function is linear in the variables (u, v). The constraint $\|v\|_2 \leq 1$ is an SOC constraint and $\|u\|_\infty \leq 1$ can be viewed as coordinate by coordinate SOC constraints. The constraint $A^\top u + \mu v = 0$ can be viewed as coordinate by coordinate constraints saying that each coordinate is bounded in absolute value by 0 and these are SOC constraints.

It is also instructive to see that the same dual can be arrived at by following the procedure in Sec. 10.1.3 in the textbook of Calafiore and El Ghaoui, starting with the primal problem in SOCP form as in (2). For this, we write

$$\begin{aligned} p^* &= \min_{x,z,t} \max_{\alpha \geq 0, \beta \geq 0} \; z^\top \mathbb{1} + \mu t + \sum_{i=1}^m \alpha_i (|(Ax)_i - y_i| - z_i) + \beta(\|x\|_2 - t) \\ &= \min_{x,z,t} \max_{\|w\|_2 \leq \beta, |u_i| \leq \alpha_i, 1 \leq i \leq m} \; \sum_{i=1}^m u_i((Ax)_i - y_i) - \sum_{i=1}^m (\alpha_i - 1) z_i + w^\top x - (\beta - \mu) t. \end{aligned}$$

Interchanging the max and min leads us to study

$$\min_{x,z,t} \; \sum_{i=1}^m u_i((Ax)_i - y_i) - \sum_{i=1}^m (\alpha_i - 1) z_i + w^\top x - (\beta - \mu) t,$$

which is $-\infty$ unless $A^\top u + w = 0$, $\beta = \mu$, and $\alpha_i = 1$ for all $i = 1, \ldots, m$. This leads to the dual problem

$$\begin{aligned} \max_{u,w} \; & -u^\top y \\ \text{s.t. } & A^\top u + w = 0, \\ & \|u\|_\infty \leq 1, \;\; \|w\|_2 \leq \mu, \end{aligned}$$

which is the same as (3) after identifying w with μv.

The process by which we found the dual of the primal problem in SOCP form, i.e. (2), basically involved conic duality on the conic constraints, and so the resulting dual problem is called the *conic dual*. This is of course the same as the dual problem we initially found, starting with the primal problem in the form (1).

The dual in (3) can be simplified further by noting that the equality constraint fully restricts the value of v — rewriting it, $v = -\frac{A^\top u}{\mu}$ — we can plug this value into the third constraint and eliminate v from our optimization problem altogether, which gives

$$\begin{aligned} d^* = \max_{u} \; & -u^\top y \\ \text{s.t. } & \|A^\top u\|_2 \leq \mu \\ & \|u\|_\infty \leq 1, \end{aligned}$$

which is also an SOCP, as can be checked.

c) Assume strong duality holds and that $m = 100$ and $n = 10^6$, i.e., A is 100×10^6. Which problem would you choose to solve using a numerical solver: the primal or the dual? Justify your answer.

Solution:
To determine the rough computational complexity of each problem, we examine the number of variables and the number of constraints in each problem. The primal SOCP has $\sim 10^6$ variables and 201 constraints (if we count each constraint $|(Ax)_i - y_i|$ as two linear inequality constraints and count the constraint $\|x\|_2 \leq t$ as a single inequality constraint), while the dual has 100 variables and 201 constraints (if we consider the constraint $\|u\|_\infty \leq 1$ as comprised of two linear inequality constraints for each coordinate of u and, after eliminating v, which eliminates the equality constraint, we count the constraint $\|A^\top u\|_2 \leq \mu$ as a single inequality constraint). The dual problem would thus be expected to be much more efficient to solve.

Remark (optional): Section 10.1.3 of the textbook of Calafiore and El Ghaoui consider a primal SOCP in the form:

$$\begin{aligned} \min_{x} \quad & c^T x \\ \text{s.t.} \quad & \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \;\; i = 1, \ldots, m, \end{aligned} \tag{4}$$

and shows that the conic dual of this SOCP can be written in the form:

$$\begin{aligned} \max_{u,\lambda} \quad & \sum_{i=1}^{m} (u_i^T b_i - \lambda_i d_i) \\ \text{s.t.} \quad & \sum_{i=1}^{m} (A_i^T u_i - \lambda_i c_i) = -c, \\ & \|u_i\|_2 \leq \lambda_i, \;\; i = 1, \ldots, m, \end{aligned} \tag{5}$$

which is also an SOCP (note that the equality constraints for the dual in the textbook are wrongly written). The process by which we got from our primal problem in the form of (2) to the dual problem in (3), i.e. the conic dual of the primal, is the same as that in the textbook.

In fact the conic dual of the primal problem in SOCP form is equivalent to the traditional Lagrangian dual of a problem that is equivalent to the primal. To see this, note that we can write the primal SOCP in (4) in the equivalent form:

$$\begin{aligned} \min_{x,y,t} \quad & c^T x \\ \text{s.t.} \quad & \|y_i\|_2 \leq t_i, \;\; i = 1, \ldots, m, \\ & y_i = A_i x + b_i, \;\; i = 1, \ldots, m, \\ & t_i = c_i^T x + d_i, \;\; i = 1, \ldots, m. \end{aligned} \tag{6}$$

We can then write the Lagrangian for this problem, which has primal variables (x, y, t), using dual variables $(\lambda, \nu_1, \ldots, \nu_m, \mu)$, where $\lambda \in \mathbb{R}^m$ is the vector of dual variables for the first m inequality constraints, $\nu_i \in \mathbb{R}^{k_i}$ is the vector of dual variables for the i-th of the first set of m vector equality constraints, $1 \leq i \leq m$, and $\mu \in \mathbb{R}^m$ is the vector of dual variables for the last set of equality constraints. We can then go through the usual process of finding the dual objective function, which will turn out to be

$$g(\lambda, \nu_1, \ldots, \nu_m, \mu) = \begin{cases} -\sum_{i=1}^{m} (\nu_i^T b_i + \mu_i d_i) & \text{if } \sum_{i=1}^{m} (A_i^T \nu_i + \mu_i c_i) = c, \; \|\nu_i\|_2 \leq \lambda_i, \text{ and } \mu = \lambda, \\ -\infty & \text{otherwise.} \end{cases}$$

From this it will immediately follow that the resulting dual problem is equivalent to the conic dual in (5) by substituting for ν_i with $-u_i$ and eliminating μ (i.e. replacing it with λ).

If we go through this process in the case of the primal in SOCP form in (2), we will see that the resulting form of the primal satsifies Slater's condition. This explains why strong duality holds.

This is of course a rather elaborate and painful way to show strong duality. One can show that strong duality holds much more easily by appealing to a conic version of Slater's condition (see pg. 265 of the textbook of Boyd and Vandenberghe), which is beyond the scope of this class.

3 Robust machine learning

We consider a binary classification problem, where the prediction label associated with a test point $x \in \mathbb{R}^n$ is the form $\hat{y}(x) = \text{sign}(w^T x + v)$, with $(w, v) \in \mathbb{R}^n \times \mathbb{R}$ being the classifier weights. Given a training set X, y, with $X = [x_1, \ldots, x_m] \in \mathbb{R}^{n \times m}$ the data matrix, with data points $x_i \in \mathbb{R}^n$, $i = 1, \ldots, m$, and $y \in \{-1, 1\}^m$ the vector of corresponding labels, the training problem is to minimize the so-called hinge loss function:

$$\min_{w,v} \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i(x_i^T w + v)). \tag{7}$$

We seek to find a classifier (w, v) that can be implemented with low precision (say, as an integer vector). To this end, we modify the training problem so that it accounts for the implementation error, when approximating the original optimal (full precision) weight vector w_* with a low-precision one, $\tilde{w}$. We bound the corresponding error as $\|\tilde{w} - w_*\|_\infty \leq \epsilon$ for some given absolute error bound $\epsilon > 0$; for example, if $\tilde{w}$ is the nearest integer vector, the error is bounded by $\epsilon = 0.5$. Then, we seek to solve the *robust counterpart* to (7):

$$\min_{w,v} \max_{\tilde{w} \,:\, \|\tilde{w} - w\|_\infty \leq \epsilon} \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i(x_i^T \tilde{w} + v)). \tag{8}$$

(a) Justify the use of the hinge loss function in problem (7); in particular, explain geometrically what it means to have a zero loss.

Solution:
The hinge loss function is an upper bound on the so-called $0 - 1$ error loss,

$$\frac{1}{m} \sum_{i=1}^{m} E(y_i(x_i^T w + v)),$$

where E is the function with values $E(\xi) = 1$ if $\xi < 0$, 0 otherwise. An alternate justification is that the function minimizes the distance of wrongly classified points to the decision boundary, which is the hyperplane $\mathcal{H}$ described by the equation $w^T x + v = 0$.

If the loss is zero, we have

$$\forall\, i = 1, \ldots, m \;:\; y_i(x_i^T w + v) \geq 0,$$

which means geometrically that the hyperplane $\mathcal{H}$ separates the positive and negative classes.

(b) Show that without loss of generality, we can set $v = 0$, which we will do henceforth.

Hint: Think about adding a dimension to the data.

Solution:
We can simply append a 1 at the end of each data point x_i, $i = 1, \ldots, m$.

(c) Explain how to obtain a low-precision classifier once problem (8) is solved. What guarantees do we have on the training error?

Solution:
Once w^* is found, we simply replace it by its closest low-precision approximation $\tilde{w}$. Since that approximation satisfies the bound $\|\tilde{w} - w_*\|_\infty \leq \epsilon$, we know that the training error corresponding to the low-precision classifier:

$$\frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i(x_i^T \tilde{w}))$$

is no worse than the worst-case training error, that is, the optimal value of the robust problem (8).

(d) Show that the optimal value of problem (8) is bounded above by

$$\min_w \frac{1}{m}\sum_{i=1}^m \max(0, 1 - y_i(x_i^T w) + \epsilon\|x_i\|_1). \tag{9}$$

Hint: Solve the problem $\max_\delta \{\delta^T z \ : \ \|\delta\|_\infty \leq \epsilon\}$ first.

Solution:
We have

$$\max_{\delta \, : \, \|\delta\|_\infty \leq \epsilon} \delta^T z = \epsilon \cdot \sum_{j=1}^n \max_{\delta_j \, : \, |\delta_j| \leq 1} \delta_j z_j = \epsilon \|z\|_1.$$

Therefore

$$\max_{\|\delta\|_\infty \leq \epsilon} \max(0, 1 - y_i(x_i^T(w + \delta)) = \max(0, 1 - y_i x_i^T w + \epsilon\|x_i\|_1). \tag{10}$$

We now turn to the full loss function. We have

$$\begin{aligned}
&\max_{\tilde{w} \, : \, \|\tilde{w} - w\|_\infty \leq \epsilon} \frac{1}{m}\sum_{i=1}^m \max(0, 1 - y_i(x_i^T \tilde{w})) \\
\leq \ &\frac{1}{m}\sum_{i=1}^m \max_{\tilde{w} \, : \, \|\tilde{w} - w\|_\infty \leq \epsilon} \max(0, 1 - y_i(x_i^T \tilde{w})) \\
= \ &\frac{1}{m}\sum_{i=1}^m \max_{\delta \, : \, \|\delta\|_\infty \leq \epsilon} \max(0, 1 - y_i(x_i^T w + \delta)),
\end{aligned}$$

which, in view of (10), leads to the desired result.

(e) Assume that the data set is normalized, in the sense that $\|x_i\|_1 = 1$, $i = 1, \ldots, m$. How would you solve problem (9) if you had code to solve (7) only?

Solution:
When the data is normalized, problem (9) reads

$$\min_w \frac{1}{m}\sum_{i=1}^m \max(0, 1 - y_i(x_i^T w) + \epsilon).$$

We can divide each term by $1 + \epsilon$, and set $\bar{w} = w/(1 + \epsilon)$; the new problem reads just as (7).