The University of Texas at Austin
Department of Electrical and Computer Engineering

EE381K: Large Scale Optimization — Fall 2015

PROBLEM SET SIX SOLUTIONS

Constantine Caramanis

Written Problems

1. **Compressed Sensing** *Consider the following optimization problem*

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|x\|_1 \\ s.t. \quad & Ax = b \end{aligned}$$

Write this as a linear program. Find its dual.

Solution Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Introducing a new variable $y = |x|$, we can write the optimization problem as

$$\begin{aligned} \min_{x\in\mathbb{R}^n, y\in\mathbb{R}^n} \quad & \mathbf{1}^\top y \\ \text{s.t.} \quad & Ax = b, \\ & x_i + y_i \geq 0, -x_i + y_i \geq 0, \quad i \in \{1,\ldots,n\}. \end{aligned}$$

where $\mathbf{1} = [1, 1, \ldots, 1]^\top$ is the length-n all-ones vector.

The dual of the above LP is given by

$$\begin{aligned} \max_{\lambda\in\mathbb{R}^m, \mu^+, \mu^-\in\mathbb{R}^n} \quad & b^\top\lambda \\ \text{s.t.} \quad & A^\top\lambda + \mu^+ - \mu^- = 0, \quad \mu^+ + \mu^- = \mathbf{1}, \quad \mu^+, \mu^- \geq 0. \end{aligned}$$

2. *Problem 5.7 in the textbook, Boyd and Vandenberghe.*

Solution (a) The Lagrangian dual function is written as

$$\begin{aligned} q(\lambda) &= \inf_{x\in\mathbb{R}^n, y\in\mathbb{R}^m} \left\{ \left(\max_{i=1,\ldots,m} y_i \right) + \sum_{i=1}^m \lambda_i (a_i^\top x + b_i - y_i) \right\} \\ &= \inf_{x\in\mathbb{R}^n, y\in\mathbb{R}^m} \left\{ b^\top\lambda + \left(\sum_{i=1}^m \lambda_i a_i^\top \right) x + \left(\max_{i=1,\ldots,m} y_i - \sum_{i=1}^m \lambda_i y_i \right) \right\}. \end{aligned} \tag{1}$$

We claim that $q(\lambda) = -\infty$ except when

$$\sum_{i=1}^m \lambda_i a_i = 0, \quad \sum_{i=1}^m \lambda_i = 1, \quad \lambda_i \geq 0, \ \forall i \in \{1,\ldots,m\}$$

If the first condition is violated, the second term in (1) can tend to $-\infty$. The rest of the conditions makes the third term in (1) bounded below. To see this, we have

$$\sum_{i=1}^m \lambda_i = 1,\ \lambda \geq 0 \quad \Rightarrow \quad \max_{i=1,\ldots,m} y_i - \sum_{i=1}^m \lambda_i y_i = \sum_{i=1}^m \lambda_i \left(\max_{i=1,\ldots,m} y_i - y_i \right) \geq 0$$

On the other hand, if $\sum_{i=1}^m \lambda_i \neq 1$, we can choose $y_1 = y_2 = \cdots = y_m \to \mathrm{sign}(\sum_{i=1}^m \lambda_i - 1) \cdot \infty$ to obtain

$$\max_{i=1,\ldots,m} y_i - \sum_{i=1}^m \lambda_i y_i = \left(1 - \sum_{i=1}^m \lambda_i\right) \max_{i=1,\ldots,m} y_i + \sum_{i=1}^m \lambda_i \left(\max_{i=1,\ldots,m} y_i - y_i \right) \to -\infty.$$

If $\lambda_j < 0$ for some j, we can choose $y_j \to -\infty$ and $y_i = 0$ for $i \neq j$ to get

$$\max_{i=1,\ldots,m} y_i - \sum_{i=1}^m \lambda_i y_i = -\lambda_j y_j \to -\infty,$$

which completes the proof of the claim.

Regarding this dual feasibility as constraints, we get the dual problem expressed as

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m} \quad & b^\top \lambda \\ \text{s.t.} \quad & \sum_{i=1}^m a_i \lambda_i = 0, \quad \sum_{i=1}^m \lambda_i = 1, \quad \lambda_i \geq 0,\ \forall i \in \{1, \ldots, m\} \end{aligned}$$

(b) The piecewise-linear minimization problem can be reformulated as an LP

$$\begin{aligned} \min_{x \in \mathbb{R}^n, y \in \mathbb{R}} \quad & y \\ \text{s.t.} \quad & y - a_i^\top x \geq b_i, \quad i \in \{1, \ldots, m\} \end{aligned}$$

The dual problem is given by

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m} \quad & b^\top \lambda \\ \text{s.t.} \quad & \sum_{i=1}^m \lambda_i = 1, \quad \sum_{i=1}^m a_i \lambda_i = 0, \quad \lambda_i \geq 0,\ \forall i \in \{1, \ldots, m\} \end{aligned} \tag{2}$$

This is identical to the dual problem derived in (a).

(c) (5.62) is written in our notation as

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m} \quad & b^\top \lambda - \sum_{i=1}^m \lambda_i \log \lambda_i \\ \text{s.t.} \quad & \sum_{i=1}^m \lambda_i = 1, \quad \sum_{i=1}^m a_i \lambda_i = 0, \quad \lambda_i \geq 0,\ \forall i \in \{1, \ldots, m\} \end{aligned} \tag{3}$$

Both (5.105) and (5.106) hold strong duality because they are either an LP or an unconstrained feasible problem. Thus, the p^*_{pal} and p^*_{gp} are equal to the optimal values of (2) and (3), respectively.

Let λ^*_{pwl} and λ^*_{gp} denote the optimal points of (2) and (3). Since $-\sum_{i=1}^m \lambda_i \log \lambda_i$ is nonnegative over the feasible set, the optimum of (3) is always greater than that of (2), i.e.,

$$p^*_{\mathrm{gp}} - p^*_{\mathrm{pwl}} \geq 0.$$

We also have

$$\begin{aligned}
p^*_{\mathrm{gp}} - p^*_{\mathrm{pwl}} &= b^\top \lambda^*_{\mathrm{gp}} - \sum_{i=1}^m \lambda^*_{\mathrm{gp},i} \log \lambda^*_{\mathrm{gp},i} - b^\top \lambda^*_{\mathrm{pwl}} \\
&\leq - \sum_{i=1}^m \lambda^*_{\mathrm{gp},i} \log \lambda^*_{\mathrm{gp},i} \\
&\leq \log m
\end{aligned}$$

where the first inequality follows from that λ^*_{pwl} maximizes $b^\top \lambda$ over the feasible set. The second inequality can be proved by solving

$$\log m = \min_{\lambda \geq 0} \left(- \sum_{i=1}^m \lambda_i \log \lambda_i \right).$$

using the first-order optimality condition.

(d) Consider the equivalent problem

$$\begin{aligned}
\min_{x \in \mathbb{R}^n, y \in \mathbb{R}^m} \quad & \frac{1}{\gamma} \log \left(\sum_{i=1}^m \exp(\gamma y_i) \right) \\
\text{s.t.} \quad & a_i^\top x + b_i = y_i, \quad i \in \{1, \ldots, m\}
\end{aligned}$$

The Lagrangian is given by

$$\begin{aligned}
L(x, y, \lambda) &= \frac{1}{\gamma} \log \left(\sum_{i=1}^m \exp(\gamma y_i) \right) + \sum_{i=1}^m \lambda_i (a_i^\top x + b_i - y_i) \\
&= b^\top \lambda + \left(\sum_{i=1}^m \lambda_i a_i^\top \right) x + \left(\frac{1}{\gamma} \log \left(\sum_{i=1}^m \exp(\gamma y_i) \right) - \sum_{i=1}^m \lambda_i y_i \right). \qquad (4)
\end{aligned}$$

The third term is minimized over y if the gradient is zero, i.e.,

$$\frac{\partial}{\partial y_j} \left(\frac{1}{\gamma} \log \left(\sum_{i=1}^m \exp(\gamma y_i) \right) - \sum_{i=1}^m \lambda_i y_i \right) = \frac{\exp(\gamma y_j)}{\sum_{i=1}^m \exp(\gamma y_j)} - \lambda_j = 0 \qquad (5)$$

for $j \in \{1, \ldots, m\}$. Note that λ should satisfy $\sum_{i=1}^m \lambda_i = 1$ so that we have zero gradient at some y. It follows from (5) that we should have $y_i = \frac{1}{\gamma} \log(\lambda_i \sum_j \exp(\gamma y_j))$. The Lagrangian dual function is then given by

$$
\begin{aligned}
q(\lambda) &= b^\top \lambda + \left(\frac{1}{\gamma} \log \left(\sum_{i=1}^m \exp(\gamma y_i) \right) - \frac{1}{\gamma} \sum_{i=1}^m \lambda_i \log \left(\lambda_i \sum_j \exp(\gamma y_j) \right) \right) \\
&= b^\top \lambda - \frac{1}{\gamma} \sum_{i=1}^m \lambda_i \log \lambda_i
\end{aligned}
$$

where $\mathrm{dom}\, q = \{\lambda : \sum_{i=1}^m \lambda_i = 1, \lambda \geq 0\}$. The dual problem is given by

$$
\begin{aligned}
\max_{\lambda \in \mathbb{R}^m} \quad & b^\top \lambda - \frac{1}{\gamma} \sum_{i=1}^m \lambda_i \log \lambda_i \\
\text{s.t.} \quad & \sum_{i=1}^m \lambda_i = 1, \quad \sum_{i=1}^m a_i \lambda_i = 0, \quad \lambda_i \geq 0, \ \forall i \in \{1, \ldots, m\}
\end{aligned}
$$

As we did in (c), the difference between p^*_{pwl} and the optimal value is bounded by $\frac{1}{\gamma} \log m$. Therefore, the optimal value approaches p^*_{pwl} as γ increases.

3. **Exponential Families** *In this problem we investigate the natural motivation for an important class of distributions: exponential families. Let X be a discrete*[1] *random variable, with possible values $x \in \mathcal{X}$. Given a set of functions $\{\phi_k(x)\}$, the corresponding exponential family is all probability mass functions of the form*

$$
p(x) \;=\; \frac{1}{Z(\theta)} \exp \left(\sum_k \theta_k \phi_k(x) \right) \tag{6}
$$

where all $\theta_k \in \mathbb{R}$ and $Z(\theta)$ is a normalizing constant. Examples include bernoulli, exponential, gaussian, poisson etc.

(a) *Consider the entropy function $H(p) := -\sum_x p(x) \log p(x)$. As is well known, the higher the entropy of a random variable, the "more random" it is. Show that $H(\cdot)$ is a concave function of p.*

Solution Let h denote the function $h(x) = -x \log x$. $h(x)$ is a concave function on $x \geq 0$ because $h''(x) = -1/x < 0$ for $x > 0$, and $h(\lambda x + (1-\lambda)0) = h(\lambda x) \geq \lambda h(x)$ for $x \geq 0$ and $\lambda \in [0, 1]$. Then we get

$$
\begin{aligned}
H(\lambda p + (1-\lambda) q) &= -\sum_{x \in \mathcal{X}} h(\lambda p(x) + (1-\lambda) q(x)) \\
&\geq -\sum_{x \in \mathcal{X}} \lambda h(p(x)) + (1-\lambda) h(q(x)) \\
&= \lambda H(p) + (1-\lambda) H(q),
\end{aligned}
$$

which shows that $H(p)$ is a concave function.

[1]The same property holds for all random variables, but we will keep it discrete here for simplicity.

(b) *Consider the following optimization problem, which maximizes entropy subject to moment constraints on certain functions:*

$$\begin{aligned} \max_p \quad & H(p) \\ s.t. \quad & E_p[\phi_k(X)] = a_k \quad \textit{for all } k \end{aligned}$$

where $E_p[\cdot]$ is the expectation when X has pmf p. Why is this a convex program ?

Solution The above problem can be rewritten as

$$\begin{aligned} \min_p \quad & -H(p) \\ s.t. \quad & E_p[\phi_k(X)] = \sum_{x \in \mathcal{X}} \phi_k(x) p(x) = a_k \quad \text{for all } k \end{aligned}$$

As we have shown that $H(p)$ is a concave function, this problem is to minimize a convex function of p where the feasible set is convex because it is the intersection of affine sets. This shows that the optimization problem is a convex program.

(c) *Given a set of functions $\{\phi_k(x)\}$, show that the optimum of the convex program above is a pmf in the corresponding exponential family (6).*

Solution The Lagrangian function is given by

$$L(p, \lambda) = -H(p) + \sum_k \lambda_k \left(\sum_{x \in \mathcal{X}} \phi_k(x) p(x) - a_k \right)$$

where $\lambda_k \in \mathbb{R}$ for every k. Since the problem is a convex program, the optimal distribution satisfies the first-order optimality condition, given by

$$\begin{aligned} \frac{\partial L(p, \lambda)}{\partial p(x)} &= -(p(x) \log p(x))' + \sum_k \lambda_k \phi_k(x) \\ &= -(\log p(x) + 1) + \sum_k \lambda_k \phi_k(x) = 0 \end{aligned}$$

for some $\lambda_k \in \mathbb{R}$. It follows that $p(x)$ is in the form

$$p(x) = \frac{1}{Z(\lambda)} \exp\left(\sum_k \lambda_k \phi_k(x) \right),$$

which is a pmf in the exponential family (6). Here $Z(\lambda)$ is a normalization constant.

4. **Fast-Mixing Markov Chains** *A doubly-stochastic matrix P is a symmetric matrix with non-negative entries such that every row and every column sums up to 1. Its leading eigenvalue is always 1, corresponding to the eigenvector $\mathbf{1}$. Consider the absolute values of all the other eigenvalues of P, say $\lambda_2(P) \geq \ldots \geq \lambda_n(P)$, and let $\mu(P) := \max_{i \neq 1} |\lambda_i(P)|$ denote the largest such absolute value*[2].

[2] $1 - \mu(P)$ governs the time it takes for a Markov chain, with probability transition matrix P, to converge to the unique stationary distribution $\frac{1}{n}\mathbf{1}$.

(a) *Show that $\mu(P)$ is a convex function of P. (Hint: $\mu(P) = \max\{\lambda_2(P), -\lambda_n(P)\}$).*

Solution We have for $\gamma \in [0, 1]$

$$\begin{aligned}
\gamma\mu(P_1) + (1-\gamma)\mu(P_2) &= \gamma \max\{\lambda_2(P_1), -\lambda_n(P_1)\} + (1-\gamma)\max\{\lambda_2(P_2), -\lambda_n(P_2)\} \\
&\geq \max\{\gamma\lambda_2(P_1) + (1-\gamma)\lambda_2(P_2), -\gamma\lambda_n(P_1) - (1-\gamma)\lambda_n(P_2)\} \\
&\geq \max\{\lambda_2(\gamma P_1 + (1-\gamma)P_2), -\lambda_n(\gamma P_1 + (1-\gamma)P_2)\} \\
&= \mu(\gamma P_1 + (1-\gamma)P_2)
\end{aligned}$$

where the first inequality follows from that the max function is convex, and the second inequality follows from that $\lambda_2(\cdot)$ can be written as a convex function

$$\lambda_2(P) = \max_{x:\|x\|=1, \mathbf{1}^\top x = 0} x^\top P x,$$

and that $\lambda_n(\cdot)$ is given by a concave function

$$\lambda_n(P) = \min_{x:\|x\|=1} x^\top P x.$$

(b) *Write $\mu(P)$ as the spectral norm of P minus another matrix. (Recall: for symmetric matrices, spectral norm is the largest absolute value of an eignevalue.)* (Hint: all eigenvectors are orthogonal.)

Solution Let $d_1 = \mathbf{1}, d_2, \ldots, d_n$ denote the eigenvectors of P, and let $\lambda_1 = 1, \lambda_2, \ldots, \lambda_n$ denote the eigenvalues of P. Since all eigenvectors are orthogonal, we have

$$\begin{aligned}
P &= U\Lambda U^\top \\
&= \left[\begin{array}{c|c|c|c} \frac{1}{\sqrt{n}}\mathbf{1} & d_2 & \cdots & d_n \end{array}\right] \cdot \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \cdot \left[\begin{array}{c} (1/\sqrt{n})\mathbf{1}^\top \\ \hline d_2^\top \\ \hline \vdots \\ \hline d_n^\top \end{array}\right] \\
&= \left[\begin{array}{c|c|c|c} \frac{1}{\sqrt{n}}\mathbf{1} & \operatorname{sign}(\lambda_2)d_2 & \cdots & \operatorname{sign}(\lambda_n)d_n \end{array}\right] \cdot \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & |\lambda_2| & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & |\lambda_n| \end{bmatrix} \cdot \left[\begin{array}{c} (1/\sqrt{n})\mathbf{1}^\top \\ \hline d_1^\top \\ \hline \vdots \\ \hline d_n^\top \end{array}\right].
\end{aligned} \tag{7}$$

We can see that the first and third matrices in (7) are unitary, so (7) is a singular value decomposition of P. Therefore, we can write

$$\mu(P) = \max_{i \neq 1} |\lambda_i(P)| = \left\| P - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \right\|.$$

5. **Duality in graph theory** *Given a graph with edge weights $w_{ij} \geq 0$, the max-weight matching problem is: find the heaviest set of disjoint edges (i.e no two edges in the set share a node).*

The min-weight vertex cover problem is: put weights u_i on each vertex, so that (a) *for every edge we have $w_{ij} \leq u_i + u_j$, and* (b) *the total node weights $\sum_i u_i$ is minimized. Show that these two problems are the duals of each other.*

Solution Let $A_{ij}^{(k)}$ be such that

$$A_{ij}^{(k)} = \begin{cases} 1, & \text{if } k = i \text{ or } k = j \\ 0, & \text{otherwise} \end{cases}$$

The max-weight matching problem can be written as an integer program

$$\begin{aligned} \max_{x_{ij}} \quad & \sum_{i,j} w_{ij} x_{ij} \\ \text{s.t.} \quad & \sum_{i,j} A_{ij}^{(k)} x_{ij} \leq 1, \ \forall k, \\ & x_{ij} \in \{0, 1\} \quad \forall i, j. \end{aligned} \tag{P1}$$

Relaxing (P1) to an LP, we get

$$\begin{aligned} \max_{x_{ij}} \quad & \sum_{i,j} w_{ij} x_{ij} \\ \text{s.t.} \quad & \sum_{i,j} A_{ij}^{(k)} x_{ij} \leq 1, \ \forall k, \\ & 0 \leq x_{ij} \leq 1, \ \forall i, j. \end{aligned} \tag{P2}$$

The dual problem of (P2) is given by

$$\begin{aligned} \min_{u_k} \quad & \sum_{k} u_k \\ \text{s.t.} \quad & \sum_{k} A_{ij}^{(k)} u_k \geq w_{ij}, \forall i, j \\ & u_k \geq 0, \forall k. \end{aligned} \tag{D}$$

This is exactly the min-weight vertex cover problem. Since the dual of the dual is the primal itself for linear programs, the dual problem of the min-weight vertex cover problem (D) is the relaxed max-weight matching problem (P2).

6. **Robust Optimization**. *Recall the Robust Optimization framework we introduced in class. We have a linear program,*

$$\begin{aligned} \min: \quad & c^\top x \\ \text{s.t.}: \quad & a_i^\top x \leq b, \quad \forall a_i \in \mathcal{U}_i, \quad i = 1, \ldots, m, \end{aligned}$$

where $\mathcal{U}_i$ represents the uncertainty set. In class we considered polyhedral and ellipsoidal uncertainty sets. Now consider the following cardinality-constrained robust problem:[3] *Each*

[3] As motivation, consider the following: if you are modeling measurements, it may make sense to assume that all entries may be off by some amount. But if you are modeling, say, faulty components, where something either fails or does not, it may make more sense to consider the case where at most some finite number of components fail, and the others operate perfectly.

constraint, $a_i^\top x \le b_i$, has some integer r_i of its entries that may deviate from some nominal value, while the remaining $(n - r_i)$ entries are known exactly. Thus we have:

$$\mathcal{U}_i = \{a = a_i^0 + \hat{a}_i \,:\, |\hat{a}_{ij}| \le \Delta_{ij},\ |\mathrm{supp}(\hat{a}_i)| \le r_i\}.$$

That is, $\hat{a}_i$ is non-zero on at most r_i entries. Here, $\hat{a}_{ij}$ is the j^{th} entry of the vector $\hat{a}_i$.

Show that the robust linear program can be rewritten as a linear program. Note that you have a non-convex problem to deal with, because of the cardinality constraint. The final outcome, however, is just a linear program.

Solution Let n denote the length of x. x is a feasible solution if for every $i \in \{1, \ldots, m\}$

$$b - a_i^{0\top} x \ge \begin{pmatrix} \max_{\hat{a}_i \in \mathbb{R}^n} & \hat{a}_i^\top x \\ \text{s.t.} & |\hat{a}_{ij}| \le \Delta_{ij}, \\ & |\mathrm{supp}(\hat{a}_i)| \le r_i \end{pmatrix}.$$

The maximum is obtained by setting $\hat{a}_{ij} = \mathrm{sign}(x_j)\Delta_{ij}$ for the indices j corresponding to the r_i largest values of $\{|x_1|, \ldots, |x_n|\}$ and $\hat{a}_{ij} = 0$ for the rest. The value is given by

$$\sum_{j \in \mathcal{M}} \Delta_{ij} |x_j|$$

where $\mathcal{M}$ is the set of the indices of the r_i largest values of $\{|x_1|, \ldots, |x_n|\}$. This is equal to the maximum of the following linear program

$$\begin{aligned} \max_{d \in \mathbb{R}^n} \quad & \sum_{j=1}^{n} \Delta_{ij} |x_j| d_j \\ \text{s.t.} \quad & \sum_{j=1}^{m} d_j = r_i,\ 0 \le d_j \le 1 \end{aligned}$$

whose dual problem is given by

$$\begin{aligned} \min_{\lambda \in \mathbb{R}^n, \mu \in \mathbb{R}} \quad & \sum_{j=1}^{m} \lambda_j + \mu r_i \\ \text{s.t.} \quad & \lambda_j \ge \max\{\Delta_{ij}|x_j| - \mu, 0\},\ \forall j \in \{1, \ldots, m\}. \end{aligned}$$

Putting the dual problems for all $i = 1, \ldots, m$ together, we rewrite the original problem as

$$\begin{aligned} \min_{x \in \mathbb{R}^n, \lambda \in \mathbb{R}^{m \times n}, \mu \in \mathbb{R}^n} \quad & c^\top x \\ \text{s.t.} \quad & a_i^{0\top} x + \sum_{j=1}^{m} \lambda_{ij} + \mu_i r_i \le b,\ \forall i \in \{1, \ldots, m\}, \\ & \lambda_{ij} \ge \max\{\Delta_{ij}|x_j| - \mu_i, 0\},\ \forall i \in \{1, \ldots, m\},\ \forall j \in \{1, \ldots, n\}. \end{aligned}$$

Note that the satisfaction of the last constraint is equivalent to simultaneous satisfaction of the following linear constraints:

$$\lambda_{ij} \ge 0, \lambda_{ij} + \mu_i \ge \Delta_{ij} x_j, \lambda_{ij} + \mu_i \ge -\Delta_{ij} x_j.$$

This gives an LP.