Solution Manual

Optimization in Data Science and Machine Learning

Li Chen

chen_l@u.nus.edu

September 11, 2022

This is the solution manual for the assignments of Prof. Yinyu Ye's optimization course in NUS graduate summer school. The solution manual is revised from the course at Stanford CME307/MS&E311 Optimization in Data Science and Machine Learning. If you find any errors or typos in the solution, please let us know through the e-mail.

Notations We use $[n]$ to denote the running index $\{1, 2, \ldots, n\}$ for n a known integer.

Contents

1 Assignment 0

1. Consider the iterative process

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right),$$

where $a > 0$. Assuming the process converges, to what does it converge?

Solution Taking the limit, we have

$$x^* = \frac{1}{2}(x^* + \frac{a}{x^*})$$

Solve this equation, we have $x^* = \pm\sqrt{a}$. It's obvious that the iterations don't change the signs of x_k, so we have 1) if $x_0 > 0$, then $x_k \to \sqrt{a}$; 2) if $x_0 < 0$, then $x_k \to -\sqrt{a}$.

2. Let $\{(\boldsymbol{a}_i, c_i)\}_{i=1}^m$ be a given dataset where $\boldsymbol{a}_i \in \mathbb{R}^n$, $c_i \in \{\pm 1\}$.

(a) Compute the gradient of the following log-logistic-loss function,

$$f(\boldsymbol{x}, x_0) = \sum_{i:c_i=1} \log\left(1 + \exp(-\boldsymbol{a}_i^\top \boldsymbol{x} - x_0)\right) + \sum_{i:c_i=-1} \log\left(1 + \exp(\boldsymbol{a}_i^\top \boldsymbol{x} + x_0)\right),$$

where $\boldsymbol{x} \in \mathbb{R}^n$ and $x_0 \in \mathbb{R}$.

(b) Consider the following data set

$$\boldsymbol{a}_1 = (0; 0), \quad \boldsymbol{a}_2 = (1; 0), \quad \boldsymbol{a}_3 = (0; 1), \quad \boldsymbol{a}_4 = (0; 0), \quad \boldsymbol{a}_5 = (-1; 0), \quad \boldsymbol{a}_6 = (0; -1),$$

with label

$$c_1 = c_2 = c_3 = 1, \quad c_4 = c_5 = c_6 = -1,$$

show that there is no solution for $\nabla f(\boldsymbol{x}, x_0) = \boldsymbol{0}$.

Solution

(a) (Here we treat the gradient vector as a row vector.) For $c_i = 1$,

$$\nabla \log\left(1 + \exp(-\boldsymbol{a}_i^\top \boldsymbol{x} - x_0)\right) = \frac{\exp(-\boldsymbol{a}_i^\top \boldsymbol{x} - x_0)}{1 + \exp(-\boldsymbol{a}_i^\top \boldsymbol{x} - x_0)}(-\boldsymbol{a}_i^\top, \ -1);$$

and for $c_i = -1$,

$$\nabla \log\left(1 + \exp(\boldsymbol{a}_i^\top \boldsymbol{x} + x_0)\right) = \frac{\exp(\boldsymbol{a}_i^\top \boldsymbol{x} + x_0)}{1 + \exp(\boldsymbol{a}_i^\top \boldsymbol{x} + x_0)}(\boldsymbol{a}_i^\top, \ 1).$$

Thus, the gradient vector $\nabla f(\boldsymbol{x}, x_0)$ is

$$\sum_{i,c_i=1} \frac{\exp(-\boldsymbol{a}_i^\top \boldsymbol{x} - x_0)}{1 + \exp(-\boldsymbol{a}_i^\top \boldsymbol{x} - x_0)}(-\boldsymbol{a}_i^\top, \ -1) + \sum_{i,c_i=-1} \frac{\exp(\boldsymbol{a}_i^\top \boldsymbol{x} + x_0)}{1 + \exp(\boldsymbol{a}_i^\top \boldsymbol{x} + x_0)}(\boldsymbol{a}_i^\top, \ 1)$$

(b) We show by contradiction that a (finite) solution does not exist (Recall that in optimization most proofs are done by contradiction!). First, note that the objective is non-negative, and hence 0 is a lower bound. Then, we observe that taking $\boldsymbol{x} = (t, t)^\top$, $x_0 = 0$ and letting $t \to \infty$ leads to $f(x; x_0) \to 0$. Hence 0 is the infimum of the objective function. Nevertheless, for any finite $\boldsymbol{x}$ and x_0, obviously the objective is strictly positive. Hence we conclude that the problem has no (finite) solution.

You can also prove it by straightforward calculation: substitute the problem data and make $\nabla f(\boldsymbol{x}, x_0) = \boldsymbol{0}$. The key observation is $\boldsymbol{a}_1 = -\boldsymbol{a}_4$, $\boldsymbol{a}_2 = -\boldsymbol{a}_5$, $\boldsymbol{a}_3 = -\boldsymbol{a}_6$ so that $\nabla f(\boldsymbol{x}, x_0) = \boldsymbol{0}$ is equivalent to

$$\sum_{i\in[3]} \frac{\exp(-\boldsymbol{a}_i^\top \boldsymbol{x} - x_0)}{1 + \exp(-\boldsymbol{a}_i^\top \boldsymbol{x} - x_0)}(-\boldsymbol{a}_i^\top,\ -1) = \sum_{i\in[3]} \frac{\exp(-\boldsymbol{a}_i^\top \boldsymbol{x} + x_0)}{1 + \exp(-\boldsymbol{a}_i^\top \boldsymbol{x} + x_0)}(\boldsymbol{a}_i^\top,\ -1).$$

By checking the last coordinate, we must have $x_0 = 0$. Given $x_0 = 0$, the first two coordinates of the above equations is equivalent to

$$\boldsymbol{0} = 2\sum_{i\in[3]} \frac{\exp(-\boldsymbol{a}_i^\top \boldsymbol{x})}{1 + \exp(-\boldsymbol{a}_i^\top \boldsymbol{x})}\boldsymbol{a}_i,$$

which is never true as the RHS is strictly positive.

3. Given a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ s.t. $\boldsymbol{A}$ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, show that for every $k = 1, 2, \cdots, n$, we have:

$$\lambda_k = \max_U \left\{ \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U, \boldsymbol{x} \neq \boldsymbol{0} \right\} \middle| U \text{ is a linear subspace of } \mathbb{R}^n \text{ of dimension } k \right\} \tag{1}$$

$$= \min_U \left\{ \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U, \boldsymbol{x} \neq \boldsymbol{0} \right\} \middle| U \text{ is a linear subspace of } \mathbb{R}^n \text{ of dimension } n-k+1 \right\} \tag{2}$$

Solution This result is known as the *Courant-Fischer Minimax Theorem*. See Theorem 8.1.2 of [GVL13] for a sample proof and Wikipedia for the related topics.

We prove Equation (1) by induction on k. Equation (2) can be proved similarly. Let $\{\boldsymbol{v}_k\}_{k=1}^n$ denote a set of orthonormal eigenbasis of $\boldsymbol{A}$, with $\boldsymbol{A}\boldsymbol{v}_k = \lambda_k \boldsymbol{v}_k$. Moreover, $\boldsymbol{A} = \sum_{k=1}^n \lambda_k \boldsymbol{v}_k \boldsymbol{v}_k^\top$. Then $\frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\sum_{k\in[n]} \lambda_k x_k^2}{\sum_{k\in[n]} x_k^2}$ if $\boldsymbol{x} = \sum_{k\in[n]} x_k \boldsymbol{v}_k \neq \boldsymbol{0}$.

When $k = 1$, the expression reduces to $\lambda_1 = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}$, which is true for symmetric matrices, with one maximizer U^1 being spanned by $\boldsymbol{v}_1$.

Now suppose we have shown Equation (1) for some k and that the maximizer U^k can be taken to be the span of the first k eigenvectors, and we need to show a maximizer for λ_{k+1} is $U^{k+1} := U^k \cup \text{Span}\{\boldsymbol{v}_{k+1}\}$. To this end, note that

$$\lambda_{k+1} = \min_{x \in U^{k+1}} \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}$$

so that $\lambda_{k+1} \leq RHS$ of (1). On the other hand, for any subspace U of dimension $k+1$ that is not spanned by the first $k+1$ eigenvectors of $\boldsymbol{A}$, minimization in RHS will choose an eigenvector corresponding to an eigenvalue smaller than λ_{k+1}.

4. Given symmetric matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n \times n}$ such that $\boldsymbol{A}$ has eigenvalues $a_1 \geq a_2 \geq \cdots \geq a_n$, $\boldsymbol{B}$ has eigenvalues $b_1 \geq b_2 \geq \cdots \geq b_n$ and $\boldsymbol{C}$ has eigenvalues $c_1 \geq c_2 \geq \cdots \geq c_n$, if $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{C}$, show that for every $k = 1, 2, \cdots, n$, we have:

$$b_k + c_n \leq a_k \leq b_k + c_1. \tag{3}$$

Solution We show that $a_k \leq b_k + c_1$. According to (1), define U_k to be the dim-k linear subspace such that

$$a_k = \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \boldsymbol{0} \right\}$$

and let $\boldsymbol{x}^*$ be the minimizer of $\min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \boldsymbol{0} \right\}$. It follows that

$$\begin{aligned}
a_k &= \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top (\boldsymbol{B}+\boldsymbol{C}) \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \boldsymbol{0} \right\} \leq \frac{\boldsymbol{x}^{*\top} (\boldsymbol{B}+\boldsymbol{C}) \boldsymbol{x}^*}{\boldsymbol{x}^{*\top} \boldsymbol{x}^*} \\
&\leq \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \boldsymbol{0} \right\} + \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{C}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \boldsymbol{0} \right\} \\
&\leq \max_{U} \left\{ \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U, \boldsymbol{x} \neq \boldsymbol{0} \right\} | \dim(U) = k \right\} + \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{C}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \boldsymbol{0} \right\} \\
&= b_k + c_1.
\end{aligned}$$

Similarly, According to (2), define U_k to be the $(n-k+1)$-dimensional linear subspace such that

$$a_k = \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \boldsymbol{0} \right\}$$

and let $\boldsymbol{x}^*$ be the maximizer of $\max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \boldsymbol{0} \right\}$. It follows that

$$
\begin{aligned}
a_k &= \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top (\boldsymbol{B}+\boldsymbol{C})\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \boldsymbol{0} \right\} \geq \frac{\boldsymbol{x}^{*\top} (\boldsymbol{B}+\boldsymbol{C})\boldsymbol{x}^*}{\boldsymbol{x}^{*\top} \boldsymbol{x}^*} \\
&\geq \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \boldsymbol{0} \right\} + \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{C}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \boldsymbol{0} \right\} \\
&\geq \min_{U} \left\{ \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in U, \boldsymbol{x} \neq \boldsymbol{0} \right\} \middle| \dim(U) = n-k+1 \right\} + \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^\top \boldsymbol{C}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \middle| \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \boldsymbol{0} \right\} \\
&= b_k + c_n.
\end{aligned}
$$

5. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a positive-semidefinite matrix with Schur decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, where $\boldsymbol{Q} = [\boldsymbol{Q}_1 | \cdots | \boldsymbol{Q}_n]$ is an orthogonal matrix, $\boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$ satisfies $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n \geq 0$. Show that for any $k = 1, \ldots, n$,

$$
\min_{\operatorname{rank}(\boldsymbol{B})=k} \|\boldsymbol{A} - \boldsymbol{B}\|_2 = \|\boldsymbol{A} - \boldsymbol{A}_k\|_2 = \lambda_{k+1}, \tag{4}
$$

and

$$
\min_{\operatorname{rank}(\boldsymbol{B})=k} \|\boldsymbol{A} - \boldsymbol{B}\|_F = \|\boldsymbol{A} - \boldsymbol{A}_k\|_F = \sqrt{\sum_{j=k+1}^{n} \lambda_j^2}, \tag{5}
$$

where $\boldsymbol{A}_k$ is defined as

$$
\boldsymbol{A}_k := \sum_{j=1}^{k} \lambda_j \boldsymbol{Q}_j \boldsymbol{Q}_j^\top.
$$

Here $\|\cdot\|_2$ stands for the spectrum norm and $\|\cdot\|_F$ stands for the Frobenius norm.

Solution This result is a special case of the *Eckhart-Young Theorem*. See Theorem 2.4.8 of [GVL13] for a sample proof for the general case. This is also well-known as the best low-rank matrix approximation and the results generalize to general matrices (using SVD decomposition), see Wikipedia. Interestingly, both Problem (4) and (5) are non-convex problems (due to the rank constraints) but admit closed-form solutions.

We give a sketch of the special case here. We first show (4). Let $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ be any rank-k matrix. By rank-nullity theorem we can find orthonormal vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-k}$ that span the null space of $\boldsymbol{B}$. In the vector space $\mathbb{R}^n$, the null space of $\boldsymbol{B}$ which is $n-k$ dimensional, and the span of $\{\boldsymbol{q}_i\}_{i=1}^{k+1}$, which is $k+1$ dimensional, have non-zero intersection.

Let $\boldsymbol{z}$ be a unit norm vector in this intersection. We then have

$$\begin{aligned}\|\boldsymbol{A}-\boldsymbol{B}\|_2^2 &\geq \|(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{z}\|_2^2 = \|\boldsymbol{A}\boldsymbol{z}\|_2^2 \\ &= \sum_{i=1}^{k+1} \lambda_i^2 (\boldsymbol{q}_i^\top \boldsymbol{z})^2 \geq \lambda_{k+1}^2\end{aligned}$$

where in the last inequality we have used that $\sum_{i=1}^{k+1}(\boldsymbol{q}_i^\top \boldsymbol{z})^2 = \|\boldsymbol{z}\|^2 = 1$, since $\boldsymbol{z}$ is in the span of $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{k+1}$. We next prove (5), we use the identity that

$$\begin{aligned}\|\boldsymbol{C}\|_F^2 &= \mathrm{Tr}(\boldsymbol{C}^\top \boldsymbol{C}) \\ &= \mathrm{Tr}(\boldsymbol{C}^\top \boldsymbol{C} \sum_{j=1}^{n} \boldsymbol{v}_j \boldsymbol{v}_j^\top) \\ &= \sum_{j=1}^{n} (\boldsymbol{v}_j^\top \boldsymbol{C}^\top \boldsymbol{C} \boldsymbol{v}_j) = \sum_{j=1}^{n} \|\boldsymbol{C}\boldsymbol{v}_j\|^2\end{aligned}$$

for any orthonormal basis $\{\boldsymbol{v}_j\}_{j=1}^n$ and write

$$\begin{aligned}\|\boldsymbol{A}-\boldsymbol{B}\|_F^2 &= \sum_{j=1}^{n} \|(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{x}_j\|^2 \\ &= \sum_{j=1}^{n-k} \|\boldsymbol{A}\boldsymbol{x}_j\|^2 + \sum_{j=n-k+1}^{n} \|(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{x}_j\|^2 \\ &\geq \sum_{j=1}^{n-k} \|\boldsymbol{A}\boldsymbol{x}_j\|^2\end{aligned}$$

where we assume again $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-k}$ span the null space of $\boldsymbol{B}$.

Finally, we claim that

$$\sum_{j=1}^{n-k} \|\boldsymbol{A}\boldsymbol{x}_j\|^2 \geq \sum_{j=k+1}^{n} \|\boldsymbol{A}\boldsymbol{q}_j\|^2 = \sum_{j=k+1}^{n} \lambda_j^2,$$

which is implied by a fact that projections onto any $n-k$ dimensional subspace (LHS) is bounded below by the projection onto the $n-k$ dimensional subspace spanned by $\{\boldsymbol{q}_j\}_{j=k+1}^n$ (RHS). Equivalently, $\{\boldsymbol{q}_j\}_{j=1}^k$ span the best fit k-dimensional subspace for $\boldsymbol{A}$, in the sense that

$$\sum_{j=1}^{k} \|\boldsymbol{A}\boldsymbol{x}_j\|^2 \leq \sum_{j=1}^{k} \|\boldsymbol{A}\boldsymbol{q}_j\|^2$$

for any orthonormal system $\{\boldsymbol{x}_j\}_{j=1}^k$.

To prove

$$\sum_{j=1}^{k} \|\boldsymbol{A}\boldsymbol{x}_j\|^2 \leq \sum_{j=1}^{k} \|\boldsymbol{A}\boldsymbol{q}_j\|^2, \tag{6}$$

we use the important fact that

$$\boldsymbol{q}_j \in \arg \max_{\boldsymbol{v} \perp \mathrm{Span}\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{j-1}\}} \|\boldsymbol{A}\boldsymbol{v}\|^2,$$

that is the j-th unit eigenvector of $\boldsymbol{A}$ maximizes $\|\boldsymbol{A}\boldsymbol{v}\|^2$ among all unit vectors that are not in the span of the first $j-1$ eigenvectors. Clearly the inequality (6) holds for $k = 1$. Suppose for the sake of induction we have shown it for some k. Let $\{\boldsymbol{y}_j\}_{j=1}^{k+1}$ be a solution to

$$\max_{\text{orthonormal } \{\boldsymbol{x}_j\}} \sum_{j=1}^{k+1} \|\boldsymbol{A}\boldsymbol{x}_j\|^2$$

W.l.o.g, we can let $\boldsymbol{y}_{k+1}$ be orthogonal to the span of $\{\boldsymbol{q}_j\}_{j=1}^{k}$. Then $\|\boldsymbol{A}\boldsymbol{y}_{k+1}\|^2 \leq \|\boldsymbol{A}\boldsymbol{q}_{k+1}\|^2$, so that

$$\sum_{j=1}^{k+1} \|\boldsymbol{A}\boldsymbol{y}_j\|^2 \leq \sum_{j=1}^{k+1} \|\boldsymbol{A}\boldsymbol{q}_j\|^2$$

completing the induction step.

2 Assignment 1

Reading. Read selected sections in [LY21] Chapters 1, 2, 6 and Appendices A, B.

1. (15') Show the followings:

   (a) (5') Consider the set

   $$F := \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}\},$$

   where data matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and vector $\boldsymbol{b} \in \mathbb{R}^m$. Prove that F is a convex set.

   (b) (5') Fix data matrix $\boldsymbol{A}$ and consider the $\boldsymbol{b}$-data set for F defined in part (a):

   $$B := \{\boldsymbol{b} \in \mathbb{R}^m :\ F \text{ is not empty}\}.$$

   Prove that B is a convex set.

   (c) (5') Fix data matrix $\boldsymbol{A}$ and consider the linearly constrained convex minimization problem

   $$\begin{aligned} z(\boldsymbol{b}) := \max \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0} \end{aligned}$$

   where $f(\boldsymbol{x})$ is a concave function, and the maximal value function $z(\boldsymbol{b})$ is an implicit function of $\boldsymbol{b}$. Prove that $z(\boldsymbol{b})$ is a concave function of $\boldsymbol{b} \in B$, where B is defined in part (b).

Solution Typical strategies of proving the convexity of the functions or sets:

(i) Prove by definition;

(ii) Prove by convex calculus: convexity preserved operations for sets and functions;

(iii) Prove convex functions by first-order or second-order conditions;

(iv) Be aware of the relation of convex sets and convex functions via the epigraph.

(a) Take any two points $\boldsymbol{x}', \boldsymbol{x}'' \in F$, that is, $\boldsymbol{A}\boldsymbol{x}' = \boldsymbol{b}, \boldsymbol{x}' \geq \boldsymbol{0}$ and $\boldsymbol{A}\boldsymbol{x}'' = \boldsymbol{b}, \boldsymbol{x}'' \geq \boldsymbol{0}$. Then, for any $\alpha \in [0, 1]$ we have

$$\alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}'' \geq 0.$$

and

$$\boldsymbol{A}(\alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}'') = \alpha\boldsymbol{A}\boldsymbol{x}' + (1-\alpha)\boldsymbol{A}\boldsymbol{x}'' = \alpha\boldsymbol{b} + (1-\alpha)\boldsymbol{b} = \boldsymbol{b}.$$

Thus, $\alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}'' \in F$.

(b) Take any two points $\boldsymbol{b}', \boldsymbol{b}'' \in B$. Then we must have $\boldsymbol{x}' \geq \mathbf{0}$ and $\boldsymbol{x}'' \geq \mathbf{0}$ such that $\boldsymbol{A}\boldsymbol{x}' = \boldsymbol{b}'$ and $\boldsymbol{A}\boldsymbol{x}'' = \boldsymbol{b}''$. Now we prove that the convex combination $\alpha\boldsymbol{b}' + (1-\alpha)\boldsymbol{b}''$ is also in B for any $\alpha \in [0,1]$. Consider the convex combination $\boldsymbol{x} = \alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}''$. Obviously, $\boldsymbol{x} \geq \mathbf{0}$. Furthermore,

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}(\alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}'') = \alpha\boldsymbol{A}\boldsymbol{x}' + (1-\alpha)\boldsymbol{A}\boldsymbol{x}'' = \alpha\boldsymbol{b}' + (1-\alpha)\boldsymbol{b}'',$$

which give the desired proof.

(c) Take any two points $\boldsymbol{b}', \boldsymbol{b}'' \in B$, and let $\boldsymbol{x}'$ and $\boldsymbol{x}''$ be two minimizers for $\boldsymbol{b} = \boldsymbol{b}'$ and $\boldsymbol{b} = \boldsymbol{b}''$, respectively. That is, $z(\boldsymbol{b}') = f(\boldsymbol{x}')$ and $z(\boldsymbol{b}'') = f(\boldsymbol{x}'')$. Then, consider $z(\alpha\boldsymbol{b}' + (1-\alpha)\boldsymbol{b}'')$. Since $\alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}'' \geq \mathbf{0}$ and $\boldsymbol{A}(\alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}'') = \alpha\boldsymbol{b}' + (1-\alpha)\boldsymbol{b}''$, $\alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}''$ is a feasible solution for problem with $\boldsymbol{b} = \alpha\boldsymbol{b}' + (1-\alpha)\boldsymbol{b}''$. Thus, the maximum value must be greater or equal to the objective value at a feasible solution, *i.e.*,

$$z(\alpha\boldsymbol{b}' + (1-\alpha)\boldsymbol{b}'') \geq f(\alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}'') \geq \alpha f(\boldsymbol{x}') + (1-\alpha) f(\boldsymbol{x}'') = \alpha z(\boldsymbol{b}') + (1-\alpha) z(\boldsymbol{b}''),$$

where the second inequality is from f is a concave function.

2. (10') Show that the dual cone of the n-dimensional nonnegative orthant cone $\mathbb{R}^n_+$ is itself, that is,

$$(\mathbb{R}^n_+)^* = \mathbb{R}^n_+.$$

(Hint: show that $\mathbb{R}^n_+ \subset (\mathbb{R}^n_+)^*$ and $(\mathbb{R}^n_+)^* \subset \mathbb{R}^n_+$.)

Solution Prove $\mathbb{R}^n_+ \subset (\mathbb{R}^n_+)^*$: Let any $\boldsymbol{y} \in \mathbb{R}^n_+$. Then $\boldsymbol{x}^\top \boldsymbol{y} = \sum_i x_i y_i \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^n_+$ since each of the product in the sum is nonnegative.

Prove $(\mathbb{R}^n_+)^* \subset \mathbb{R}^n_+$: Suppose this is not true, that is, there exists $\boldsymbol{y} \in (\mathbb{R}^n_+)^*$ but $\boldsymbol{y} \notin \mathbb{R}^n_+$. Then at least one entry of $\boldsymbol{y}$ is negative, w.l.o.g., say $y_1 < 0$. Now we select $\boldsymbol{e}_1 = (1; 0; ...; 0) \in \mathbb{R}^n_+$ so that

$$\boldsymbol{e}_1^\top \boldsymbol{y} = y_1 < 0$$

which contradicts that $\boldsymbol{y} \in (\mathbb{R}^n_+)^*$.

Remark: We have shown $\mathbb{R}^n_+$ is self-dual. One can also show the positive semidefinite cone $\mathbb{S}^n_+$ is self-dual easily (by showing that $\mathbb{S}^n_+ \subseteq (\mathbb{S}^n_+)^*$ and $(\mathbb{S}^n_+)^* \subseteq \mathbb{S}^n_+$).

3. (10') Let $g_1, \ldots, g_m$ be a collection of concave functions on $\mathbb{R}^n$ such that

$$S = \{\boldsymbol{x} : g_i(\boldsymbol{x}) > 0 \text{ for } i = 1, \ldots, m\} \neq \emptyset.$$

Show that for any positive constant μ and any convex function f on $\mathbb{R}^n$, the function (called Barrier function)

$$h(\boldsymbol{x}) = f(\boldsymbol{x}) - \mu \sum_{i=1}^{m} \log(g_i(\boldsymbol{x}))$$

is convex over S. (Hint: directly apply the convex/concave function definition or analyze the Hessian of $h(\boldsymbol{x})$.)

Solution It is easy to verify that S is convex. We know that the positively weighted sum of convex functions having a common domain is convex on that domain. The given conditions imply that the function

$$h(\boldsymbol{x}) = f(\boldsymbol{x}) - \mu \sum_{i=1}^{m} \log(g_i(\boldsymbol{x}))$$

is a positively weighted sum of the convex functions. To see this, one can prove that a nondecreasing concave function of a concave function is concave, then it follows that $\log g_i(\boldsymbol{x})$ is a concave function.

To prove it, take any two points $\boldsymbol{x}'$ and $\boldsymbol{x}''$ in S, then for each i

$$g_i(\alpha \boldsymbol{x}' + (1-\alpha)\boldsymbol{x}'') \geq \alpha g_i(\boldsymbol{x}') + (1-\alpha) g_i(\boldsymbol{x}'').$$

Since log is nondecreasing,

$$\log(g_i(\alpha \boldsymbol{x}' + (1-\alpha)\boldsymbol{x}'')) \geq \log(\alpha g_i(\boldsymbol{x}') + (1-\alpha) g_i(\boldsymbol{x}'')).$$

Moreover, log is a concave function, so that

$$\log(g_i(\alpha \boldsymbol{x}' + (1-\alpha)\boldsymbol{x}'')) \geq \log(\alpha g_i(\boldsymbol{x}') + (1-\alpha) g_i(\boldsymbol{x}'')) \geq \alpha \log(g_i(\boldsymbol{x}')) + (1-\alpha) \log(g_i(\boldsymbol{x}'')),$$

which completes the proof.

Hence its negative $-\log g_i(\boldsymbol{x})$ is convex. Thus, we see that

$$h(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} [\mu(-\log g_i(\boldsymbol{x}))]$$

is convex on S.

4. (10') (Lipschitz Functions) Prove the following two implication inequalities:

   (a) (5') Assume f is a first-order β-Lipschitz function, namely there is a positive number β such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$:

   $$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|,$$

   then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

   $$|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y})| \leq \frac{\beta}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

   (b) (5') Assume f is a second-order β-Lipschitz function, namely there is a positive number β such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$:

   $$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) - \nabla^2 f(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y})\| \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|^2,$$

   then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

   $$|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{y})^\top \nabla^2 f(\boldsymbol{y}) (\boldsymbol{x} - \boldsymbol{y})| \leq \frac{\beta}{3} \|\boldsymbol{x} - \boldsymbol{y}\|^3.$$

Solution The key tool is Taylor's formula with integral remainder, see the Wikipedia.

Let $\boldsymbol{d} := \boldsymbol{y} - \boldsymbol{x}$ and $\phi(t) = f(\boldsymbol{x} + t\boldsymbol{d})$ where t is a scalar variable. Then we have $\phi(0) = f(\boldsymbol{x})$ and $\phi(1) = f(\boldsymbol{x} + \boldsymbol{d}) = f(\boldsymbol{y})$. Moreover,

$$f(\boldsymbol{x} + \boldsymbol{d}) - f(\boldsymbol{x}) = \phi(1) - \phi(0) = \int_0^1 \mathrm{d}\phi(t) = \int_0^1 \boldsymbol{d}^\top \nabla f(\boldsymbol{x} + t\boldsymbol{d}) \mathrm{d}t.$$

For the first implication inequality, note that $\boldsymbol{d}^\top \nabla f(\boldsymbol{x}) = \int_0^1 \boldsymbol{d}^\top \nabla f(\boldsymbol{x}) \mathrm{d}t$, we have

$$\begin{aligned}
&|f(\boldsymbol{x} + \boldsymbol{d}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top \boldsymbol{d}| = \left| \int_0^1 \boldsymbol{d}^\top \left(\nabla f(\boldsymbol{x} + t\boldsymbol{d}) - \nabla f(\boldsymbol{x}) \right) \mathrm{d}t \right| \\
&\leq \int_0^1 \left| \boldsymbol{d}^\top \left(\nabla f(\boldsymbol{x} + t\boldsymbol{d}) - \nabla f(\boldsymbol{x}) \right) \right| \mathrm{d}t \\
&\leq \int_0^1 \|\boldsymbol{d}\| \, \|\nabla f(\boldsymbol{x} + t\boldsymbol{d}) - \nabla f(\boldsymbol{x})\| \, \mathrm{d}t \quad \text{(Cauchy-Schwartz inequality)} \\
&= \|\boldsymbol{d}\| \int_0^1 \|\nabla f(\boldsymbol{x} + t\boldsymbol{d}) - \nabla f(\boldsymbol{x})\| \, \mathrm{d}t \\
&\leq \|\boldsymbol{d}\| \int_0^1 \beta \|t\boldsymbol{d}\| \mathrm{d}t \quad \text{(the first-order Lipschitz condition)} \\
&= \|\boldsymbol{d}\| \beta \|\boldsymbol{d}\| \int_0^1 t \mathrm{d}t = \frac{\beta}{2} \|\boldsymbol{d}\|^2.
\end{aligned}$$

For the second implication inequality, note that $\frac{1}{2}\boldsymbol{d}^\top\nabla^2 f(\boldsymbol{x})\boldsymbol{d} = \boldsymbol{d}^\top\nabla^2 f(\boldsymbol{x})\boldsymbol{d}\int_0^1 t\mathrm{d}t$, we have

$$
\begin{aligned}
&|f(\boldsymbol{x}+\boldsymbol{d}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top\boldsymbol{d} - \frac{1}{2}\boldsymbol{d}^\top\nabla^2 f(\boldsymbol{x})\boldsymbol{d}| \\
&= \left|\int_0^1 \boldsymbol{d}^\top\left(\nabla f(\boldsymbol{x}+t\boldsymbol{d}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})(t\boldsymbol{d})\right)\mathrm{d}t\right| \\
&\le \int_0^1 \left|\boldsymbol{d}^\top\left(\nabla f(\boldsymbol{x}+t\boldsymbol{d}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})(t\boldsymbol{d})\right)\right|\mathrm{d}t \\
&\le \int_0^1 \|\boldsymbol{d}\|\|\nabla f(\boldsymbol{x}+t\boldsymbol{d}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})(t\boldsymbol{d})\|\mathrm{d}t \quad \text{(Cauchy-Schwartz inequality)} \\
&= \|\boldsymbol{d}\|\int_0^1 \|\nabla f(\boldsymbol{x}+t\boldsymbol{d}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})(t\boldsymbol{d})\|\mathrm{d}t \\
&\le \|\boldsymbol{d}\|\int_0^1 \beta\|t\boldsymbol{d}\|^2\mathrm{d}t \quad \text{(the second-order Lipschitz condition)} \\
&= \beta\|\boldsymbol{d}\|^3\int_0^1 t^2\mathrm{d}t = \frac{\beta}{3}\|\boldsymbol{d}\|^3.
\end{aligned}
$$

5. (10') Consider the following SOCP problem:

$$
\begin{aligned}
\min \quad & 2x_1 + x_2 + x_3 \\
\text{s.t.} \quad & x_1 + x_2 + x_3 = 1, \\
& x_1 - \sqrt{x_2^2 + x_3^2} \ge 0.
\end{aligned}
$$

   (a) (5') Show that the feasible region is a convex set.

   (b) (5') Try to find a minimizer of the problem and "argue" [1] why it is a minimizer.

Solution

(a) It is clear that the plane set $\{\boldsymbol{x} : \boldsymbol{e}^\top\boldsymbol{x} = 1\}$ is a convex set. Let $\boldsymbol{x}_{-1} = (x_2; x_3; ...; x_n)$. Then we like to prove that

$$\{\boldsymbol{x} : \|\boldsymbol{x}_{-1}\| \le x_1\}$$

is a convex set. Consider any two points $\boldsymbol{x}'$ and $\boldsymbol{x}''$ in the set. For any $\alpha \in [0\ 1]$, we have, by triangle inequality,

$$\|\alpha\boldsymbol{x}'_{-1} + (1-\alpha)\boldsymbol{x}''_{-1}\| \le \|\alpha\boldsymbol{x}'_{-1}\| + \|(1-\alpha)\boldsymbol{x}''_{-1}\| = \alpha\|\boldsymbol{x}'_{-1}\| + (1-\alpha)\|\boldsymbol{x}''_{-1}\|.$$

[1] We recommend to prove this directly, namely without using duality argument which will be introduced in the following lectures.

But $\|\boldsymbol{x}'_{-1}\| \le x'_1$ and $\|\boldsymbol{x}''_{-1}\| \le x''_1$, so that

$$\|\alpha\boldsymbol{x}'_{-1} + (1-\alpha)\boldsymbol{x}''_{-1}\| \le \alpha x'_1 + (1-\alpha)x''_1;$$

that is, the convex combination point is also in the set. This implies that the set is a convex set. The feasible region is the intersection of two convex sets, so it is also a convex set.

(b) The problem can be treated as

$$\begin{aligned} \min \quad & x_1 \\ \text{s.t.} \quad & x_1 + x_2 + x_3 = 1, \\ & x_1 - \sqrt{x_2^2 + x_3^2} \ge \mathbf{0}; \end{aligned}$$

which is as the same as

$$\begin{aligned} \max \quad & x_2 + x_3 \\ \text{s.t.} \quad & x_2 + x_3 + \sqrt{x_2^2 + x_3^2} \le 1. \end{aligned}$$

For any fixed positive value of $x_2 + x_3$, $\sqrt{x_2^2 + x_3^2}$ would be minimized when $x_2 = x_3$. Thus, we consider the case $x_2 = x_3$: which is as the same as

$$\begin{aligned} \max \quad & 2x_2 \\ \text{s.t.} \quad & 2x_2 + \sqrt{2}x_2 \le 1. \end{aligned}$$

That is, $x_2 = \frac{1}{2+\sqrt{2}}$. Thus, the minimal value of the original problem is $2 - 2x_2 = \sqrt{2}$.

6. (10') Prove that the set $\{\boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^n_+\}$ is a closed and convex cone. (Hint: apply Carathéodory's theorem in Lecture Note to prove the closedness.)

Solution Let $C = \{\boldsymbol{A}\boldsymbol{x} : \ \boldsymbol{x} \ge \mathbf{0} \in \mathbb{R}^n\}$.

It is easy to see that C is a cone. Take any $\boldsymbol{b} \in C$, then $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ for some $\boldsymbol{x} \ge \mathbf{0}$. Now consider $\beta\boldsymbol{b}$ for any $\beta \ge \mathbf{0}$. But $\beta\boldsymbol{b} = \beta(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}(\beta\boldsymbol{x})$ and $\beta\boldsymbol{x} \ge \mathbf{0}$, so that $\beta\boldsymbol{b} \in C$.

The convexity is easy to prove. Take $\boldsymbol{b}^1 \in C$ and $\boldsymbol{b}^2 \in C$. Then we must have $\boldsymbol{x}^1 \ge \mathbf{0}$ and $\boldsymbol{x}^2 \ge \mathbf{0}$ such that $\boldsymbol{b}^1 = \boldsymbol{A}\boldsymbol{x}^1$ and $\boldsymbol{b}^2 = \boldsymbol{A}\boldsymbol{x}^2$. Then for any $\alpha \in [0\ 1]$,

$$\alpha\boldsymbol{b}^1 + (1-\alpha)\boldsymbol{b}^2 = \alpha(\boldsymbol{A}\boldsymbol{x}^1) + (1-\alpha)(\boldsymbol{A}\boldsymbol{x}^2) = \boldsymbol{A}(\alpha\boldsymbol{x}^1 + (1-\alpha)\boldsymbol{x}^2).$$

since $\alpha\boldsymbol{x}^1 + (1-\alpha)\boldsymbol{x}^2 \ge \mathbf{0}$, we have $\alpha\boldsymbol{b}^1 + (1-\alpha)\boldsymbol{b}^2 \in C$.

Now we prove C is closed. Let $\{\boldsymbol{b}^k = \boldsymbol{A}\boldsymbol{x}^k : \boldsymbol{x}^k \geq \boldsymbol{0}\}_{k\in\{1,2,...\}}$ be a convergent sequence with the limit point $\bar{\boldsymbol{b}}$. We need to prove $\bar{\boldsymbol{b}} \in C$.

From Carathéodory's theorem, we can assume that $\boldsymbol{x}^k$ is a basic feasible solution, that is, for some basis $B^k \subseteq [n]$, we have $\boldsymbol{A}_{B^k}\boldsymbol{x}^k_{B^k} = \boldsymbol{b}^k$, and the rest of entries in $\boldsymbol{x}^k$ are all zeros. Then, $\boldsymbol{x}^k$ is bounded for all $k \in \{1, 2, ...\}$. Thus, there must be a convergent subsequence of $\{\boldsymbol{x}^k \geq \boldsymbol{0}\}_{k\in\mathcal{K}}$ with the limit point $\bar{\boldsymbol{x}}$. Then we have $\bar{\boldsymbol{x}} \geq \boldsymbol{0}$. Now consider the subsequence $\{\boldsymbol{b}^k = \boldsymbol{A}\boldsymbol{x}^k\}_{k\in\mathcal{K}}$, which is a convergent sequence with limit $\boldsymbol{A}\bar{\boldsymbol{x}} \in C$. But the limit of any convergent subsequence of $\boldsymbol{b}^k$ is $\bar{\boldsymbol{b}}$, so that $\bar{\boldsymbol{b}} = \boldsymbol{A}\bar{\boldsymbol{x}} \in C$.

7. (15') Farkas' lemma can be used to derive many other (named) theorems of the alternative. This problem concerns a few of these pairs of systems. Using Farkas's lemma, prove each of the following results.

   (a) (5') Gordan's Theorem. Exactly one of the following systems has a solution:

   $$\begin{aligned} &\text{(i)} \quad \boldsymbol{A}\boldsymbol{x} > \boldsymbol{0} \\ &\text{(ii)} \quad \boldsymbol{y}^\top\boldsymbol{A} = \boldsymbol{0}, \quad \boldsymbol{y} \geq \boldsymbol{0}, \quad \boldsymbol{y} \neq \boldsymbol{0}. \end{aligned}$$

   (b) (5') Stiemke's Theorem. Exactly one of the following systems has a solution:

   $$\begin{aligned} &\text{(i)} \quad \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{0}, \quad \boldsymbol{A}\boldsymbol{x} \neq \boldsymbol{0} \\ &\text{(ii)} \quad \boldsymbol{y}^\top\boldsymbol{A} = \boldsymbol{0}, \quad \boldsymbol{y} > \boldsymbol{0} \end{aligned}$$

   (c) (5') Gale's Theorem. Exactly one of the following systems has a solution:

   $$\begin{aligned} &\text{(i)} \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ &\text{(ii)} \quad \boldsymbol{y}^\top\boldsymbol{A} = \boldsymbol{0}, \quad \boldsymbol{y}^\top\boldsymbol{b} < 0, \quad \boldsymbol{y} \geq \boldsymbol{0} \end{aligned}$$

Solution

(a) Gordan's Theorem. Let $\boldsymbol{b}$ denote a positive vector. Then, (i) is equivalent to $\boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}$ and it can be written as

$$\boldsymbol{A}\boldsymbol{x}' - \boldsymbol{A}\boldsymbol{x}'' - \boldsymbol{z} = \boldsymbol{b}, \ (\boldsymbol{x}'; \boldsymbol{x}''; \boldsymbol{z}) \geq \boldsymbol{0}$$

By Farkas' lemma, it is alternative system is

$$\boldsymbol{y}^\top(\boldsymbol{A}, \ -\boldsymbol{A}, \ -\boldsymbol{I}) \leq \boldsymbol{0}, \boldsymbol{y}^\top\boldsymbol{b} = 1$$

which is equivalent to (ii).

(b) Stiemke's Theorem. Let b denote a positive vector. Then, (i) is equivalent to $\boldsymbol{Ax} \geq \mathbf{0},\ \boldsymbol{b}^\top \boldsymbol{Ax} = 1$ and it can be written as:

$$\begin{pmatrix} \boldsymbol{A} & -\boldsymbol{A} & -\boldsymbol{I} \\ \boldsymbol{b}^\top \boldsymbol{A} & -\boldsymbol{b}^\top \boldsymbol{A} & \mathbf{0} \end{pmatrix} (\boldsymbol{x}'; \boldsymbol{x}''; \boldsymbol{z}) = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix},\ (\boldsymbol{x}'; \boldsymbol{x}''; \boldsymbol{z}) \geq \mathbf{0}$$

By Farkas' lemma, its alternative system is on $(\boldsymbol{y}'; \tau)$ such that:

$$(\boldsymbol{y}'; \tau)^\top \begin{pmatrix} \boldsymbol{A} & -\boldsymbol{A} & -\boldsymbol{I} \\ \boldsymbol{b}^\top \boldsymbol{A} & -\boldsymbol{b}^\top \boldsymbol{A} & \mathbf{0} \end{pmatrix} \leq \mathbf{0},\ (\boldsymbol{y}'; \tau)^\top (\mathbf{0}; 1) = 1$$

Let $\boldsymbol{y} = \boldsymbol{y}' + \tau \cdot \boldsymbol{b}$. Then, it is a solution to (ii).

(c) Gale's Theorem. Note that (i) can be written as:

$$\boldsymbol{Ax}' - \boldsymbol{Ax}'' + \boldsymbol{z} = \boldsymbol{b},\ (\boldsymbol{x}'; \boldsymbol{x}''; \boldsymbol{z}) \geq \mathbf{0}$$

By Farkas' lemma, its alternative system is

$$\boldsymbol{y}^\top (\boldsymbol{A},\ -\boldsymbol{A},\ \boldsymbol{I}) \leq \mathbf{0},\ \boldsymbol{y}^\top \boldsymbol{b} = 1$$

Then, $-\boldsymbol{y}$ is a solution to (ii).

8. (20') Consider the sensor localization problem on plane $\mathbb{R}^2$ with one sensor $\boldsymbol{x}$ and three anchors $\boldsymbol{a}_1 = (1;0)$, $\boldsymbol{a}_2 = (-1;0)$ and $\boldsymbol{a}_3 = (0;2)$. Suppose the Euclidean distances from the sensor to the three anchors are d_1, d_2 and d_3 respectively and known to us.Then, from the anchor and distance information, we can locate the second by finding $\boldsymbol{x} \in \mathbb{R}^2$ such that

$$\|\boldsymbol{x} - \boldsymbol{a}_i\|^2 = d_i^2,\ i = 1, 2, 3.$$

Do the following numerical experiments using CVX (or cvxpy, convex.jl) or MOSEK and answer the questions:

(a) (10') Generate any sensor point **in the convex hull** of the three anchors, compute its distances to three anchors d_i, $i = 1, 2, 3$, respectively. Then solve the SOCP relaxation problem

$$\|\boldsymbol{x} - \boldsymbol{a}_i\|^2 \le d_i^2,\ i = 1, 2, 3.$$

Did you find the correct location? What about if the sensor point was in the **outside** of the convex hull? Try a few different locations of the sensor and identify the pattern.

(b) (10') Now try the SDP relaxation

$$(\boldsymbol{a}_i; -1)(\boldsymbol{a}_i; -1)^\top \bullet \begin{pmatrix} \boldsymbol{I} & \boldsymbol{x} \\ \boldsymbol{x}^\top & \boldsymbol{Y} \end{pmatrix} = d_i^2,\ i = 1, 2, 3;\ \begin{pmatrix} \boldsymbol{I} & \boldsymbol{x} \\ \boldsymbol{x}^\top & \boldsymbol{Y} \end{pmatrix} \succeq \boldsymbol{0} \in \mathbb{S}^3,$$

which can be written in the standard form

$$\begin{aligned} (1;0;0)(1;0;0)^\top \bullet \boldsymbol{Z} &= 1, \\ (0;1;0)(0;1;0)^\top \bullet \boldsymbol{Z} &= 1, \\ (1;1;0)(1;1;0)^\top \bullet \boldsymbol{Z} &= 2, \\ (\boldsymbol{a}_i; -1)(\boldsymbol{a}_i; -1)^\top \bullet \boldsymbol{Z} &= d_i^2,\ i = 1, 2, 3, \\ \boldsymbol{Z} &\succeq \boldsymbol{0} \in \mathbb{S}^3. \end{aligned}$$

Did you find the correct location everywhere on the plane? Try a few different locations of the sensor and identify the pattern.

You can use CVX (or cvxpy, convex.jl) to solve these numerical problems.

Solution

Both the SOCP and SDP relaxations exactly find the sensor location if the sensor is contained in the convex hull of the anchor points.

However, when the sensor is located outside of the convex hull, the SOCP relaxation will fail to find the sensor correctly. This is due to the relaxed $\|\boldsymbol{x}-\boldsymbol{a}_i\| \leq d_i$ constraint, which allows regions of the convex hull to be feasible even if $\boldsymbol{x}^*$ is outside of the convex hull. Thus the SOCP relaxation will tend to return solutions within the convex hull. On the other hand, the SDP relaxation is always exact since it strictly requires that $\|\boldsymbol{x}-\boldsymbol{a}_i\| = d_i$.

Experimental MATALB code is given below. The Julia code is online here.

```
%%  Homework 1 Problem 8

%% Each column is anchor point
A = [1 -1 0;
     0  0 2];

%% Generate sensor in convex hull of 3 anchors
%% SOCP relaxation
alpha = rand(3,1);
alpha = alpha/norm(alpha,1);
s_true = A*alpha;
d = norms(A - s_true*ones(1,3));

cvx_begin quiet
    variable s(2)
    minimize( 0 )
    subject to
        norms(A - s*ones(1,3)) <= d;
cvx_end

fprintf('SOCP - Inside of Convex Hull\n');
fprintf('True sensor location     : (%f, %f)\n', s_true(1), ...
    s_true(2));
fprintf('Recovered sensor location: (%f, %f)\n', s(1), s(2));
fprintf('Difference : %f\n\n', norm(s_true - s));

%% SDP relaxation
cvx_begin sdp quiet
    variable X(3,3) semidefinite
    minimize( 0 )
    subject to
        X(1:2,1:2) == eye(2)
```

```
        for i=1:3
            [A(:,i);-1]'*X*[A(:,i);-1] == d(i)^2
        end
cvx_end

s = X(1:2,3);

fprintf('SDP - Inside of Convex Hull\n');
fprintf('True sensor location     : (%f, %f)\n', s_true(1), ...
    s_true(2));
fprintf('Recovered sensor location: (%f, %f)\n', s(1), s(2));
fprintf('Difference : %f\n\n', norm(s_true - s));

%% Generate sensor outside of convex hull of 3 anchors
alpha = 10*rand(3,1);
s_true = A*alpha;
d = norms(A - s_true*ones(1,3));

cvx_begin quiet
    variable s(2)
    minimize( 0 )
    subject to
        norms(A - s*ones(1,3)) <= d;
cvx_end

fprintf('SOCP - Outside of Convex Hull\n');
fprintf('True sensor location     : (%f, %f)\n', s_true(1), ...
    s_true(2));
fprintf('Recovered sensor location: (%f, %f)\n', s(1), s(2));
fprintf('Difference : %f\n\n', norm(s_true - s));

%% SDP relaxation
cvx_begin sdp quiet
    variable X(3,3) semidefinite
    minimize( 0 )
    subject to
        X(1:2,1:2) == eye(2)
        for i=1:3
            [A(:,i);-1]'*X*[A(:,i);-1] == d(i)^2
        end
cvx_end
```

```
71
72  s = X(1:2,3);
73
74  fprintf('SDP - Outside of Convex Hull\n');
75  fprintf('True sensor location     : (%f, %f)\n', s_true(1), ...
        s_true(2));
76  fprintf('Recovered sensor location: (%f, %f)\n', s(1), s(2));
77  fprintf('Difference : %f\n\n', norm(s_true - s));
```

9. (20') Consider the sensor localization problem on plane $\mathbb{R}^2$ with **two** sensors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and three anchors $\boldsymbol{a}_1 = (1;0)$, $\boldsymbol{a}_2 = (-1;0)$ and $\boldsymbol{a}_3 = (0;2)$. Suppose that we know the (Euclidean) distances from one sensor $\boldsymbol{x}_1$ to $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, denoted by d_{11} and d_{12}; distances of the other sensor $\boldsymbol{x}_2$ to $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$, denoted by d_{22} and d_{23}; and the distance between the two sensors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, denoted by $\hat{d}_{12}$. Then, from the anchor and distance information we would like to locate the sensor positions $\boldsymbol{x}_1, \boldsymbol{x}_2 \in R^2$.

   Do the following numerical experiments using CVX (or cvxpy, convex.jl) or MOSEK and answer the questions:

   (a) (10') Generate two sensor points anywhere and try the SOCP relaxation model

$$\begin{aligned} \|\boldsymbol{x}_1 - \boldsymbol{a}_i\|^2 &\le d_{1i}^2, \ i = 1, 2 \\ \|\boldsymbol{x}_2 - \boldsymbol{a}_i\|^2 &\le d_{2i}^2, \ i = 2, 3 \\ \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 &\le \hat{d}_{12}^2. \end{aligned}$$

   Did you find the correct locations? What have you observed? Try a few different locations of the sensor pairs and identify the pattern.

   (b) (10') Now try the SDP relaxation: find $\boldsymbol{X} = [\boldsymbol{x}_1, \ \boldsymbol{x}_2] \in \mathbb{R}^{2\times 2}$ and

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{X} \\ \boldsymbol{X}^\top & \boldsymbol{Y} \end{pmatrix} \in S^4$$

   to meet the constraints in the standard form:

$$\begin{aligned} (1;0;0;0)(1;0;0;0)^\top \bullet \boldsymbol{Z} &= 1, \\ (0;1;0;0)(0;1;0;0)^\top \bullet \boldsymbol{Z} &= 1, \\ (1;1;0;0)(1;1;0;0)^\top \bullet \boldsymbol{Z} &= 2, \\ (\boldsymbol{a}_i;-1;0)(\boldsymbol{a}_i;-1;0)^\top \bullet \boldsymbol{Z} &= d_{1i}^2, \ i = 1, 2, \\ (\boldsymbol{a}_i;0;-1)(\boldsymbol{a}_i;0;-1)^\top \bullet \boldsymbol{Z} &= d_{2i}^2, \ i = 2, 3, \\ (0;0;1;-1)(0;0;1;-1)^\top \bullet \boldsymbol{Z} &= \hat{d}_{12}^2, \\ \boldsymbol{Z} &\succeq 0 \in S^4. \end{aligned}$$

Did you find the correct locations? What have you observed? Can you conclude with something? Try a few different locations of the sensor pairs and identify the pattern.

Solution

(a) For the SOCP formulation, in general we're not able to find the correct locations, even if both the sensors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are in the convex hull of the anchors $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$. To enable exact recovery, we need to require that $\boldsymbol{x}_1$ *is inside the convex hull of* $\boldsymbol{x}_2, \boldsymbol{a}_1, \boldsymbol{a}_2$ *and* $\boldsymbol{x}_2$ *is inside the convex hull of* $\boldsymbol{x}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$. We will validate this claim by numerical results.

Specifically, *if at least one sensor is outside the convex hull of the anchors* $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$, *we generally cannot ensure exact recovery.* So the exact recovery condition here is *stronger* than just requiring both sensors being inside the convex hull of the anchors.

We will revisit the problem and provide analysis of exact recovery in Assignment 2. Basically, we write down the first-order KKT conditions and keep in mind the non-positiveness of the multipliers. In particular, when the sensors are in the interior of the convex hulls, the corresponding multipliers are positive and hence by complementarity the inequalities become tight, which leads to exact recovery. On the other hand, when the sensors are on the boundaries of the convex hulls, then the inequality constraints already uniquely determine the points, and hence again we obtain exact recovery.

In contrast, if any of the convex hull inclusions of the two sensors is violated, then the corresponding (three) multipliers must all be 0. This will potentially lead to non-tight inequalities, disabling exact recovery.

In the experiments, we provide three options of generating sensors $\boldsymbol{x}_1, \boldsymbol{x}_2$ for facility of usage. The first is to generate both by `randn`. The second is to start by generating $\boldsymbol{x}_1$ inside the convex hull of $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$, and then generate $\boldsymbol{x}_2$ inside the convex hull of $\boldsymbol{x}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$. The third is to generate $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ independently inside the convex hull of $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$. Notice that the second option does not ensure that $\boldsymbol{x}_1$ is inside $\boldsymbol{x}_2, \boldsymbol{a}_1, \boldsymbol{a}_2$.

We showcase the five possible situations in Figure 1.

(1) $\boldsymbol{x}_1 \in \text{convhull}(\boldsymbol{x}_2, \boldsymbol{a}_1, \boldsymbol{a}_2), \boldsymbol{x}_2 \in \text{convhull}(\boldsymbol{x}_1, \boldsymbol{a}_2, \boldsymbol{a}_3), (\boldsymbol{x}_1, \boldsymbol{x}_2 \in \text{convhull}(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3))$;

(2) $\boldsymbol{x}_1 \notin \text{convhull}(\boldsymbol{x}_2, \boldsymbol{a}_1, \boldsymbol{a}_2), \boldsymbol{x}_2 \in \text{convhull}(\boldsymbol{x}_1, \boldsymbol{a}_2, \boldsymbol{a}_3), \boldsymbol{x}_1, \boldsymbol{x}_2 \in \text{convhull}(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$;

(3) $\boldsymbol{x}_1 \notin \text{convhull}(\boldsymbol{x}_2, \boldsymbol{a}_1, \boldsymbol{a}_2), \boldsymbol{x}_2 \notin \text{convhull}(\boldsymbol{x}_1, \boldsymbol{a}_2, \boldsymbol{a}_3), \boldsymbol{x}_1, \boldsymbol{x}_2 \in \text{convhull}(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$;

(4) one of $\boldsymbol{x}_1, \boldsymbol{x}_2$ outside $\text{convhull}(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$, and the other inside;

(5) both $\boldsymbol{x}_1, \boldsymbol{x}_2$ outside $\text{convhull}(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$.

We see that apart from the first case, we almost always lose exact recovery.

Also notice that switching the generation of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (i.e. first generate $\boldsymbol{x}_2$ inside the convex hull of $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ and then generate $\boldsymbol{x}_1$ inside the convex hull of $\boldsymbol{x}_2, \boldsymbol{a}_1, \boldsymbol{a}_2$) still results in the same observations, i.e. we obtain exact recovery if $\boldsymbol{x}_2$ is also inside the convex hull of $\boldsymbol{x}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$, and vice versa.

MATLAB code is attached below. The Julia code is online here.

```
clear all; close all;
%% Initialization
A = [1,-1,0;0,0,2];
%%% 1) random initialization choice (remove comment to enable)
% x1 = randn(2, 1);
% x2 = randn(2, 1);
 %%% 2) special initialization choice (add comment to disable)
 lambda1 = rand(3,1);
 x1 = A * lambda1 / sum(lambda1);
 lambda2 = rand(3,1);
 x2 = [x1, A(:,2:3)] * lambda2 / sum(lambda2);
%
% lambda2 = rand(3,1);
% x2 = A * lambda2 / sum(lambda2);
% lambda1 = rand(3,1);
% x1 = [x2, A(:,1:2)] * lambda1 / sum(lambda1);
%%% 3) random initialization choice inside the conv-hull ...
    (remove comment to
%%% enable)
% lambda1 = rand(3,1);
% x1 = A * lambda1 / sum(lambda1);
% lambda2 = rand(3,1);
% x2 = A * lambda2 / sum(lambda2);

%% Plot the figure
scatter(A(1,:),A(2,:), 'k', 'filled');
hold on;
```

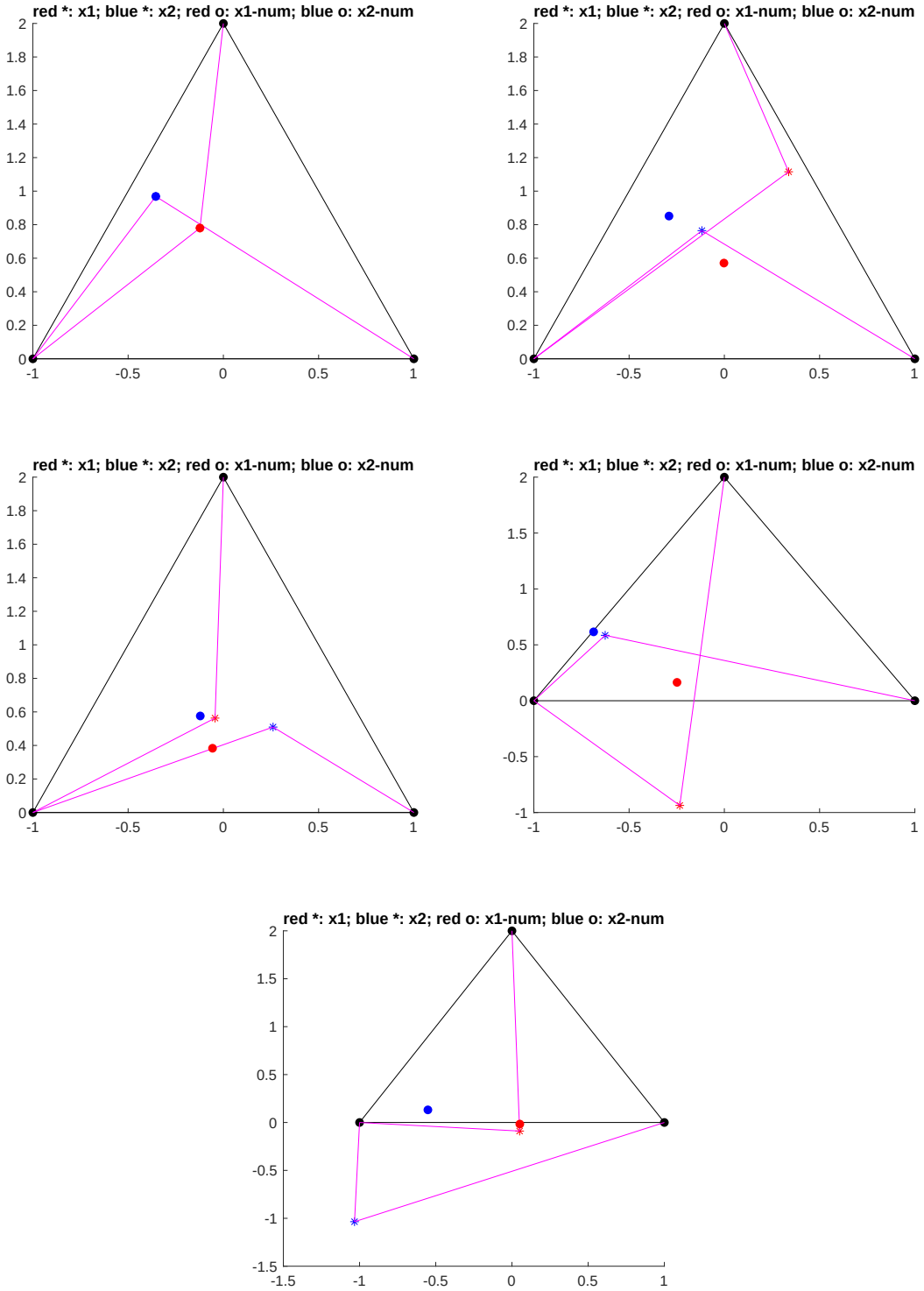


Figure 1: Top left: exact recovery, case 1. Top right: inexact recovery, case 2. Mid left: inexact recovery, case 3. Mid right: inexact recovery, case 4. Bottom: inexact recovery, case 5.

```
scatter(x1(1), x1(2), 'r*');
scatter(x2(1), x2(2), 'b*');
plot([-1,1], [0,0], 'k-');
plot([-1,0], [0,2], 'k-');
plot([1,0], [0,2], 'k-');

plot([x1(1),-1], [x1(2),0], 'm-');
plot([x1(1),0], [x1(2),2], 'm-');

plot([x2(1),1], [x2(2),0], 'm-');
plot([x2(1),-1], [x2(2),0], 'm-');

%% data generation
d11 = norm(x1-A(:,1));
d12 = norm(x1-A(:,2));
d22 = norm(x2-A(:,2));
d23 = norm(x2-A(:,3));
d12h = norm(x1-x2);
%% SOCP
cvx_begin
variables z1(2) z2(2)
minimize(0)
subject to
norm(z1-A(:,1))<=d11
norm(z1-A(:,2))<=d12
norm(z2-A(:,2))<=d22
norm(z2-A(:,3))<=d23
norm(z1-z2)<=d12h
cvx_end
fprintf('x1 error = %3.4e\n', norm(z1-x1));
fprintf('x2 error = %3.4e\n', norm(z2-x2));
scatter(z1(1), z1(2), 'ro', 'filled');
scatter(z2(1), z2(2), 'bo', 'filled');
title('red *: x1; blue *: x2; red o: x1-num; blue o: x2-num');
hold off
```

(b) For the SDP case, the recovery is still not always exact even if both sensors are inside the convex hull of the anchors $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$. But there are *more chances of recovering exactly than the SOCP formulation.*

When both $\boldsymbol{x}_1, \boldsymbol{x}_2$ are inside the convex hull of $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$, as long as one of the following two cases holds: 1) $\boldsymbol{x}_1$ is inside the convex hull of $\boldsymbol{x}_2, \boldsymbol{a}_1, \boldsymbol{a}_2$; 2) $\boldsymbol{x}_2$ is

inside the convex hull of $\boldsymbol{x}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$*, then we obtain exact recovery.* The trickier case is when $\boldsymbol{x}_1, \boldsymbol{x}_2$ are not both inside the convex hull of $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$, but (exactly) one of 1) and 2) holds. In this case, we sometimes get exact recovery while sometimes not. Then we may need some algebraic characterizations. In fact, the exact recovery is obtained if the optimal dual slack matrix (which can be retrieved e.g. using `dual variable` command in CVX) has rank n (the number of sensors), as shown in Lecture Note 5. This implies that any primal optimal/feasible solution $\boldsymbol{Z}$ has rank 2, and hence gives exact recovery.

Again, we will validate this claim by numerical experiments. In the numerical experiments, we again provide three options of generating sensors $\boldsymbol{x}_1, \boldsymbol{x}_2$ as before. We showcase three possible situations below with both $\boldsymbol{x}_1, \boldsymbol{x}_2$ inside the convex hull of $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ in Figure 2.

(1) $\boldsymbol{x}_1 \in \text{convhull}(\boldsymbol{x}_2, \boldsymbol{a}_1, \boldsymbol{a}_2), \boldsymbol{x}_2 \in \text{convhull}(\boldsymbol{x}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$, $(\boldsymbol{x}_1, \boldsymbol{x}_2 \in \text{convhull}(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3))$;

(2) $\boldsymbol{x}_1 \notin \text{convhull}(\boldsymbol{x}_2, \boldsymbol{a}_1, \boldsymbol{a}_2), \boldsymbol{x}_2 \in \text{convhull}(\boldsymbol{x}_1, \boldsymbol{a}_2, \boldsymbol{a}_3), \boldsymbol{x}_1, \boldsymbol{x}_2 \in \text{convhull}(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$;

(3) $\boldsymbol{x}_1 \notin \text{convhull}(\boldsymbol{x}_2, \boldsymbol{a}_1, \boldsymbol{a}_2), \boldsymbol{x}_2 \notin \text{convhull}(\boldsymbol{x}_1, \boldsymbol{a}_2, \boldsymbol{a}_3), \boldsymbol{x}_1, \boldsymbol{x}_2 \in \text{convhull}(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$;

In all other cases, we (almost) always lose exact recovery.

For example, when one of $\boldsymbol{x}_1, \boldsymbol{x}_2$ are outside the convex hull of $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$, things become much trickier, see Figure 3.

When exactly one of 1) or 2) holds, both exact and inexact recovery are possible, see Figure 4.

The MATLAB code is attached below. The Julia code is online here.

```
clear all; close all;
%% Initialization
A = [1,-1,0;0,0,2];
%%% 1) random initialization choice (remove comment to enable)
% x1 = randn(2, 1);
% x2 = randn(2, 1);
 %%% 2) special initialization choice (add comment to disable)
 lambda1 = rand(3,1);
 x1 = A * lambda1 / sum(lambda1);
 lambda2 = rand(3,1);
 x2 = [x1, A(:,2:3)] * lambda2 / sum(lambda2);
%
% lambda2 = rand(3,1);
```

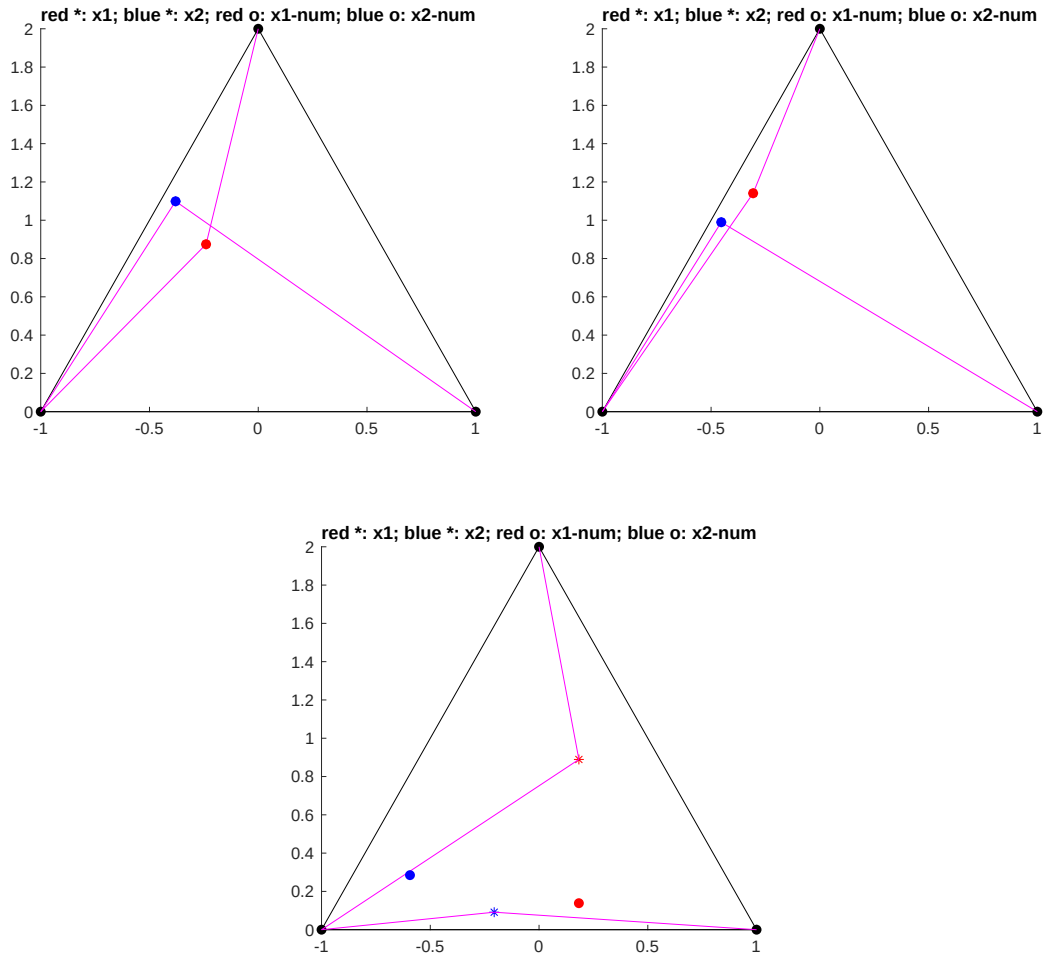


Figure 2: Top left: exact recovery, case 1. Top right: exact recovery, case 2. Bottom: inexact recovery, case 3.

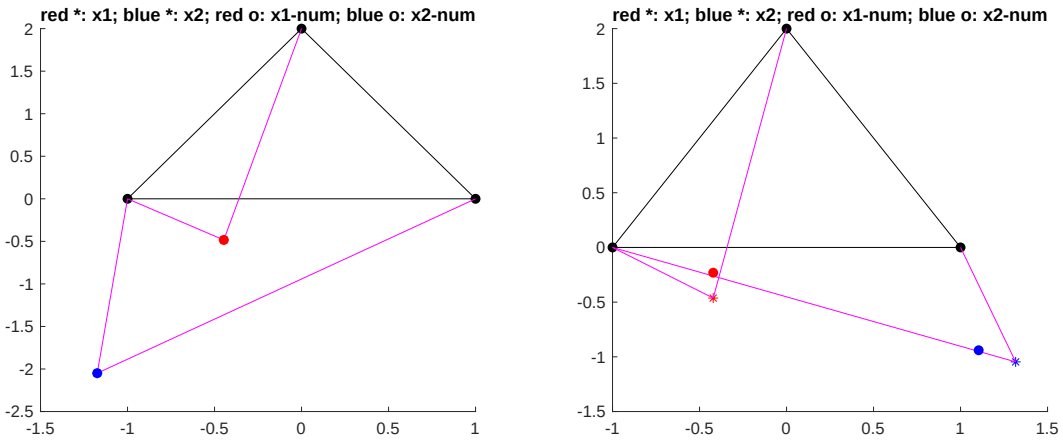


Figure 3: Exact and inexact recovery when $\boldsymbol{x}_1, \boldsymbol{x}_2$ are outside the convex hull of $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$.

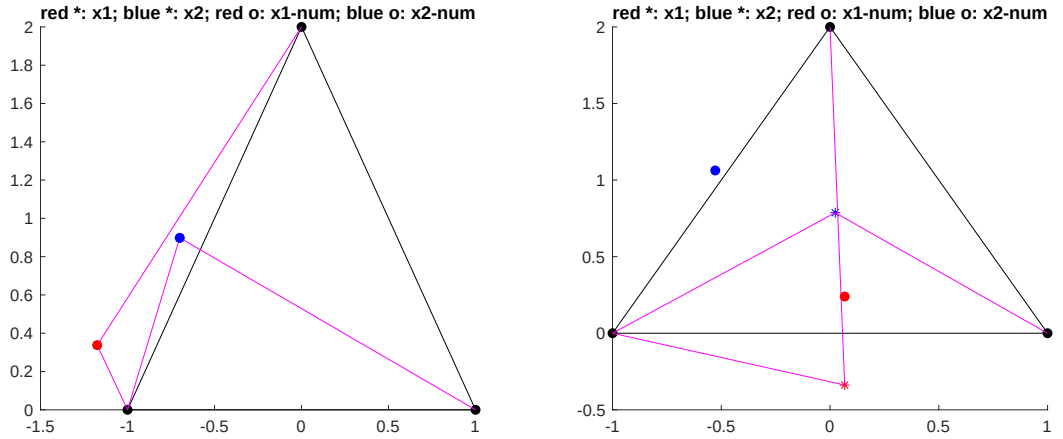


Figure 4: Exact and inexact recovery when one of 1) and 2) holds.

```
% x2 = A * lambda2 / sum(lambda2);
% lambda1 = rand(3,1);
% x1 = [x2, A(:,1:2)] * lambda1 / sum(lambda1);
%%% 3) random initialization choice inside the conv-hull ...
    (remove comment to
%%% enable)
% lambda1 = rand(3,1);
% x1 = A * lambda1 / sum(lambda1);
% lambda2 = rand(3,1);
% x2 = A * lambda2 / sum(lambda2);

%% Plot the figure
scatter(A(1,:),A(2,:), 'k', 'filled');
hold on;
scatter(x1(1), x1(2), 'r*');
scatter(x2(1), x2(2), 'b*');
plot([-1,1], [0,0], 'k-');
plot([-1,0], [0,2], 'k-');
plot([1,0], [0,2], 'k-');

plot([x1(1),-1], [x1(2),0], 'm-');
plot([x1(1),0], [x1(2),2], 'm-');

plot([x2(1),1], [x2(2),0], 'm-');
plot([x2(1),-1], [x2(2),0], 'm-');

%% data generation
d11 = norm(x1-A(:,1));
d12 = norm(x1-A(:,2));
d22 = norm(x2-A(:,2));
d23 = norm(x2-A(:,3));
d12h = norm(x1-x2);
%% SDP
a1 = A(:,1);
a2 = A(:,2);
a3 = A(:,3);
cvx_begin
variable Z(4,4) semidefinite
minimize(0)
subject to
Z(1:2,1:2) == eye(2, 2);
```

```
%%% constraint formulation 1
% [a1;-1;0]' * Z * [a1;-1;0] == d11^2;
% [a2;-1;0]' * Z * [a2;-1;0] == d12^2;
% [a2;0;-1]' * Z * [a2;0;-1] == d22^2;
% [a3;0;-1]' * Z * [a3;0;-1] == d23^2;
% [0;0;1;-1]' * Z * [0;0;1;-1] == d12h^2;
%%% constraint formulation 2
sum(sum(([a1;-1;0]*[a1;-1;0]').* Z)) == d11^2;
sum(sum(([a2;-1;0]*[a2;-1;0]').* Z)) == d12^2;
sum(sum(([a2;0;-1]*[a2;0;-1]').* Z)) == d22^2;
sum(sum(([a3;0;-1]*[a3;0;-1]').* Z)) == d23^2;
sum(sum(([0;0;1;-1]*[0;0;1;-1]').* Z)) == d12h^2;
cvx_end
z1 = Z(1:2,3);
z2 = Z(1:2,4);
fprintf('x1 error = %3.4e\n', norm(z1-x1));
fprintf('x2 error = %3.4e\n', norm(z2-x2));
scatter(z1(1), z1(2), 'ro', 'filled');
scatter(z2(1), z2(2), 'bo', 'filled');
title('red *: x1; blue *: x2; red o: x1-num; blue o: x2-num');
hold off
```

10. (10') For the Maze Runner example in Lecture Note #1, suppose that the blue-action at State 3 has a probability 0.5 leading to State 4 and 0.5 leading to State 5; and the only action at State 5 leads to State 0. Reformulate the MDP-LP problem with $\gamma = 0.9$ and solve it using any LP solver.

Solution LP formulation

$$
\begin{aligned}
\text{maximize}_{\boldsymbol{y}} \quad & y_0 + y_1 + y_2 + y_3 + y_4 + y_5, \\
\text{subjectto} \quad & y_0 \le \min\{0 + \gamma y_1, 0 + \gamma(0.5y_2 + 0.25y_3 + 0.125y_4 + 0.125y_5)\}, \\
& y_1 \le \min\{0 + \gamma y_2, 0 + \gamma(0.5y_3 + 0.25y_4 + 0.25y_5)\}, \\
& y_2 \le \min\{0 + \gamma y_3, 0 + \gamma(0.5y_4 + 0.5y_5)\}, \\
& y_3 \le \min\{0 + \gamma y_4, 0 + \gamma(0.5y_4 + 0.5y_5)\}, \\
& y_4 \le 1 + \gamma y_5, \\
& y_5 \le 0 + \gamma y_0.
\end{aligned} \tag{7}
$$

Solution

$$
\begin{aligned}
y_0^* &= 0.747207793362298, \qquad \pi_0^* = \text{Red.} \\
y_1^* &= 0.830230881521939, \qquad \pi_1^* = \text{Red.} \\
y_2^* &= 0.922478757256612, \qquad \pi_2^* = \text{Red.} \\
y_3^* &= 1.024976396962329, \qquad \pi_3^* = \text{Blue.} \\
y_4^* &= 1.605238312580964, \\
y_5^* &= 0.672487014018313.
\end{aligned} \tag{8}
$$

The sample MATLAB code is as follows. The Julia code is online here.

```
gamma = 0.9

cvx_begin
    variables y0 y1 y2 y3 y4 y5
    maximize y0 + y1 + y2 + y3 + y4 + y5
    subject to
       y0 ≤ 0 + gamma * y1
       y0 ≤ 0 + gamma * (0.5 * y2 + 0.25 * y3 + 0.125 * y4 + 0.125 ...
           * y5)
       y1 ≤ 0 + gamma * y2
       y1 ≤ 0 + gamma * (0.5 * y3 + 0.25 * y4 + 0.25 * y5)
       y2 ≤ 0 + gamma * y3
       y2 ≤ 0 + gamma * (0.5 * y4 + 0.5 * y5)
       y3 ≤ 0 + gamma * y4
       y3 ≤ 0 + gamma * (0.5 * y4 + 0.5 * y5)
       y4 ≤ 1 + gamma * y5
       y5 ≤ 0 + gamma * y0
cvx_end
```

3 Assignment 2

1. (15') Consider Problem 5 Assignment 1 where the second-order cone is replaced by the p-th order cone for $p \geq 1$:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & 2x_1 + x_2 + x_3 \\ \text{s.t.} \quad & x_1 + x_2 + x_3 = 1, \\ & x_1 - \|(x_2,\ x_3)\|_p \geq 0. \end{aligned}$$

   (a) (5') Write out the conic dual problem.

   (b) (5') Compute the dual optimal solution $(y^*, \boldsymbol{s}^*)$.

   (c) (5') Using the zero duality condition to compute the primal optimal solution $\boldsymbol{x}^*$.

Solution

(a) Following Lecture Note 3, the dual is

$$\max\ y \quad \text{s.t.} \quad y\boldsymbol{e} + \boldsymbol{s} = (2,1,1)^\top,\ s_1 - \|(s_2,\ s_3)\|_q \geq 0$$

or

$$\max\ y \quad \text{s.t.} \quad (2-y) - (2|1-y|^q)^{1/q} \geq 0$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

(b) If $y \geq 1$, the constraint can be written as $(2-y) - 2^{1/q}(y-1) \geq 0$ so that the maximal value is

$$y^* = \frac{2 + 2^{1/q}}{1 + 2^{1/q}} \geq 1.$$

Hence there is no need to consider the other case when $y < 1$. And $\boldsymbol{s}^* = (2 - y^*; 1 - y^*; 1 - y^*)$. For $p = 1, 2, \infty$, we have $y^* = 3/2, \sqrt{2}, y^* = 4/3$, respectively.

(c) From the zero duality condition, we have $2x_1^* + x_2^* + x_3^* = y^*$, and together with the constraints $x_1^* + x_2^* + x_3^* = 1$, we have

$$x_1^* = y^* - 1 = \frac{1}{1 + 2^{1/q}}, \quad x_2^* + x_3^* = \frac{2^{1/q}}{1 + 2^{1/q}}.$$

When $x_2^* = x_3^* = \frac{2^{1/q}}{2(1+2^{1/q})} > 0$,

$$\|(x_2^*; x_3^*)\|_p^p = 2\left(\frac{2^{1/q}}{2(1+2^{1/q})}\right)^p = \frac{1}{(1+2^{1/q})^p} 2^{1-p+p/q} = \frac{1}{(1+2^{1/q})^p} \leq (x_1^*)^p$$

so that it is feasible and, consequently, optimal.

This optimal solution is also unique, as we have

$$\frac{2^{1/q}}{1+2^{1/q}} = x_2^* + x_3^* \le \|(x_2^*; x_3^*)\|_p \|(1;1)\|_q = 2^{1/q}\|(x_2^*; x_3^*)\|_p$$

by Hölder's inequality, which implies that

$$\|(x_2^*; x_3^*)\|_p \ge \frac{1}{1+2^{1/q}} = x_1^*$$

and the equality is obtained iff $x_2^* = x_3^* = \frac{2^{1/q}}{2(1+2^{1/q})}$.

2. (20') Consider the distributionally robust optimization (DRO) problem

$$\min_{\boldsymbol{x}\in X} \max_{\boldsymbol{d}\in D} \sum_{k=1}^{N} (\hat{p}_k + d_k) h(\boldsymbol{x}, \boldsymbol{\xi}_k) \tag{9}$$

where the distribution set D is now given by

$$D = \left\{ \boldsymbol{d} : \sum_{k=1}^{N} d_k = 0,\ \|\boldsymbol{d}\|^2 \le 1/N,\ \hat{p}_k + d_k \ge 0,\ \forall k \in [N].\} \right\}$$

(a) (3') What is the interpretation of D? Answer within 2 sentences.

(b) (4') Represent D in standard conic form. (Hint: one set of the slack variables are in the second-order cone and the others are in the non-negative orthant cone.)

(c) (7') Construct the conic dual of the inner max-problem.

(d) (6') Replace the inner max-problem (9) by its dual, and simplify the DRO problem as much as possible.

Solution

(a) D denotes a set of bounded perturbations $\boldsymbol{d}$ (or slack variables) which keep the resulting $p_k := \hat{p}_k + d_k$, $k \in [N]$ a probability vector.

(b) The conic representation of D is

$$\left\{ (d_0; \boldsymbol{d}) : d_0 = 1/\sqrt{N}, \sum_{k=1}^{N} d_k = 0, \hat{p}_k + d_k = p_k, p_k \ge 0, \|\boldsymbol{d}\| \le d_0 \right\}$$

(c) Denoting $\boldsymbol{h} := (h(\boldsymbol{x}, \boldsymbol{\xi}_1); \ldots; h(\boldsymbol{x}, \boldsymbol{\xi}_N))$ and $\hat{\boldsymbol{p}} := (\hat{p}_1; \ldots; \hat{p}_N)$, and ignoring the constants $\sum_{k=1}^{N} \hat{p}_k h(\boldsymbol{x}, \boldsymbol{\xi}_k)$, the primal problem can be abbreviated as the following CLP:

$$\begin{aligned} \max_{d_0; \boldsymbol{d}; \boldsymbol{y}} \ & \boldsymbol{h}^\top \boldsymbol{d} \\ \text{s.t.} \ & d_0 = 1/\sqrt{N}, \boldsymbol{e}^\top \boldsymbol{d} = 0, \boldsymbol{d} - \boldsymbol{y} = -\hat{\boldsymbol{p}} \\ & (d_0; \boldsymbol{d}) \in SOC^{N+1}, \boldsymbol{y} \geq \boldsymbol{0} \end{aligned}$$

Suppose that $\lambda_0, \lambda_1, \boldsymbol{\lambda}_2$ are the multipliers for the corresponding equality constraints, then the dual problem is

$$\begin{aligned} \min_{\lambda_0, \lambda_1, \boldsymbol{\lambda}_2} \ & \lambda_0/\sqrt{N} - \boldsymbol{\lambda}_2^\top \hat{\boldsymbol{p}} \\ \text{s.t.} \ & (1; \boldsymbol{0}; \boldsymbol{0})\lambda_0 + (0; \boldsymbol{e}; \boldsymbol{0})\lambda_1 + (0; \boldsymbol{I}; -\boldsymbol{I})\boldsymbol{\lambda}_2 - (s_0; \boldsymbol{s}; \boldsymbol{z}) = (0; \boldsymbol{h}; \boldsymbol{0}) \\ & (s_0; \boldsymbol{s}) \in SOC^{N+1}, \boldsymbol{z} \geq 0 \end{aligned}$$

or equivalently,

$$\begin{aligned} \min_{\lambda_0, \lambda_1, \boldsymbol{\lambda}_2} \ & \lambda_0/\sqrt{N} - \boldsymbol{\lambda}_2^\top \hat{\boldsymbol{p}} \\ \text{s.t.} \ & \|\lambda_1 \boldsymbol{e} + \boldsymbol{\lambda}_2 - \boldsymbol{h}\| \leq \lambda_0 \\ & \boldsymbol{\lambda}_2 \leq \boldsymbol{0} \end{aligned}$$

which can be further simplified to

$$\begin{aligned} \min_{\lambda_1, \boldsymbol{\lambda}_2} \ & \|\lambda_1 \boldsymbol{e} + \boldsymbol{\lambda}_2 - \boldsymbol{h}\|/\sqrt{N} - \boldsymbol{\lambda}_2^\top \hat{\boldsymbol{p}} \\ \text{s.t.} \ & \boldsymbol{\lambda}_2 \leq \boldsymbol{0} \end{aligned}$$

(d) Replacing the inner-max problem with its dual in (c), we can reformulate the DRO problem as follows:

$$\begin{aligned} \min_{\boldsymbol{x} \in X, \lambda_1, \boldsymbol{\lambda}_2} \ & \hat{\boldsymbol{p}}^\top \boldsymbol{h} + \|\lambda_1 \boldsymbol{e} + \boldsymbol{\lambda}_2 - \boldsymbol{h}\|/\sqrt{N} - \boldsymbol{\lambda}_2^\top \hat{\boldsymbol{p}} \\ \text{s.t.} \ & \boldsymbol{\lambda}_2 \leq \boldsymbol{0} \end{aligned}$$

where $\hat{\boldsymbol{p}}$ and $\boldsymbol{h}$ are defined as in (c). When $\boldsymbol{x} \in X$ and $\boldsymbol{\lambda}_2 \leq \boldsymbol{0}$ are fixed, λ_1 can be partially solved out as

$$\lambda_1 = \boldsymbol{e}^\top (\boldsymbol{h} - \boldsymbol{\lambda}_2)/N = \frac{1}{N} \sum_{k=1}^{N} (h(\boldsymbol{x}, \boldsymbol{\xi}_k) - \boldsymbol{\lambda}_2^k)$$

and hence we finally arrive at

$$\begin{aligned} \min_{\boldsymbol{x}\in X,\boldsymbol{\lambda}_2} \quad & \hat{\boldsymbol{p}}^\top \boldsymbol{h} + \|\boldsymbol{H}_n(\boldsymbol{h}-\boldsymbol{\lambda}_2)\|/\sqrt{N} - \boldsymbol{\lambda}_2^\top \hat{\boldsymbol{p}} \\ \text{s.t.} \quad & \boldsymbol{\lambda}_2 \le \boldsymbol{0} \end{aligned}$$

where $\boldsymbol{H}_n := \boldsymbol{I} - \dfrac{\boldsymbol{e}\boldsymbol{e}^\top}{N}$ is the centralization matrix.

3. (10') Consider the SOCP relaxation in Problem 8 of Assignment 1:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^2} \quad & \boldsymbol{0}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \|\boldsymbol{x}-\boldsymbol{a}_i\|^2 \le d_i^2,\ i=1,2,3, \end{aligned}$$

   (a) (4') Write down the first-order KKT optimality conditions.

   (b) (3') Interpret (with no more than 2 sentences) the three optimal multipliers when the true position of the sensor is inside the convex hull of the three anchors.

   (c) (3') Could the true position $\bar{\boldsymbol{x}} \in \mathbb{R}^2$ of the sensor satisfy the optimality conditions if it is outside the convex hull of the three anchors? What would be the multiplier values?

Solution Let the Lagrangian or dual multipliers be $y_i \le 0$, $i = 1, 2, 3$.

   (a) Then, writing down the (first-order) KKT conditions, the optimal solution would satisfy

$$\sum_{i\in[3]} y_i(\boldsymbol{x}-\boldsymbol{a}_i) = \boldsymbol{0},$$

   and complementarity

$$y_i(d_i^2 - \|\boldsymbol{x}-\boldsymbol{a}_i\|^2) = 0,\ i \in [3].$$

   (b) When the true position $\bar{\boldsymbol{x}} \in \mathbb{R}^2$ is inside the convex hull, then y_i represents a force pulling $\bar{\boldsymbol{x}}$ from $\boldsymbol{a}_i$. The three forces balance at $\bar{\boldsymbol{x}}$ as the conditions indicated. In particular, when y_i's are not all zero, then we have $\bar{\boldsymbol{x}} = \frac{y_1\boldsymbol{a}_1+y_2\boldsymbol{a}_2+y_3\boldsymbol{a}_3}{y_1+y_2+y_3}$. Moreover, if all the forces are nonzero, then we find the correct solution. This is because the complementarity conditions then indicate that each constraint is tight, that is,

$$d_i^2 - \|\boldsymbol{x}-\boldsymbol{a}_i\|^2 = 0,\ \forall i = 1,2,3$$

   which mean that you find the $\boldsymbol{x}$ that satisfies all the original equality constraints. In this case, the relaxation is exact.

(c) It still satisfies the optimality conditions. But all multipliers must have 0 values, since otherwise we will have $\bar{\boldsymbol{x}} = \frac{y_1\boldsymbol{a}_1+y_2\boldsymbol{a}_2+y_3\boldsymbol{a}_3}{y_1+y_2+y_3}$ with $y_i \leq 0$, which is a point inside the convex hull. This leads to a contradiction. In this case, the $\boldsymbol{x}$ you find may not have all the constraints active, *i.e.*,

$$d_i^2 - \|\boldsymbol{x} - \boldsymbol{a}_i\|^2 = 0, \ \forall i = 1, 2, 3$$

may not all hold.

4. (10') Consider the following parametric QCQP problem for a parameter $\kappa > 0$:

$$\begin{aligned} \min \quad & (x_1 - 1)^2 + x_2^2 \\ \text{s.t.} \quad & -x_1 + \tfrac{x_2^2}{\kappa} \geq 0 \end{aligned}$$

(a) (5') Is $\boldsymbol{x} = \mathbf{0}$ a first-order KKT solution?

(b) (5') Is $\boldsymbol{x} = \mathbf{0}$ a second-order KKT necessary or sufficient solution for some value of κ?

Solution Define $f(\boldsymbol{x}) := (x_1 - 1)^2 + x_2^2$, $c(\boldsymbol{x}) = -x_1 + \frac{x_2^2}{\kappa}$. Then the Lagrangian function for this problem is

$$L(\boldsymbol{x}, y) = f(\boldsymbol{x}) - yc(\boldsymbol{x}) = (x_1 - 1)^2 + x_2^2 - y\left(-x_1 + \frac{x_2^2}{\kappa}\right), \ y \geq 0.$$

(a) Firstly, $\boldsymbol{x} = \mathbf{0}$ is feasible with $c(\boldsymbol{x}) = 0$. Moreover,

$$\nabla f(\mathbf{0}) = (-2;\ 0), \quad \nabla c(\mathbf{0}) = (-1;\ 0)$$

Thus $y = 2$ makes $\nabla f(\mathbf{0}) = 2\nabla c(\mathbf{0})$ so that $\boldsymbol{x} = 0$ is a first-order KKT solution.

(b) Since the constraint is active, the tangent space is

$$T = \{\boldsymbol{d} : \ \boldsymbol{d} \in \mathbb{R}^2, \ (-1, \ 0)\boldsymbol{d} = 0\}.$$

The second-order necessary condition implies that for all $\boldsymbol{d} \in T$

$$\boldsymbol{d}^\top \nabla_x^2 L(\bar{\boldsymbol{x}}, \bar{y})\boldsymbol{d} \geq 0,$$

where

$$\nabla_x^2 L(\mathbf{0}, 2) = \begin{pmatrix} 2 & 0 \\ 0 & 2 - \frac{4}{\kappa} \end{pmatrix}$$

Thus, when $\kappa \geq 2$, the Hessian matrix of the Lagrangian is PSD so that $\boldsymbol{x} = \mathbf{0}$ is a second-order KKT solution. Otherwise, $\boldsymbol{x} = \mathbf{0}$ cannot be a local minimizer.

5. (20') (Central-Path and Potential) Given standard LP problem

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \quad \boldsymbol{x} \geq \boldsymbol{0}. \end{aligned} \tag{LP}$$

The **Analytic Center** of the primal feasible region $\mathcal{F}_p := \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$ is defined as the solution of the following linear-constrained convex optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & -\sum_{j=1}^{n} \log x_j, \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \quad \boldsymbol{x} > \boldsymbol{0}. \end{aligned} \tag{PB}$$

The **Central Path** $\boldsymbol{x}(\mu)$ of (LP) is defined as the solution of the following Barrier LP problem (where $\mu > 0$ is a parameter):

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} - \mu \cdot \sum_{j=1}^{n} \log x_j, \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \quad \boldsymbol{x} > \boldsymbol{0}. \end{aligned} \tag{BLP}$$

Part I Now consider the following example:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^3} \quad & x_1 + x_2, \\ \text{s.t.} \quad & x_1 + x_2 + x_3 = 1, \\ & (x_1, x_2, x_3) \geq \boldsymbol{0}. \end{aligned} \tag{10}$$

(a) (4') What is the analytic center of the primal feasible region in (10)?

(b) (4') Find the central path $\boldsymbol{x}(\mu) = (x_1(\mu), x_2(\mu), x_3(\mu))$ for (10).

(c) (4') Show that as μ decreases to 0, $\boldsymbol{x}(\mu)$ converges to the unique optimal solution of (10).

Part II Consider another example with different objective but the same feasible region:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^3} \quad & x_1 \\ \text{s.t.} \quad & x_1 + x_2 + x_3 = 1 \\ & (x_1, x_2, x_3) \geq \boldsymbol{0} \end{aligned} \tag{11}$$

(d) (4') Find the central path $\boldsymbol{x}(\mu) = (x_1(\mu), x_2(\mu), x_3(\mu))$ for (11).

(e) (4') Which point does the central path converge to now (as $\mu \to 0+$)?

Solution

(a) The analytic center is the vector that minimizes the potential function:

$$-\sum_{j=1}^{3} \log x_j$$

and satisfies $\sum_{j=1}^{3} x_j = 1$, $\boldsymbol{x} > 0$. Thus the analytic center is $(1/3; 1/3; 1/3)$.

(b) From the central path condition we derive a quadratic equation for x_1:

$$2x_1^2 - (3\mu + 1)x_1 + \mu = 0.$$

Taking the non-negative root gives

$$x_1 = \frac{3\mu + 1 - \sqrt{9\mu^2 + 1 - 2\mu}}{4}.$$

Other conditions give $x_2 = x_1$ and $x_3 = 1 - 2x_1$.

(c) The set of optimal solution is a singleton $(0; 0; 1)$. When μ decreases to zero, we know from the expression that $x_1(\mu) = x_2(\mu) \to 0$. Also, since $\sum_i x_i = 1$ always holds, we know that $x_3 \to 1$. We know that $(0, 0, 1)$ is going to be the optimal solution, because $f(\boldsymbol{x}) = x_1 + x_2 \geq 0$, and $(0, 0, 1)$ attains the value 0. The uniqueness is easily proved: to attain optimal value, x_1, x_2 has to be zero, so x_3 have to be 1, because of the equality constraint.

Thus, as μ goes to zero, $\boldsymbol{x}(\mu)$ converges to the unique optimal solution.

(d)(e) Just repeat the above stuff. The only thing to be noted is that now the optimal solution to the original problem is **not unique**, so the problem description in (c) needs to be slightly changed. But everything else is the same.

6. (15') Consider the following SVM problem, where $\mu \geq 0$ is a prescribed constant:

$$\begin{aligned} \min \quad & \beta + \mu \|\boldsymbol{x}\|^2 \\ \text{s.t.} \quad & \boldsymbol{a}_i^\top \boldsymbol{x} + x_0 + \beta \geq 1, \ \forall i, \\ & \boldsymbol{b}_j^\top \boldsymbol{x} + x_0 - \beta \leq -1, \ \forall j, \\ & \beta \geq 0. \end{aligned}$$

(a) (8') Write out the Lagrangian dual problem of the SVM problem. Write it as explicit as possible (at least remove the inner minimization). (Hint: You may want to consider two separate cases: $\mu = 0$ and $\mu > 0$)

(b) (7') Suppose that we have 6 training data in $\mathbb{R}^2$: $\boldsymbol{a}_1 = (0; 0)$, $\boldsymbol{a}_2 = (1; 0)$, $\boldsymbol{a}_3 = (0; 1)$ and $\boldsymbol{b}_1 = (0; 0)$, $\boldsymbol{b}_2 = (-1; 0)$, $\boldsymbol{b}_3 = (0; -1)$. Use the optimality conditions (or any approach you want) to find optimal solutions for $\mu = 0$ and $\mu = 10^{-5}$, respectively. Are the two optimal solutions unique for the given μ? Prove your claim.

Solution

(a) Let the multipliers for $\boldsymbol{a}_i$ constraints be $y_i^a \geq 0$ and those for $\boldsymbol{b}_j$ constraints be $y_j^b \leq 0$, and $\beta \geq 0$ be $y^\beta \geq 0$. Then, the Lagrangian function is

$$L(\boldsymbol{x}, x_0, \beta, \boldsymbol{y}^a, \boldsymbol{y}^b, y^\beta) = \beta + \mu\|\boldsymbol{x}\|^2 - \sum_i y_i^a(\boldsymbol{a}_i^\top \boldsymbol{x} + x_0 + \beta - 1) - \sum_j y_j^b(\boldsymbol{b}_j^\top \boldsymbol{x} + x_0 - \beta + 1) - y^\beta \beta.$$

The dual must have constraint (by taking derivative w.r.t. x_0 and β)

$$\sum_i y_i^a + \sum_j y_j^b = 0$$

and

$$1 - y^\beta - \sum_i y_i^a + \sum_j y_j^b = 0,$$

since otherwise the primal can choose x_0 or β to make the Lagrangian function unbounded from below.

1) If $\mu = 0$, then we also have

$$\sum_i y_i^a \boldsymbol{a}_i + \sum_j y_j^b \boldsymbol{b}_j = \boldsymbol{0},$$

since otherwise the primal can choose $\boldsymbol{x}$ to make the Lagrangian function unbounded below.

Hence, the dual problem is

$$\begin{array}{ll} \max & \sum_i y_i^a - \sum_j y_j^b, \\ \text{s.t.} & \sum_i y_i^a + \sum_j y_j^b = 0, \\ & 1 - y^\beta - \sum_i y_i^a + \sum_j y_j^b = 0, \\ & \sum_i y_i^a \boldsymbol{a}_i + \sum_j y_j^b \boldsymbol{b}_j = \boldsymbol{0}, \\ & \boldsymbol{y}^a \geq \boldsymbol{0}, \boldsymbol{y}^b \leq \boldsymbol{0}, \ y^\beta \geq 0. \end{array}$$

2) For $\mu > 0$, from Lagrangian Derivative Conditions (w.r.t. $\boldsymbol{x}$) we know

$$2\mu\boldsymbol{x} = \sum_i y_i^a \boldsymbol{a}_i + \sum_j y_j^b \boldsymbol{b}_j.$$

Thus,

$$\phi(\boldsymbol{y}^a, \boldsymbol{y}^b, y^\beta) = -\frac{1}{4\mu}\|\sum_i y_i^a \boldsymbol{a}_i + \sum_j y_j^b \boldsymbol{b}_j\|^2 + \sum_i y_i^a - \sum_j y_j^b,$$

and the dual problem is

$$\begin{array}{ll} \max & \phi(\boldsymbol{y}^a, \boldsymbol{y}^b, y^\beta) \\ \text{s.t.} & \sum_i y_i^a + \sum_j y_j^b = 0, \\ & 1 - y^\beta - \sum_i y_i^a + \sum_j y_j^b = 0, \\ & \boldsymbol{y}^a \geq \boldsymbol{0},\ \boldsymbol{y}^b \leq \boldsymbol{0},\ y^\beta \geq 0. \end{array}$$

(b) Firstly, we show that for the set of $\boldsymbol{a}_i, \boldsymbol{b}_j$ given in this problem, any feasible β satisfies $\beta \geq 1$. To see this, suppose on the contrary that $\beta < 1$. Then for $\boldsymbol{a}_1 = \boldsymbol{b}_1$, we have

$$\boldsymbol{a}_1^\top \boldsymbol{x} + x_0 \geq 1 - \beta > 0 > -1 + \beta \geq \boldsymbol{b}_1^\top \boldsymbol{x} + x_0,$$

which is a contradiction. Hence the optimal value $\beta + \mu\|\boldsymbol{x}\|^2$ of the primal objective function is at least 1. Moreover, it can always be achieved by simply setting $\beta = 1$, $\boldsymbol{x} = \boldsymbol{0}$ and $x_0 = 0$. Hence we know that the optimal value is always 1 no matter whether $\mu = 0$ or not.

1) For $\mu = 0$, any point of the form $\beta = 1$, $\boldsymbol{x} = (t; t)$, $x_0 = 0$ with $t \geq 0$ is optimal, as the objective value is 1 and the constraints are satisfied. So the optimal solution is not unique.

2) For $\mu > 0$, a point is optimal iff $\beta = 1$ and $\boldsymbol{x} = \boldsymbol{0}$, since otherwise we will have $\beta + \mu\|\boldsymbol{x}\|^2 > \beta \geq 1$. In this case, we need $x_0 \geq 0 \geq x_0$, and hence $x_0 = 0$. Hence we obtain a unique optimal solution $\beta = 1$, $\boldsymbol{x} = \boldsymbol{0}$ and $x_0 = 0$.

7. (20') Consider a generalized Arrow–Debreu equilibrium problem in which the market has n agents and m goods. Agent i, $i = 1, ..., n$, has a bundle amount of $\boldsymbol{w}_i = (w_{i1}, w_{i2}, \ldots, w_{im}) \in \mathbb{R}^m_+$ goods initially and has a linear utility function whose coefficients are $\boldsymbol{u}_i = (u_{i1}, u_{i2}, \ldots, u_{im}) > \boldsymbol{0} \in \mathbb{R}^m$. The goal is to price each good so that the market clears. Note that, given the price vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_m) > \boldsymbol{0}$, agent i's

utility maximization problem is:

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{u}_i^\top \boldsymbol{x}_i \\ \text{subject to} \quad & \boldsymbol{p}^\top \boldsymbol{x}_i \le \boldsymbol{p}^\top \boldsymbol{w}_i \\ & \boldsymbol{x}_i \ge \boldsymbol{0} \end{aligned}$$

(a) (5') For a given $\boldsymbol{p} \in \mathbb{R}^m$, write down the optimality conditions for agent i's utility maximization problem. Without loss of generality, you may fix $p_m = 1$ since the budget constraints are homogeneous in $\boldsymbol{p}$.

(b) (5') Suppose that $\boldsymbol{p} \in \mathbb{R}^m$ and $\boldsymbol{x}_i \in \mathbb{R}^m$ satisfy the constraints:

$$\begin{aligned} \sum_{i=1}^n \boldsymbol{x}_i &= \sum_{i=1}^n \boldsymbol{w}_i, \\ \frac{\boldsymbol{u}_i^\top \boldsymbol{x}_i}{\boldsymbol{p}^\top \boldsymbol{w}_i} p_j &\ge u_{ij}, \quad \forall i, j, \\ \boldsymbol{p} &\ge \boldsymbol{0}, \\ \boldsymbol{x}_i &\ge \boldsymbol{0}, \quad \forall i. \end{aligned}$$

Show that $\boldsymbol{p}$ is then an equilibrium price vector.

(c) (5') For simplicity, assume that all u_{ij} are positive so that all p_j are positive. By introducing new variables $y_j = \log(p_j)$ for $j = 1, ..., m$, the conditions can be written as follows:

$$\begin{aligned} \min \quad & 0 \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^n \boldsymbol{x}_i = \sum_{i=1}^n \boldsymbol{w}_i \\ & \log(\boldsymbol{u}_i^\top \boldsymbol{x}_i) - \log\left(\textstyle\sum_{k=1}^m w_{ik} e^{y_k}\right) + y_j \ge \log(u_{ij}) \quad \forall i, j \\ & x_{ij} \ge 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \forall i, j \end{aligned}$$

Show that this problem is convex in x_{ij} and y_j. (Hint: Use the fact that $\log\left(\sum_{k=1}^m w_{ik} e^{y_k}\right)$ is a convex function in the y_k's.)

(d) (5') Consider the Fisher example on Lecture Note with two agents and two goods, where the utility coefficients are given by

$$\boldsymbol{u}_1 = (2;\ 1) \quad \text{and} \quad \boldsymbol{u}_2 = (3;\ 1),$$

while now there are no fixed budgets. Rather, let

$$\boldsymbol{w}_1 = (1;\ 0) \quad \text{and} \quad \boldsymbol{w}_2 = (0;\ 1)$$

that is, agent 1 brings in one unit good x and agent brings in one unit of good y. Find the Arrow–Debreu equilibrium prices, where you may assume $p_y = 1$.

Solution

(a) Notice that here $\boldsymbol{p}$ is fixed, and hence the problem is simply an LP. Writing down the primal feasibility, dual feasibility and zero duality gap conditions, we obtain:

$$\boldsymbol{u}_i \leq \lambda_i \boldsymbol{p}, \quad \lambda_i \geq 0, \quad \lambda_i \cdot \boldsymbol{p}^\top \boldsymbol{w}_i = \boldsymbol{u}_i^\top \boldsymbol{x}_i, \quad \boldsymbol{x}_i \geq \boldsymbol{0}, \quad \boldsymbol{p}^\top \boldsymbol{x}_i \leq \boldsymbol{p}^\top \boldsymbol{w}_i.$$

Alternative solution: write down the KKT conditions – the zero duality gap condition $\lambda_i \cdot \boldsymbol{p}^\top \boldsymbol{w}_i = \boldsymbol{u}_i^\top \boldsymbol{x}_i$ will be replaced by the zero gradient condition for the Lagrangian. Notice that these are equivalent.

(b) This proof is identical to the Lecture Note 5 for Fisher equilibrium where scalar $\boldsymbol{w}_i$ is substituted by $\boldsymbol{p}^\top \boldsymbol{w}_i$. In particular, we simply check that $\boldsymbol{x}_i$ are all optimal for the given $\boldsymbol{p}$ in their own utility maximization LPs, *i.e.*, we check that the optimality conditions in (a) are all satisfied.

Firstly, define $\lambda_i := \frac{\boldsymbol{u}_i^\top \boldsymbol{x}_i}{\boldsymbol{p}^\top \boldsymbol{w}_i}$. Then obviously we have $\lambda_i \geq 0$ and $\lambda_i \boldsymbol{p} \geq \boldsymbol{u}_i$ by the second set of constraints in (b). Moreover, by definition, we have $\lambda_i \boldsymbol{p}^\top \boldsymbol{w}_i = \boldsymbol{u}_i^\top \boldsymbol{x}_i$, and $\boldsymbol{x}_i \geq \boldsymbol{0}$ is satisfied automatically by the third set of constraints in (b).

It remains to check that $\boldsymbol{p}^\top \boldsymbol{x}_i \leq \boldsymbol{p}^\top \boldsymbol{w}_i$. To see this, multiply both sides of the first set of constraints in (b) by $\boldsymbol{p}^\top$, we have

$$\sum_{i=1}^{n} \boldsymbol{p}^\top \boldsymbol{x}_i = \sum_{i=1}^{n} \boldsymbol{p}^\top \boldsymbol{w}_i$$

On the other hand, multiplying both sides of the second set of constraints in (b) by x_{ij} and sum over j, we have

$$\frac{\boldsymbol{u}_i^\top \boldsymbol{x}_i}{\boldsymbol{p}^\top \boldsymbol{w}_i} \boldsymbol{p}^\top \boldsymbol{x}_i \geq \boldsymbol{u}_i^\top \boldsymbol{x}_i$$

and since $\boldsymbol{u}_i > \boldsymbol{0}$ by assumption, we have $\frac{\boldsymbol{u}_i^\top \boldsymbol{x}_i}{\boldsymbol{p}^\top \boldsymbol{w}_i}, p_j$ both strictly large than 0 (since otherwise the second set of constraints in (b) would be violated). In particular, we have $\boldsymbol{u}_i^\top \boldsymbol{x}_i > 0$, and hence we can divide it on both sides of the above inequality, and obtain that $\boldsymbol{p}^\top \boldsymbol{x}_i \geq \boldsymbol{p}^\top \boldsymbol{w}_i$. Combining this with the fact that $\sum_{i=1}^n \boldsymbol{p}^\top \boldsymbol{x}_i = \sum_{i=1}^n \boldsymbol{p}^\top \boldsymbol{w}_i$, we conclude that $\boldsymbol{p}^\top \boldsymbol{w}_i = \boldsymbol{p}^\top \boldsymbol{x}_i \geq \boldsymbol{p}^\top \boldsymbol{x}_i$, which finishes our proof.

(c) We first observe that the function $\log(\boldsymbol{u}_i^\top \boldsymbol{x}_i)$ is concave in $\boldsymbol{x}_i$, and that the function $g : \mathbb{R}^m \to R$ given by $g(\boldsymbol{y}) := \log\left(\sum_{k=1}^m w_{ik} e^{y_k}\right)$ is convex in $\boldsymbol{y}$. The former is

obvious. To establish the latter, we can check the Hessian of g is PSD. However, an easier way is to show that the epigraph of the function g is convex. Note that

$$\begin{aligned} & \{(y,t) : \log\left(\textstyle\sum_{k=1}^m w_{ik} e^{y_k}\right) \le t\} \\ = \; & \{(y,t) : \textstyle\sum_{k=1}^m w_{ik} e^{y_k} \le e^t\} \\ = \; & \{(y,t) : \textstyle\sum_{k=1}^m w_{ik} e^{y_k - t} \le 1\} \end{aligned}$$

Since the function $\sum_{k=1}^m w_{ik} e^{y_k - t}$ is clearly convex in $(\boldsymbol{y}, t)$ by convexity preserving operations (conic combination and affine transformation), its sublevel set must be convex.

Hence, we conclude that the inequalities:

$$\log\left(\sum_{k=1}^m w_{ik} e^{y_k}\right) - \log\left(u_i^\top \boldsymbol{x}_i\right) - y_j \le -\log(u_{ij}) \quad \forall i, j$$

define a convex set. As the remaining constraints and the objective function are linear, we conclude that the problem is a convex minimization problem.

(d) We can solve an exponential cone optimization problem in (c) to obtain the solutions:

$$p_x = 2,\ p_y = 1,\ x_1 = 1/2,\ y_1 = 1,\ x_2 = 1/2,\ y_2 = 0.$$

We can also use (b) to find $p_x, x_1, y_1, x_2, y_2 \ge 0$ with $p_y = 1$, such that:

$$\begin{aligned} x_1 + x_2 &= 1 \\ y_1 + y_2 &= 1 \\ \frac{2x_1 + y_1}{p_x} p_x &\ge 2 \\ \frac{2x_1 + y_1}{p_x} p_y &\ge 1 \\ \frac{3x_2 + y_2}{p_y} p_x &\ge 3 \\ \frac{3x_2 + y_2}{p_y} p_y &\ge 1 \end{aligned} \quad .$$

8. (Optional:) Consider the dual problem of an SDP,

$$\begin{aligned} \max_{\boldsymbol{y}, S} \quad & by \\ \text{subject to } & \boldsymbol{A}y + \boldsymbol{S} = \boldsymbol{C} \\ & \boldsymbol{S} \succeq \boldsymbol{0}, \end{aligned}$$

where $\boldsymbol{A}, \boldsymbol{C} \in \mathbb{S}^3$ is given. If $\boldsymbol{A}$ is not zero and the above problem is solvable, show that it has a solution $(\boldsymbol{y}, \boldsymbol{S})$ satisfies $\text{rank}(\boldsymbol{S}) \leq 2$. (Hint: apply Carathéodory's theorem)

Solution First, we reformulate this problem in a standard SDP form. Since $\boldsymbol{A} = \{a_{ij}\}_{i,j=1}^3$ is not a zero matrix, we first assume $a_{11} \neq 0$ w.l.o.g.. Then, we can eliminate y by the substitution $y = \frac{\langle \boldsymbol{e}_{11}, \boldsymbol{C}-\boldsymbol{S}\rangle}{a_{11}}$ so that the dual problem can be reformulated as

$$\begin{aligned} &\min \langle \boldsymbol{S} - \boldsymbol{C}, b\boldsymbol{e}_{11}\rangle \\ &\text{s.t. } \langle \boldsymbol{S} - \boldsymbol{C}, \boldsymbol{e}_{ij} - \boldsymbol{e}_{11}a_{ij}/a_{11}\rangle = 0, \ \forall 1 \leq i \leq j \leq 3 \\ &\quad \boldsymbol{S} \succeq \boldsymbol{0}, \end{aligned}$$

where $\boldsymbol{e}_{ij}$ is a matrix with one at (i, j) entry and zero otherwise. Here, the first constraint comes from $\boldsymbol{A}y + \boldsymbol{S} = \boldsymbol{C}$. Next, we apply Carathéodory's theorem (Theorem 8 in Lecture Note 5) to draw the conclusion. Notice that the condition is satisfies automatically when $i = j = 1$. We eliminate this constraint, and the new SDP problem only has 5 equality constraints. By Carathéodory's theorem, the rank r of one optimal solution satisfies

$$r(r+1) \leq 10,$$

which implies $r \leq 2$.

Groupwork (40') (group of 1-4 people):

9. (5') Let $\{(\boldsymbol{a}_i, c_i)\}_{i=1}^m$ be a given dataset where $\boldsymbol{a}_i \in \mathbb{R}^n$, $c_i \in \{\pm 1\}$. In Logistic Regression (LR), we determine $x_0 \in R$ and $\boldsymbol{x} \in \mathbb{R}^n$ by maximizing

$$\left(\prod_{i,c_i=1} \frac{1}{1+\exp(-\boldsymbol{a}_i^\top \boldsymbol{x} - x_0)}\right)\left(\prod_{i,c_i=-1} \frac{1}{1+\exp(\boldsymbol{a}_i^\top \boldsymbol{x} + x_0)}\right).$$

which is equivalent to maximizing the log-likelihood probability

$$-\sum_{i,c_i=1} \log\left(1+\exp(-\boldsymbol{a}_i^\top \boldsymbol{x} - x_0)\right) - \sum_{i,c_i=-1} \log\left(1+\exp(\boldsymbol{a}_i^\top \boldsymbol{x} + x_0)\right).$$

In this problem, we consider the quadratic regularized log-logistic-loss function

$$f(\boldsymbol{x}, x_0) = \sum_{i,c_i=1} \log\left(1+\exp(-\boldsymbol{a}_i^\top \boldsymbol{x} - x_0)\right) + \sum_{i,c_i=-1} \log\left(1+\exp(\boldsymbol{a}_i^\top \boldsymbol{x} + x_0)\right) + 0.001 \cdot \|\boldsymbol{x}\|_2^2.$$

Consider the following data set

$$\boldsymbol{a}_1 = (0;0), \quad \boldsymbol{a}_2 = (1;0), \quad \boldsymbol{a}_3 = (0;1), \quad \boldsymbol{a}_4 = (0;0), \quad \boldsymbol{a}_5 = (-1;0), \quad \boldsymbol{a}_6 = (0;-1),$$

with label

$$c_1 = c_2 = c_3 = 1, \quad c_4 = c_5 = c_6 = -1$$

use the KKT conditions to find a solution of min $f(\boldsymbol{x}, x_0)$. You can either solve it numerically (e.g., using MATLAB `fsolve`) or analytically (represent the solution by a solution of a simpler (1D) nonlinear equation).

Solution Since the problem is unconstrained, the KKT condition is nothing but setting $\nabla f(\boldsymbol{x}, x_0)$ to zero. Let $\boldsymbol{x} = (x_1; x_2)$, the KKT condition can be written coordinate-wise as

$$\begin{aligned}
0 &= \frac{-1}{1+\exp(x_0)} + \frac{-1}{1+\exp(x_0+x_1)} + \frac{-1}{1+\exp(x_0+x_2)} + \frac{1}{1+\exp(-x_0)} + \frac{1}{1+\exp(-x_0+x_1)} + \frac{1}{1+\exp(-x_0+x_2)} \\
0 &= -\frac{1}{1+\exp(x_0+x_1)} - \frac{1}{1+\exp(-x_0+x_1)} + 0.002x_1 \\
0 &= -\frac{1}{1+\exp(x_0+x_2)} - \frac{1}{1+\exp(-x_0+x_2)} + 0.002x_2
\end{aligned} \tag{12}$$

Note that if $x_0 = 0$ then the first equation of (12) automatically holds. Assuming $x_0 = 0$, the last two equations becomes

$$x_1 = \frac{1000}{1+\exp(x_1)}, \quad x_2 = \frac{1000}{1+\exp(x_2)}$$

Hence it suffices to set $x_1 = x_2$ to be the (unique) solution of nonlinear equation $z(1 + e^z) = 1000$. The approximate solution of this nonlinear equation is 5.2452. Consequently a KKT solution is

$$\boldsymbol{x}^* \approx (5.2452; 5.2452), \quad x_0^* = 0.$$

Remark: You can also numerically solve (12) using your favorite solvers (e.g., MATLAB function `fsolve`) or solve the original problem with exponential cones. The Julia code is online here.

10. (15') Consider standard LP problem

$$\begin{array}{ll} \min\limits_{\boldsymbol{x}\in\mathbb{R}^n} & \boldsymbol{c}^\top \boldsymbol{x}, \\ \text{s.t.} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \quad \boldsymbol{x} \geq \boldsymbol{0}. \end{array} \tag{LP}$$

with its dual

$$\begin{array}{ll} \max\limits_{\boldsymbol{y}\in\mathbb{R}^m, \boldsymbol{s}\in\mathbb{R}^n} & \boldsymbol{b}^\top \boldsymbol{y}, \\ \text{s.t.} & \boldsymbol{A}\boldsymbol{y} + \boldsymbol{s} = \boldsymbol{c}, \quad \boldsymbol{s} \geq \boldsymbol{0} \end{array} \tag{LD}$$

For any $\boldsymbol{x} \in \text{int } \mathcal{F}_p := \{\boldsymbol{x} \in \mathbb{R}^n : A\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} > 0\}$ and $\boldsymbol{s} \in \text{int } \mathcal{F}_d := \{\boldsymbol{s} \in \mathbb{R}^n : \boldsymbol{s} = \boldsymbol{c} - \boldsymbol{A}^\top \boldsymbol{y}, \boldsymbol{s} > \boldsymbol{0}, \boldsymbol{y} \in \mathbb{R}^m\}$, the **Primal-Dual Potential Function** is defined by

$$\psi_{n+\rho}(\boldsymbol{x}, \boldsymbol{s}) := (n + \rho) \log(\boldsymbol{x}^\top \boldsymbol{s}) - \sum_{j=1}^{n} \log(\boldsymbol{x}_j \boldsymbol{s}_j)$$

where $\rho > 0$ is a parameter.

Task: For two LP examples in Problem 5, namely (10) and (11), draw $\boldsymbol{x}$ part of the primal-dual potential function level sets

$$\psi_6(\boldsymbol{x}, \boldsymbol{s}) \leq 0 \quad \text{and} \quad \psi_6(\boldsymbol{x}, \boldsymbol{s}) \leq -10,$$

and

$$\psi_{12}(\boldsymbol{x}, \boldsymbol{s}) \leq 0 \quad \text{and} \quad \psi_{12}(\boldsymbol{x}, \boldsymbol{s}) \leq -10;$$

respectively in int $\mathcal{F}_p$ (on a plane).

Hint: To plot the $\boldsymbol{x}$ part of the level set of the potential function, say $\psi_6(\boldsymbol{x}, \boldsymbol{s}) \leq 0$, you plot

$$\{\boldsymbol{x} \in \text{int } \mathcal{F}_p : \min_{\boldsymbol{s}\in\text{int } \mathcal{F}_d} \psi_6(\boldsymbol{x}, \boldsymbol{s}) \leq 0\}.$$

This can be approximately done by sampling as follows. You randomly generate N primal points $\{\boldsymbol{x}^p\}_{p=1}^N$ from int $\mathcal{F}_p$, and N primal points of $\{\boldsymbol{s}^q\}_{q=1}^N$ from int $\mathcal{F}_d$. For each primal point $\boldsymbol{x}^p$, you find if it is true that

$$\min_{q=1,\dots,N} \psi_6(\boldsymbol{x}^p, \boldsymbol{s}^q) \le 0.$$

Then, you plot those $\boldsymbol{x}^p$ who give an "yes" answer.

Solution The Julia code is online here.

First, by sampling, it is very hard to plot $\{\psi_6(\boldsymbol{x}, \boldsymbol{s}) \le -10\}$, because here $\boldsymbol{s} = (1+y, 1+y, y) > 0$, so we need $y > 0$. But $\psi_6(\boldsymbol{x}, \boldsymbol{s}) \ge 3\log(\boldsymbol{x}^\top \boldsymbol{s}) + 3\log 3 = 3\log(x_1 + x_2 + y) + 3\log 3$. Hence $\{\psi_6(\boldsymbol{x}, \boldsymbol{s}) \le -10\}$ is too harsh for sampled points to survive. Notice that when $n + \rho$ is larger, more primal points survive, and when we look at lower level set $\{\psi \le -10\}$, even though fewer points survive, but they converge to the optimal solution (as we lower the level set continuously). See Figure 5 and 6 for the sublevel sets of Part I and Part II in Problem 5.

Here is how we do the analysis: we sample 10000 feasible $\boldsymbol{x}$ uniformly in the $\mathcal{F}_p$ (the unit simplex), we sample 10000 feasible $\boldsymbol{s}$, where $\boldsymbol{s} = [1-y, 1-y, -y]$, and for $\boldsymbol{s}$ to > 0, we sample $y = -2\texttt{rand}(1)$. Then we follow the determine rule in hint, and analyze whether $\min_{q=1,\dots,N} \psi_6(\boldsymbol{x}^p, \boldsymbol{s}^q) \le 0$ or not.

Remark. Alternatively, you can use optimization solvers like MOSEK, or fmincon.m in MATLAB, to solve the feasibility problems directly by looping over a grid of $\boldsymbol{x}$ (or uniformly sampling $\boldsymbol{x}$) and solving the partial feasibility problem in terms of $\boldsymbol{s}$. If the solver returns infeasibility, then $\boldsymbol{x}$ is not feasible. Otherwise, $\boldsymbol{x}$ is feasible. Similarly, for any sampled/chosen $\boldsymbol{x}$ that needs to be checked, we can simply minimize over $\boldsymbol{s}$ and conclude that $\boldsymbol{x}$ is feasible iff the optimal value of $\psi_{n+\rho}(\boldsymbol{x}, \cdot)$ is non-positive.

11. (10') Recall the Fisher's Equilibrium prices problem (discussed in Lecture Note 6), which we describe here again for reference. Let B be the set of buyers and G be the set of goods. Each buyer $i \in B$ has a budget $\boldsymbol{w}_i > 0$, and utility coefficients $u_{ij} \ge 0$ for each good $j \in G$. Under price $\boldsymbol{p}$, buyer $i \in B$'s optimal purchase quantity $\boldsymbol{x}_i^*(\boldsymbol{p})$ is the solution of the following optimization problem:

$$\begin{aligned} \boldsymbol{x}_i^*(\boldsymbol{p}) \in \arg\max \quad & \boldsymbol{u}_i^\top \boldsymbol{x}_i := \sum_{j \in G} u_{ij} x_{ij} \\ \text{s.t.} \quad & \boldsymbol{p}^\top \boldsymbol{x}_i := \sum_{j \in G} p_j x_{ij} \le \boldsymbol{w}_i, \\ & \boldsymbol{x}_i \ge \boldsymbol{0} \end{aligned}$$

Figure 5: Sublevel sets of $\psi_{n+\rho}(\boldsymbol{x}, \boldsymbol{s})$, $\boldsymbol{x}$ part (Problem 5 Part I)

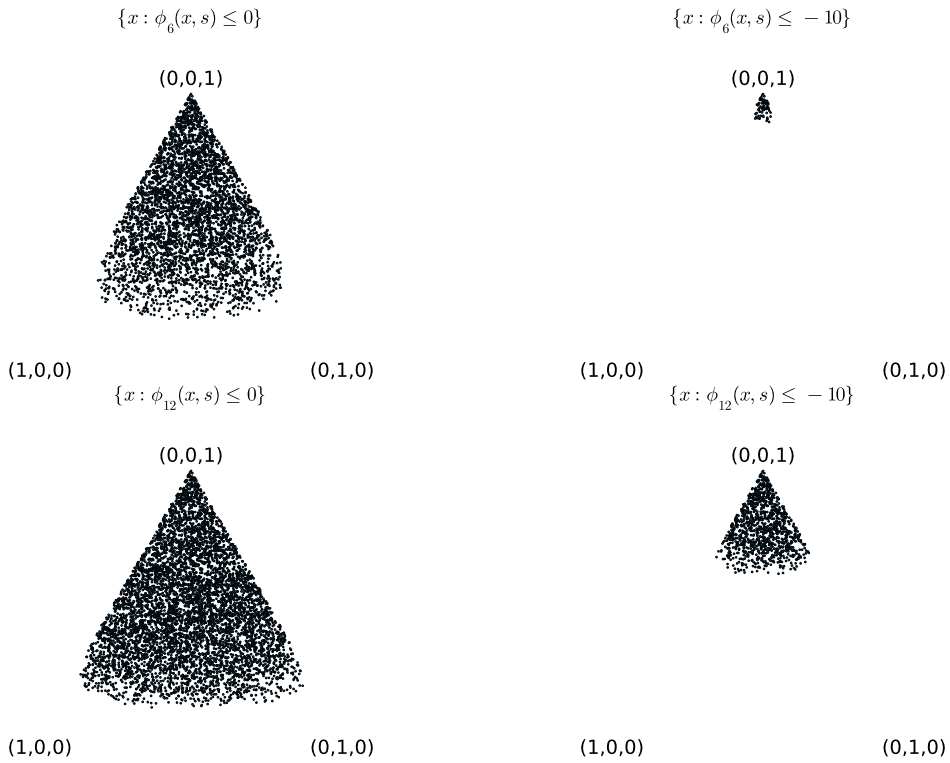
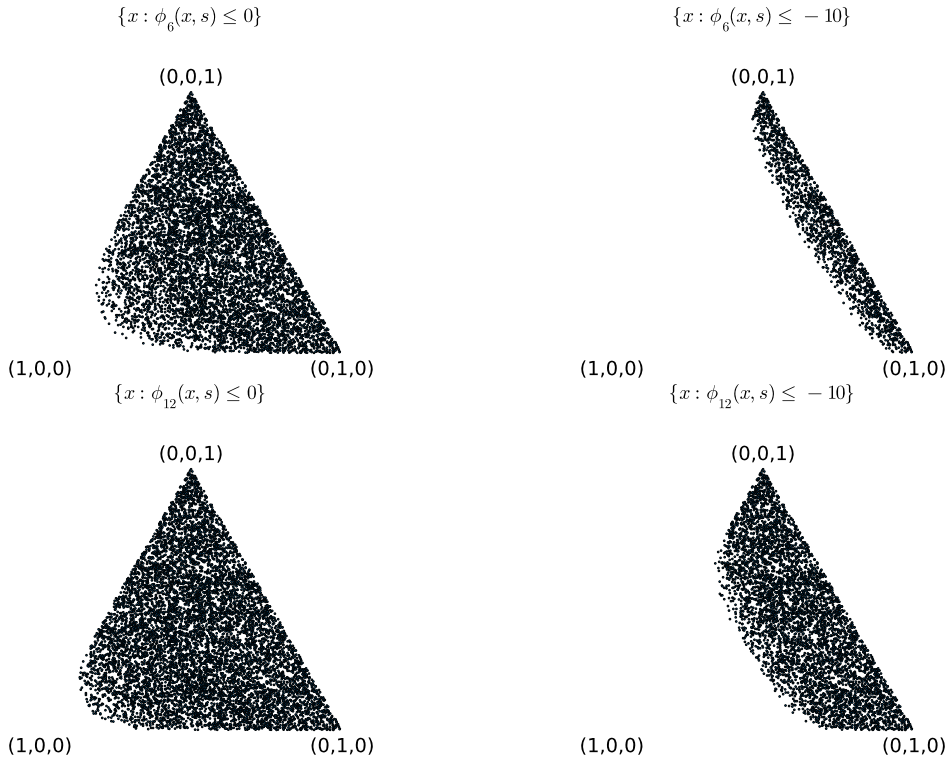


Figure 6: Sublevel sets of $\psi_{n+\rho}(\boldsymbol{x}, \boldsymbol{s})$, $\boldsymbol{x}$ part (Problem 5 Part I)



Suppose each good $j \in G$ has a supply level $\bar{s}_j$. We call a price vector $\boldsymbol{p}^*$ an **equilibrium price vector** if the market clears, namely for all $j \in G$,

$$\sum_{i \in B} x_{ij}^*(\boldsymbol{p}^*) = \bar{s}_j.$$

In the lecture, we discussed how to compute the equilibrium price $\boldsymbol{p}^*$ and buyers' activities $\{\boldsymbol{x}_i^*(\boldsymbol{p}^*)\}_{i \in B}$ under the equilibrium price based on utility coefficients $\{\boldsymbol{u}_i\}_{i \in B}$, budgets $\{\boldsymbol{w}_i\}_{i \in B}$ and supplies $\bar{\boldsymbol{s}}$:

$$(\{\boldsymbol{u}_i\}_{i \in B}, \{\boldsymbol{w}_i\}_{i \in B}, \bar{\boldsymbol{s}}) \Rightarrow (\boldsymbol{p}^*, \{\boldsymbol{x}_i^*(\boldsymbol{p}^*)\}_{i \in B}) \tag{13}$$

In this question, we consider the inverse problem of (13): suppose the market does not know the "private information" of each buyer, namely the utility $\{\boldsymbol{u}_i\}_{i \in B}$ and the budgets $\{\boldsymbol{w}_i\}_{i \in B}$, but instead you observe the equilibrium prices $\{\boldsymbol{p}^{*(k)}\}_{k=1}^K$ and their corresponding realized activities $\{\boldsymbol{x}_i^{*(k)}\}_{k=1}^K$ under K different supply levels $\bar{\boldsymbol{s}}^{(1)}, \ldots, \bar{\boldsymbol{s}}^{(K)}$. The query is to infer buyers' utility coefficients $\{\boldsymbol{u}_i\}_{i \in B}$ and their budgets $\{\boldsymbol{w}_i\}_{i \in B}$. We assume that the utility function is ℓ_1-normalized, namely $\|\boldsymbol{u}_i\|_1 = 1$ for $i \in B$.

Hint: Mathematically, the query is to find $\{\boldsymbol{u}_i\}_{i \in B}$ (s.t. $\boldsymbol{u}_i \geq \boldsymbol{0}$ and $\|\boldsymbol{u}_i\|_1 = 1$) and $\{\boldsymbol{w}_i\}_{i \in B}$ (s.t. $\boldsymbol{w}_i > 0$) such that for all $i \in B$, and $k = 1, \ldots, K$,

$$\begin{aligned} \boldsymbol{x}_i^{*(k)} = \arg\max_{\boldsymbol{x}_i} \quad & \boldsymbol{u}_i^\top \boldsymbol{x}_i \\ \text{s.t.} \quad & (\boldsymbol{p}^{*(k)})^\top \boldsymbol{x}_i \leq \boldsymbol{w}_i \\ & \boldsymbol{x}_i \geq \boldsymbol{0} \end{aligned}$$

given $\{\boldsymbol{x}_i^{*(k)}\}_{i \in B, k \in \{1,\ldots,K\}}$ and $\{\boldsymbol{p}^{*(k)}\}_{k \in \{1,\ldots,K\}}$.

Question: Now consider the following 2-buyer 2-good example and solve this inverse problem. Let $B = \{1, 2\}$ and $G = \{1, 2\}$. Suppose we observe the following 5 scenarios:

- $\boldsymbol{p}^{*(1)} = (\frac{9}{5}; \frac{3}{5})$, $\boldsymbol{x}_1^{*(1)} = (1; \frac{1}{3})$, $\boldsymbol{x}_2^{*(1)} = (0; \frac{5}{3})$;
- $\boldsymbol{p}^{*(2)} = (2; 1)$, $\boldsymbol{x}_1^{*(2)} = (1; 0)$, $\boldsymbol{x}_2^{*(2)} = (0; 1)$;
- $\boldsymbol{p}^{*(3)} = (1; 1)$, $\boldsymbol{x}_1^{*(3)} = (2; 0)$, $\boldsymbol{x}_2^{*(3)} = (0; 1)$;
- $\boldsymbol{p}^{*(4)} = (\frac{1}{2}; 1)$, $\boldsymbol{x}_1^{*(4)} = (4; 0)$, $\boldsymbol{x}_2^{*(4)} = (0; 1)$;
- $\boldsymbol{p}^{*(5)} = (\frac{3}{7}; \frac{6}{7})$, $\boldsymbol{x}_1^{*(5)} = (\frac{14}{3}; 0)$, $\boldsymbol{x}_2^{*(5)} = (\frac{1}{3}; 1)$.

Use any approach to find $\{\boldsymbol{u}_i\}_{i\in B}$ (s.t. $\boldsymbol{u}_i \geq \mathbf{0}$ and $\|\boldsymbol{u}_i\|_1 = 1$) and $\{\boldsymbol{w}_i\}_{i\in B}$ (s.t. $\boldsymbol{w}_i > 0$). Describe your approach and report the result.

Solution Solve the system of KKT conditions:

$$\begin{aligned}
&\boldsymbol{p}^{*(k)\top}\boldsymbol{x}_i^{*(k)} \leq \boldsymbol{w}_i \\
&\boldsymbol{x}_i^{*(k)} \geq \mathbf{0} \\
&\boldsymbol{u}_i \leq q_i^{*(k)}\boldsymbol{p}^{*(k)} \\
&q_i^{*(k)} \geq 0 \\
&\boldsymbol{x}_i^{*(k)} \cdot (q_i^{*(k)}\boldsymbol{p}^{*(k)} - \boldsymbol{u}_i) = \mathbf{0} \\
&q_i^{*(k)} \cdot (\boldsymbol{w}_i - \boldsymbol{p}^{*(k)\top}\boldsymbol{x}_i^{*(k)}) = 0
\end{aligned} \qquad , \forall k \in [K], i \in B$$

together with

$$\boldsymbol{u}_i \geq \mathbf{0}, \boldsymbol{e}^\top \boldsymbol{u}_i = 1, \boldsymbol{w}_i > 0$$

for $i \in B$.

In fact, the first inequality should be binding at optimal solutions, *i.e.*,, $\boldsymbol{p}^{*(k)\top}\boldsymbol{x}_i^{*(k)} = \boldsymbol{w}_i$ (Because the buyers should consume all the budget to obtain maximal utility). Hence we have $w_1 = 2$, $w_2 = 1$. Then we use complementary slackness to obtain $\boldsymbol{u}_1 = (3/4; 1/4)$, $\boldsymbol{u}_2 = (1/3; 2/3)$ by observing $\boldsymbol{x}_1^{*(1)} > \mathbf{0}$ and $\boldsymbol{x}_2^{*(5)} > \mathbf{0}$.

4 Midterm Exam

1. **Question 1** (30') [True/False] Give a *true* or *false* answer to each of the following questions and explain your choice. (If your answer is *true*, provide an argument or cite the appropriate claims from lecture notes or textbook. If your answer is *false*, provide a counterexample or cite the appropriate claims from lecture notes or textbook).

(a) (5') In classical linear programming, even if the problem is feasible and bounded, the optimal solution might not be attained.

Solution False. See page 5 in Lecture Note 4. If the problem is feasible and bounded, its dual is also feasible and bounded. Thus, both primal and dual problems have optimal solutions.

(b) (5') In classical linear programming, strong duality always holds if both of the primal and the dual problems are feasible. In contrast, in semi-definite programming, there are feasible problem pairs where strong duality does not hold.

Solution True. For a couter-example, please see page 13 in Lecture Note 4.

(c) (5') In a convex constrained optimization problem, if $\boldsymbol{x}$ is a local minimizer, then it is a KKT solution.

Solution False. For a counter-example, see Example 6 of Chapter 6.4 in [LY21], more examples are available in Section 2.3 in [DW17]. Any local minimizer is a KKT point if it is a regular point even if it is a convex optimization problem.

(d) (5') Consider a set C defined by $C := \{\boldsymbol{x} = (x_1, ..., x_n)^\top \in \mathbb{R}^n_+ : f(\boldsymbol{x}) \geq 0\}$, where $n \geq 2$. If $f(\boldsymbol{x}) = \prod_{i=1}^n x_i - 1$, then C is a convex set even $f(\boldsymbol{x})$ is a non-concave function on C.

Solution True. To verify C is a convex set, we write it in an equivalent form

$$C = \{\boldsymbol{x} : -\sum_{i=1}^n \log(x_i) \leq 0\},$$

which implies C is convex. To see $f(\boldsymbol{x})$ is not a concave function, directly check the Hessian matrix when $n = 2$. In this case, $f(\boldsymbol{x})$ is neither convex nor concave.

(e) (5') Consider a conic LP in the standard form

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{a}_i^\top \boldsymbol{x} = b_i, \ i = 1, ..., m, \ (\mathcal{A}\boldsymbol{x} = \boldsymbol{b}) \\ & \boldsymbol{x} \in K. \end{aligned}$$

Similar to the Lagrangian function, we can construct

$$L(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{s}) = \boldsymbol{c}^\top \boldsymbol{x} - \boldsymbol{y}^\top (\mathcal{A}\boldsymbol{x} - \boldsymbol{b}) - \boldsymbol{s}^\top \boldsymbol{x},$$

where $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{s} \in K^*$. Let

$$\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{s}) = \inf_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{s}).$$

Then, the conic dual problem is equivalent to the lagrangian dual problem, that is, $\max_{\boldsymbol{y}, \boldsymbol{s} \in K^*} \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{s})$.

Solution True. If $\boldsymbol{c} - \boldsymbol{s} - \mathcal{A}^\top \boldsymbol{y} \neq \boldsymbol{0}$, we have $\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{s}) = -\infty$ by taking $\boldsymbol{x} = \alpha(\boldsymbol{c} - \boldsymbol{s} - \mathcal{A}^\top \boldsymbol{y})$ and letting $\alpha \to -\infty$. Otherwise, we have

$$L(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{s}) = \boldsymbol{b}^\top \boldsymbol{y}.$$

Thus, the Lagrangian dual problem is eqivalent to

$$\max_{\boldsymbol{c} = \boldsymbol{s} + \mathcal{A}^\top \boldsymbol{y}, \boldsymbol{s} \in K^*} \boldsymbol{b}^\top \boldsymbol{y},$$

which is the conic dual problem.

(f) (5') Consider a feasible optimization problem on $\mathbb{R}^3$,

$$\begin{aligned} \min \quad & f(x_1, x_2, x_3) \\ \text{s.t.} \quad & c(x_1, x_2, x_3) \leq 0, \end{aligned}$$

where $f(\boldsymbol{x})$ and $c_i(\boldsymbol{x})$ are strongly convex functions.[2] Assume that $f(x_1, x_2, x_3) = f(x_1, x_3, x_2)$ and $c(x_1, x_2, x_3) = c(x_1, x_3, x_2)$, for any $x_1, x_2, x_3 \in \mathbb{R}$. Assume this problem has a minimizer. Then we must have $x_2 = x_3$ at the unique minimizer.

Solution TRUE. Otherwise we would have more than one minimizer, which is impossible because the function is strongly convex.

[2] A strongly convex function f satisfies $f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) < \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y})$ for distinct x, y and $\alpha \in (0, 1)$

2. **Question 2** [SVM] (20')

Recall the Supporting Vector Machine in Lecture 1 and question 6 in Homework 2. Let the red class of points contain three points

$$\boldsymbol{a}_1 = (0, 3), \boldsymbol{a}_2 = (1, 2), \boldsymbol{a}_3 = (2, 3),$$

and let the blue class of points contain three points

$$\boldsymbol{b}_1 = (1, 0), \boldsymbol{b}_2 = (2, 1), \boldsymbol{b}_3 = (1, 3),$$

which are illustrated in Figure 7.

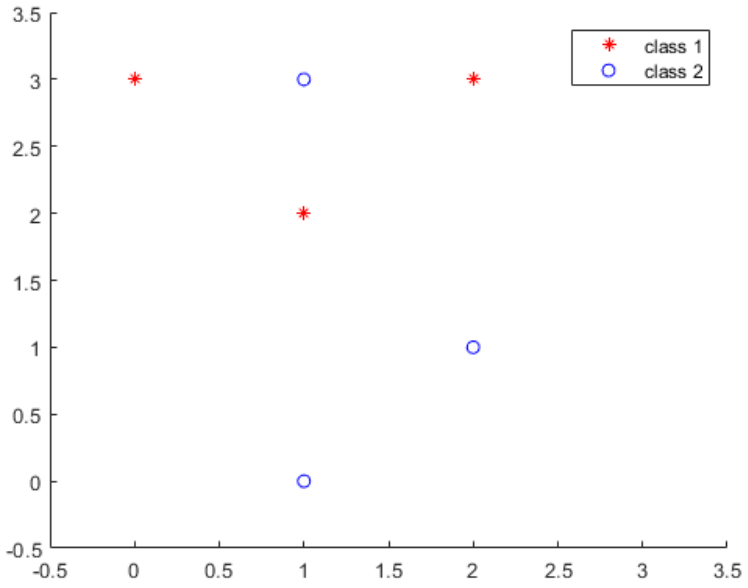


Figure 7: Scatter of all points

(a) (4') Consider the "hard-margin" SVM problem

$$\begin{aligned} \min\ & \|\boldsymbol{x}\|^2 \\ \text{s.t. } & \boldsymbol{a}_i^\top \boldsymbol{x} + x_0 \geq 1, i = 1, 2, 3 \\ & \boldsymbol{b}_j^\top \boldsymbol{x} + x_0 \leq -1, j = 1, 2, 3. \end{aligned}$$

Is this problem feasible (i.e. there is a line that can separate the two classes of points)? Explain why.

Solution It is infeasible. $\boldsymbol{b}_1^\top \boldsymbol{x} + x_0 \leq -1$ and $\boldsymbol{a}_2^\top \boldsymbol{x} + x_0 \geq 1$ implies $\boldsymbol{x}(2) \geq 0$. However, it contradicts with $\boldsymbol{b}_3^\top \boldsymbol{x} + x_0 \leq -1$.

(b) (6') Consider the following SVM problem with additional variable β (which was described in class):

$$\begin{aligned} \min\ & \beta + \|\boldsymbol{x}\|^2 \\ \text{s.t. } & \boldsymbol{a}_i^\top \boldsymbol{x} + x_0 + \beta \geq 1, i = 1, 2, 3 \\ & \boldsymbol{b}_j^\top \boldsymbol{x} + x_0 - \beta \leq -1, j = 1, 2, 3 \\ & \beta \geq 0. \end{aligned}$$

Show that this problem is feasible and find an optimal solution by hand. (Hint: you can try to show $\beta \geq 1$ at first, and then fix $\beta = 1$.)

Solution

Plugging in $\boldsymbol{a}_2 = (1, 2)$, $\boldsymbol{b}_1 = (1, 0)$, $\boldsymbol{b}_3 = (1, 3)$, we have

$$\begin{aligned} x_1 + 2x_2 + x_0 + \beta &\geq 1 \\ x_1 + x_0 - \beta &\leq -1 \\ x_1 + 3x_2 + x_0 - \beta &\leq -1 \end{aligned}$$

The first two inequalities imply that $2x_2 + 2\beta \geq 2$, while the first and the third inequalities imply that $-x_2 + 2\beta \geq 2$. It follows that $\beta \geq 1$ so that the optimal objective value is bounded below by 1. On the other hand, setting $\beta = 1$, $\boldsymbol{x} = 0$ and $x_0 = 0$ we see that this is a feasible solution. It follows that the optimal value is 1.

(c) (6') Consider the following "soft-margin" SVM problem with six additional variable β's:

$$\begin{aligned} \min\ & \frac{1}{6}(\beta_1^a + \beta_2^a + \beta_3^a + \beta_1^b + \beta_2^b + \beta_3^b) + \|\boldsymbol{x}\|^2 \\ \text{s.t. } & \boldsymbol{a}_i^\top \boldsymbol{x} + x_0 + \beta_i^a \geq 1, i = 1, 2, 3 \\ & \boldsymbol{b}_j^\top \boldsymbol{x} + x_0 - \beta_j^b \leq -1, j = 1, 2, 3 \\ & \beta_i^a, \beta_j^b \geq 0,\ i, j = 1, 2, 3. \end{aligned}$$

Please construct the dual of this "soft-margin" SVM problem. In this case, one optimal solution is

$$\begin{aligned} & \boldsymbol{x} = (-0.08, 0.33)^\top,\ x_0 = -0.50, \\ & \beta_1^a = 0.51,\ \beta_2^a = 0.92,\ \beta_3^a = 0.67, \\ & \beta_1^a = 0.41,\ \beta_2^a = 0.66,\ \beta_3^b = 1.41. \end{aligned}$$

Solution With the Lagrangian function, we can have the dual problem is

$$\begin{aligned}
\max \quad & -\frac{1}{4}\|\sum_{i=1}^{3} y_i^a \boldsymbol{a}_i + \sum_{j=1}^{3} y_j^b \boldsymbol{b}_j\|^2 + \sum_{i=1}^{3} y_i^a - \sum_{j=1}^{3} y_j^b \\
\text{s.t.} \quad & \sum_{i=1}^{3} y_i^a + \sum_{j=1}^{3} y_j^b = 0 \\
& \frac{1}{6} - y_i^a - s_i^a = 0, \text{ for } i = 1, 2, 3 \\
& \frac{1}{6} + y_j^b - s_j^b = 0, \text{ for } j = 1, 2, 3 \\
& y_i^a, s_i^a, s_j^b \geq 0 \; y_j^b \leq 0, \text{ for } i, j = 1, 2, 3
\end{aligned}$$

The soft-margin classifier is illustrated in Figure 8.

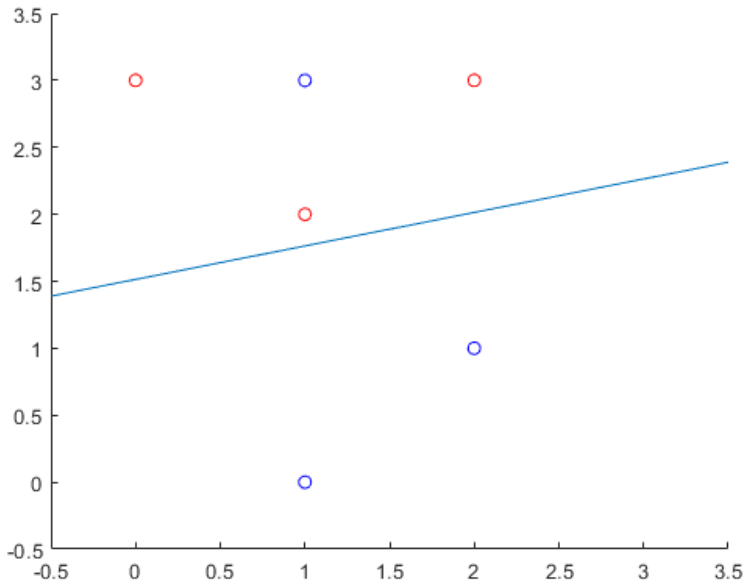


Figure 8: Scatter of all points and classifier - soft margins

(d) (4') Compare the results of part (a), (b) and (c). Explain which model would you likely to choose to use for this case and why. What about in general? (There is no right or wrong answer to this question.)

Solution I would choose the one in part (c). Moreover, the third SVM is more robust to outliers. But any reasonable answer with SVMs in parts (a) and (b) is OK for this question.

3. **Question 3** [Resource Allocation] (20')

A function $f : R^n \longrightarrow R$ is said to be *separable* if it can be written in the form

$$f(x_1, \ldots, x_n) = \sum_{j=1}^{n} f_j(x_j).$$

Such functions were studied by Gibbs in connection with work on the chemical equilibrium problem (1876). He showed a theorem which states that (in the differentiable case) a necessary condition of local optimality for a feasible point x^* of the separable nonlinear programming problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \sum_{j=1}^{n} f_j(x_j) \\ \text{s.t.} \quad & \sum_{j=1}^{n} x_j = M \quad (M > 0) \\ & x_j \geq 0 \quad j = 1, \ldots, n \end{aligned}$$

is that there exist a thresholding number λ^* such that

$$f_j'(x_j^*) = \lambda^* \quad \text{if } x_j^* > 0$$

$$f_j'(x_j^*) \geq \lambda^* \quad \text{if } x_j^* = 0$$

Here f' is the derivative of f, and given M represents the total amount of the single resource. This constraint is called "Knapsack" constraint.

(a) (5') Show that every minimizer is a KKT solution. (Hint: at lease one of variable must be positive at any feasible solution.)

Solution For any feasible solution $\boldsymbol{x} = (x_1, ..., x_n)^\top$, denote

$$A = \{j = 1, .., n : x_j = 0\}.$$

We have $|A| \leq n - 1$ since the condition $M > 0$ implies at least one of x_j is non-zero. Then, the hypersurface of active constraints is

$$\left\{ \sum_{j=1}^{n} x_j = M; x_j = 0, j \in A \right\}.$$

To check the regularity, we compute the derivative of constraint functions of active constraints, and have $\nabla(x_j) = \boldsymbol{e}_j$ and $\nabla(\sum_j x_j - M) = \boldsymbol{e}$, where $\boldsymbol{e}_j$ is the vector

with value 1 in its j-th entry and 0 otherwise and $\mathbf{1}$ is an all-one vector. Thus, all feasible solution are regular points since it is easy to check $\{\boldsymbol{e}, \boldsymbol{e}_j, j \in A\}$ is a set of linearly-independent vectors.

(b) (5') Why is this theorem true? What does λ^* represent?

Solution The Lagrangian of the problem is given by

$$L(\boldsymbol{x}, \lambda, \boldsymbol{\mu}) = \sum_i f_i(x_i) + \lambda(M - \sum_i x_i) - \sum_i \mu_i x_i$$

and the first order conditions will be

$$\frac{\partial L}{\partial x_i} = f_i'(x_i) - \lambda - \mu_i = 0$$
$$\lambda(M - \sum x_i) = 0$$
$$\mu_i x_i = 0$$
$$\mu_i \geq 0$$

From the first and last equation we get

$$f_i'(x_i) = \lambda + \mu_i \geq \lambda.$$

λ^* represents the rate of optimal objective change over the change of right-hand-side resource M. Note that if $x_i > 0$ then $\mu_i = 0$, which implies

$$f_i'(x_i) = \lambda$$

In this case, λ^* corresponds to the marginal gain from increasing each one of the f_i when $x_i^* > 0$, that is, the optimal allocation of x_i is at the point such that every function f_i has the same marginal gain λ^*.

(c) (5') Show that

$$\lambda^* = \frac{1}{M} \sum_j (x_j^* f_j'(x_j^*)).$$

Solution Multiply the equation $\frac{\partial L}{\partial x_i} = 0$ by x_i to get

$$f_i'(x_i)x_i - \lambda x_i - \mu_i x_i = 0$$

Replacing $\mu_i x_i = 0$ and summing over i

$$\sum_i f_i'(x_i)x_i = \lambda \sum_i x_i$$

And using primal feasibility condition $\sum_i x_i = M$ we get

$$\sum_i f_i'(x_i)x_i = \lambda M$$

which gives the result.

(d) (5') What is λ^* if

$$f_j(x_j) = -w_j \log(x_j),\ \forall j,$$

where w_j is a given positive constant (the minimizer of the problem is called the weighted analytic center)?

Solution In this case $f_i'(x_i)x_i = -w_i$, so

$$\lambda^* = -\frac{1}{M}\sum_i w_i.$$

4. **Question 4** [Small SNL] (20')

In this problem, we will solve the SDP relaxation of a small SNL problem by hand. Consider a sensor network localization problem with three anchors:

$$\boldsymbol{a}_1 = (1; 0),$$
$$\boldsymbol{a}_2 = (0; 1),$$
$$\boldsymbol{a}_3 = (-1; 0).$$

The sensor's true location is

$$\boldsymbol{x} = (1; 1.5).$$

However, the sensor's true location is unknown to the problem solver. In Figure 9, the three red points denote anchors and the blue point denotes the true location of the sensor.

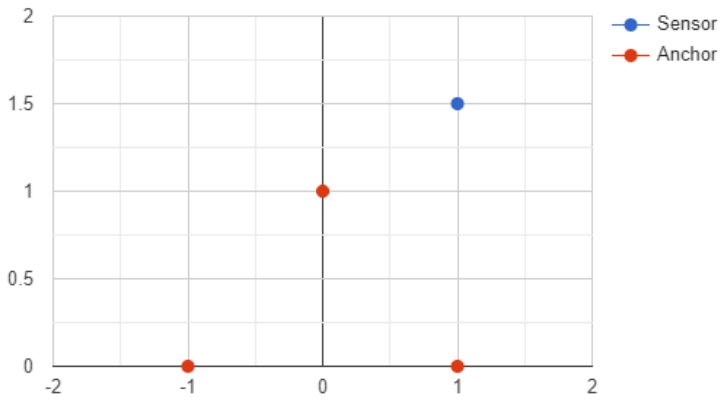


Figure 9: Location of anchors and the sensor

(a) (5') Write out the SDP relaxation problem for localizing the single sensor with null objective.

Solution We first compute the distance between the sensor, and anchors and have

$$d_1^2 = \|\boldsymbol{a}_1 - \boldsymbol{x}\|^2 = 2.25,$$
$$d_2^2 = \|\boldsymbol{a}_2 - \boldsymbol{x}\|^2 = 1.25,$$
$$d_3^2 = \|\boldsymbol{a}_3 - \boldsymbol{x}\|^2 = 6.25.$$

Then, the SDP relaxation problem is

$$\begin{aligned}
\max_{\boldsymbol{Z}} \quad & \mathbf{0} \bullet \boldsymbol{Z} \\
\text{s.t.} \quad & (1;0;0)(1;0;0)^\top \bullet \boldsymbol{Z} = 1 \\
& (0;1;0)(0;1;0)^\top \bullet \boldsymbol{Z} = 1 \\
& (1;1;0)(1;1;0)^\top \bullet \boldsymbol{Z} = 2 \\
& (\boldsymbol{a}_k;-1)(\boldsymbol{a}_k;-1)^\top \bullet \boldsymbol{Z} = d_k^2 \ , \ \forall k = 1,2,3 \\
& \boldsymbol{Z} \succeq \mathbf{0}.
\end{aligned}$$

(b) (5') Write out the dual of the SDP relaxation problem.

Solution For the primal problem,

$$\begin{aligned}
\max_{\boldsymbol{Z}} \quad & \mathbf{0} \bullet \boldsymbol{Z} \\
\text{s.t.} \quad & (1;0;0)(1;0;0)^\top \bullet \boldsymbol{Z} = 1 \ (w_1) \\
& (0;1;0)(0;1;0)^\top \bullet \boldsymbol{Z} = 1 \ (w_2) \\
& (1;1;0)(1;1;0)^\top \bullet \boldsymbol{Z} = 2 \ (w_3) \\
& (\boldsymbol{a}_k;-1)(\boldsymbol{a}_k;-1)^\top \bullet \boldsymbol{Z} = d_k^2 \ (\lambda_k), \ \text{for } k = 1,2,3 \\
& \boldsymbol{Z} \succeq \mathbf{0}.
\end{aligned}$$

The dual problem is

$$\begin{aligned}
\min \quad & w_1 + w_2 + 2w_3 + \sum_{k=1}^{3} \lambda_k d_k^2 \\
\text{s.t.} \quad & \begin{pmatrix} \begin{pmatrix} w_1 + w_3 & w_3 \\ w_3 & w_2 + w_3 \end{pmatrix} + \sum_{k=1}^{3} \lambda_k \boldsymbol{a}_k \boldsymbol{a}_k^\top & -\sum_{k=1}^{3} \lambda_k \boldsymbol{a}_k \\ -\sum_{k=1}^{3} \lambda_k \boldsymbol{a}_k^\top & \sum_{k=1}^{3} \lambda_k \end{pmatrix} \succeq \mathbf{0}.
\end{aligned}$$

(c) (5') Write out the SDP solution explicitly constructed from the true position (Lecture Note 5) and verify that it is an optimal solution to the SDP relaxation problem.

Solution In this case,

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1.5 \\ 1 & 1.5 & 3.25 \end{pmatrix}$$

and it is easy to verify that $\boldsymbol{Z}$ is a feasible solution of the primal SDP. Since the objective is always 0, it is also an optimal solution.

(d) (5') Verify that the dual has a rank-one optimal solution so that the SDP solution in (c) is the only optimal solution.

Solution We can find a rank-1 dual solution such that

$$\begin{pmatrix} \begin{pmatrix} w_1 + w_3 & w_3 \\ w_3 & w_2 + w_3 \end{pmatrix} + \sum_{k=1}^{3} \lambda_k \boldsymbol{a}_k \boldsymbol{a}_k^\top & -\sum_{k=1}^{3} \lambda_k \boldsymbol{a}_k \\ -\sum_{k=1}^{3} \lambda_k \boldsymbol{a}_k^\top & \sum_{k=1}^{3} \lambda_k \end{pmatrix} = (-\boldsymbol{x};1)(-\boldsymbol{x};1)^\top.$$

Thus, any solution to the SDP relaxation is at most rank 2. Moreover, the primal problem implies that all feasible solution is at least rank 2. Thus, all primal solutions have the same rank, and the primal solution is unique based on Theorem 3 in Lecture Note 5 (To show the uniqueness, we need to verify the linear independence.)

5. **Question 5** (10'+5'(bonus))

   Consider the primal feasible region in standard form $\{\boldsymbol{x} \in \mathbb{R}^n_+ : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$.

(a) (5') A variable x_i is said to be a *null variable* if $x_i = 0$ in every feasible solution. Prove that, if the feasible region is non-empty, x_i is a null variable if and only if there is a nonzero vector $\boldsymbol{y} \in \mathbb{R}^m$ such that $\boldsymbol{y}^\top \boldsymbol{A} \geq \boldsymbol{0}$, $\boldsymbol{y}^\top \boldsymbol{b} = 0$. and the ith component of $\boldsymbol{y}^\top \boldsymbol{A}$ is strictly positive.

Solution If the feasible region is non-empty, the following LP is feasible and bounded.

$$\begin{aligned} \max \quad & x_i \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0} \end{aligned}$$

Moreover, x_i is a null variable iff the primal optimal objective value is 0. This is equivalent to the dual problem

$$\begin{aligned} \min \quad & \boldsymbol{b}^\top \boldsymbol{y} \\ \text{s.t.} \quad & \boldsymbol{y}^\top \boldsymbol{A} \geq \boldsymbol{e}_i^\top \end{aligned}$$

has optimal value 0 by the Duality Theorem of Linear Programming, *i.e.*, there exists some $\boldsymbol{y}$ such that $\boldsymbol{y}^\top \boldsymbol{A} \geq \boldsymbol{e}_i^\top$ and $\boldsymbol{y}^\top \boldsymbol{b} = 0$, which is equivalent to the condition in this question.

(b) (5') [*Strict complementarity*] Let the feasible region be nonempty. Then there is a feasible $\boldsymbol{x}$ and vector $\boldsymbol{y} \in \mathbb{R}^m$ such that

$$\boldsymbol{y}^\top \boldsymbol{A} \geq \boldsymbol{0},\ \boldsymbol{y}^\top \boldsymbol{b} = 0,\ \boldsymbol{A}^\top \boldsymbol{y} + \boldsymbol{x} > \boldsymbol{0}.$$

Solution If the feasible region is nonempty, by Part (a) we know for each $i = 1, .., n$, there exist $\boldsymbol{x}^i \in \mathbb{R}^n, \boldsymbol{y}^i \in \mathbb{R}^m$ such that

$$\boldsymbol{A}\boldsymbol{x}^i = \boldsymbol{b}, \boldsymbol{x}^i \geq \boldsymbol{0}, \boldsymbol{A}^\top \boldsymbol{y}^i \geq \boldsymbol{0}, \boldsymbol{b}^\top \boldsymbol{y}^i = 0,$$

and either $\boldsymbol{e}_i^\top \boldsymbol{x}^i > 0$ or $\boldsymbol{e}_i^\top \boldsymbol{A}^\top \boldsymbol{y}^i > 0$, *i.e.*, $\boldsymbol{e}_i^\top (\boldsymbol{x}^i + \boldsymbol{A}^\top \boldsymbol{y}^i) > 0$.

Let $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}^i$ and $\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{y}^i$, we have

$$\boldsymbol{A}\bar{\boldsymbol{x}} = \boldsymbol{b}, \bar{\boldsymbol{x}} \geq \boldsymbol{0}, \boldsymbol{A}^\top \bar{\boldsymbol{y}} \geq \boldsymbol{0}, \boldsymbol{b}^\top \bar{\boldsymbol{y}} = 0,$$

and $\boldsymbol{e}_i^\top (\bar{\boldsymbol{x}} + \boldsymbol{A}^\top \bar{\boldsymbol{y}}) > 0$ for each $i = 1, ..., n$, *i.e.*, $\bar{\boldsymbol{x}} + \boldsymbol{A}^\top \bar{\boldsymbol{y}} > \boldsymbol{0}$.

(c) (5') (Bonus) A variable x_i is a *nonextremal variable* if $x_i > 0$ in every feasible solution. Prove that, if the feasible region is non-empty, x_i is a nonextremal variable if and only if there is $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{d} \in \mathbb{R}^n$ such that $\boldsymbol{y}^\top \boldsymbol{A} = \mathbf{d}^\top$, where $d_i = -1,\ d_j \geqslant 0$ for $j \neq i$; and such that $\boldsymbol{y}^\top \boldsymbol{b} < 0$.

Solution If the feasible region is non-empty, the following LP is feasible and bounded.

$$\begin{aligned} \min \quad & x_i \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0} \end{aligned}$$

It implies that x_i is a nonextremal variable iff the optimal objective value of the above primal problem is strictly positive. Thus, x_i is a nonextremal variable iff there is one dual feasible solution such that the objective value is strictly positive by the Duality Theorem of Linear Programming. That is, there exists a dual feasible solution such that

$$\boldsymbol{b}^\top \boldsymbol{y} > 0,\ \boldsymbol{y}^\top \boldsymbol{A}^\top \leq \boldsymbol{e}_i^\top \tag{14}$$

which is equivalent to existence of $\boldsymbol{y}$ and $\boldsymbol{d}$ such that

$$\boldsymbol{y}^\top \boldsymbol{b} < 0, \boldsymbol{y}^\top \boldsymbol{A} = \boldsymbol{d}^\top, d_i = -1, d_j \geq 0, \forall j \neq i.$$

6. **Question 6** (20')

We have studied the Fisher market equilibrium problem where there are m goods in the market and each good j has a fixed amount $\bar{s}_j (> 0)$ available. There are n buyers in the market where each buyer, say buyer $i \in \{1, ..., n\}$, is equipped with a fixed budget $w_i (> 0)$ and independently solves a linear utility maximization problem

$$\max_{\boldsymbol{x}_i = (x_{i1}, ..., x_{im})} \quad u_i(\boldsymbol{x}_i) = \sum_j u_{ij} x_{ij} \quad \text{s.t. } \boldsymbol{p}^\top \boldsymbol{x}_i \leq w_i,\ \boldsymbol{x}_i \geq \boldsymbol{0}.$$

Here $u_{ij} \geq 0$ is the coefficient of buyer i on good j, and decision variable x_{ij} is the amount of good j purchased by buyer i, and $\boldsymbol{p} \in \mathbb{R}^m$ is a given market price vector. The equilibrium prices are the prices to clear the market.

Another important utility function for each buyer i is called the Leontief utility function

$$u_i(\boldsymbol{x}_i) = \min_j \left\{ \frac{x_{ij}}{u_{ij}} \right\},$$

that is, the utility function value is the smallest x_{ij}/u_{ij}, $j = 1, ..., m$. For simplicity, assume $u_{ij} > 0$ for all i and j.

One fact of the Leontief utility function is that a good may not be able to be cleared (or all purchased) in the market so that its price should be zero from the economic theory. Thus, we look the equilibrium price $\boldsymbol{p} \in \mathbb{R}^m$ and allocation $\boldsymbol{x}_i \in \mathbb{R}^m$, $i = 1, ..., n$, such that the following conditions are met:

$$\sum_i x_{ij} \leq \bar{s}_j, \quad \text{and} \quad p_j \left(\bar{s}_j - \sum_{i=1}^{n} x_{ij} \right) = 0, \ \forall j,$$

and $\boldsymbol{x}_i \in \mathbb{R}^m$, $i = 1, ..., n$, is an optimal solution of

$$\max_{\boldsymbol{x}_i} \quad \min_j \left\{ \frac{x_{ij}}{u_{ij}} \right\} \quad \text{s.t. } \boldsymbol{p}^\top \boldsymbol{x}_i \leq w_i, \ \boldsymbol{x}_i \geq \boldsymbol{0}. \tag{15}$$

(a) (5') What is the interpretation of the Leontief utility function?

Solution (Buying Proportionally) Example: suppose x_1 is the number of left shoes and x_2 the number of right shoes; but a consumer can only use pairs of shoes. Hence, his utility is $\min\{x_1, x_2\}$. Or the cocktail example mentioned in class.

(b) (5') Write down the optimality conditions of buyer ith's maximization problem (15). (Hint: you may simplify the problem by define a scalar utility objective value z_i, then $x_{ij} = u_{ij} z_i$ so that the problem become a one-variable problem.)

Solution

Solution 1 : We can reformulate (15) as the following equivalent LP:

$$\max_{\boldsymbol{x}_i, z_i} \quad z_i \quad \text{s.t. } z_i u_{ij} \leq x_{ij}, \ \boldsymbol{p}^\top \boldsymbol{x}_i \leq w_i, \ \boldsymbol{x}_i \geq \boldsymbol{z}. \tag{16}$$

and the optimality conditions are simply the KKT conditions

$$\begin{cases} z_i u_{ij} \le x_{ij} \\ \boldsymbol{p}^\top \boldsymbol{x}_i \le w_i \\ \lambda_i p = \boldsymbol{\mu}_i + \boldsymbol{\nu}_i \\ \boldsymbol{\mu}_i^\top \boldsymbol{u}_i = 1 \\ \lambda_i (\boldsymbol{p}^\top \boldsymbol{x}_i - w_i) = 0 \\ \mu_{ij}(u_{ij} z_i - x_{ij}) = 0 \\ \nu_{ij} x_{ij} = 0 \\ \boldsymbol{x}_i, \lambda_i, \boldsymbol{\mu}_i, \boldsymbol{\nu}_i \ge \boldsymbol{0} \end{cases} \tag{17}$$

where λ_i is the multiplier of $\boldsymbol{p}^\top \boldsymbol{x}_i - w_i \le 0$, μ_{ij} is the multiplier of $z_i u_{ij} - x_{ij} \le 0$ and ν_{ij} is the multiplier of $-x_{ij} \le 0$. Also $\boldsymbol{u}_i$ is the length m vector with j-th component u_{ij}.

Solution 2 (5 pts): Alternatively, if we write out the dual problem

$$\begin{aligned} \min_{\lambda_i, \boldsymbol{\mu}_i, \boldsymbol{\nu}_i} \quad & \lambda_i w_i \\ \text{s.t.} \quad & \boldsymbol{\mu}_i^\top \boldsymbol{u}_i = 1 \\ & \lambda_i \boldsymbol{p} = \boldsymbol{\mu}_i + \boldsymbol{\nu}_i \\ & \lambda_i, \boldsymbol{\mu}_i, \boldsymbol{\nu}_i \ge 0 \end{aligned}$$

then the equivalent optimality conditions are listed as follows (*i.e.*, primal/dual feasibility and zero duality gap):

$$\begin{cases} \lambda_i w_i = z_i \\ \boldsymbol{p}^\top \boldsymbol{x}_i \le w_i \\ z_i u_{ij} \le x_{ij} \\ \boldsymbol{\mu}_i^\top \boldsymbol{u}_i = 1 \\ \lambda_i \boldsymbol{p} = \boldsymbol{\mu}_i + \boldsymbol{\nu}_i \\ \boldsymbol{x}_i, \lambda_i, \boldsymbol{\mu}_i, \boldsymbol{\nu}_i \ge \boldsymbol{0} \end{cases} \tag{18}$$

Solution 3 (5 pts): Again we can write out the dual problem and the same optimality conditions as in Solution 2, but as we know that for the primal problem to be optimal, we must have $z_i = \min_j \left\{ \frac{x_{ij}}{u_{ij}} \right\}$, hence substituting it into the above

equations we have

$$\begin{cases} \lambda_i w_i = \min_j \left\{ \frac{x_{ij}}{u_{ij}} \right\} \\ \boldsymbol{p}^\top \boldsymbol{x}_i \le w_i \\ \lambda_i \boldsymbol{p} \ge \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^\top \boldsymbol{u}_i = 1 \\ \boldsymbol{x}_i, \lambda_i, \boldsymbol{\mu}_i \ge \boldsymbol{0} \end{cases} \tag{19}$$

All the above conditions are equivalent, but the latter two are more convenient to use in (c) and (d).

Solution 4 (5 pts) (Simplest): The simplest way is to let $x_{ij} = z_i u_{ij}$ for all j, since there is no sense to buy good j more than necessary ($x_{ij} > z_i u_{ij}$), because decreasing x_{ij} does not alter the optimal value for each individual i (Think about the intuition from Part (a)). Then we have the optimality conditions as

$$x_{ij}^* = z_i^* u_{ij} \ \forall i \quad \text{where } z_i^* = \frac{w_i}{\sum_i u_{ij} p_j}; \quad \forall i.$$

One can verify that $\lambda_i^* = z_i^*/w_i$, $\boldsymbol{\mu}_i^* = \lambda_i^* \boldsymbol{p}$ are feasible to Equation (19).

(c) (5') Derive the equilibrium (price $\boldsymbol{p}$ and allocation $\boldsymbol{x}_i,\ i = 1, ..., n$) conditions for the Leontoef market.

<u>Solution</u>

Solution 1 (rigorous version): This question serves as a stepping stone for (d). The key is to prove that $\boldsymbol{p}^\top \boldsymbol{x}_i = w_i$. For any $i \in [n]$, $j \in [m]$, from Equation (18) we know

$$\begin{aligned} & x_{ij} \ge u_{ij} z_i = u_{ij} \lambda_i w_i \ge u_{ij} \lambda_i \boldsymbol{p}^\top \boldsymbol{x}_i \\ \Longrightarrow \quad & \mu_{ij} x_{ij} \ge \mu_{ij} u_{ij} \lambda_i \boldsymbol{p}^\top \boldsymbol{x}_i \\ \Longrightarrow \quad & \boldsymbol{\mu}_i^\top \boldsymbol{x}_i = \textstyle\sum_{i=1}^m \mu_{ij} x_{ij} \ge \sum_{i=1}^m \mu_{ij} u_{ij} \lambda_i \boldsymbol{p}^\top \boldsymbol{x}_i = \boldsymbol{\mu}_i^\top \boldsymbol{u}_i \lambda_i \boldsymbol{p}^\top \boldsymbol{x}_i = \lambda_i \boldsymbol{p}^\top \boldsymbol{x}_i \end{aligned}$$

But we also have

$$\boldsymbol{\mu}_i \le \lambda_i \boldsymbol{p},\ \boldsymbol{x}_i \ge \boldsymbol{0} \Longrightarrow \lambda_i \boldsymbol{p}^\top \boldsymbol{x}_i \ge \boldsymbol{\mu}_i^\top \boldsymbol{x}_i.$$

Hence all the above inequalities become equalities, and hence in particular we have $\boldsymbol{p}^\top \boldsymbol{x}_i = w_i$ and $x_{ij} = z_i u_{ij}$ for all i and j.

Now we can solve for z_i and x_{ij} for a given $\boldsymbol{p}$ by

$$w_i = \sum_{j=1}^m p_j x_{ij} = \sum_{i=1}^m p_j u_{ij} z_i = \boldsymbol{p}^\top \boldsymbol{u}_i z_i$$

so that $z_i = \frac{w_i}{\boldsymbol{p}^\top \boldsymbol{u}_i}$ and $x_{ij} = z_i u_{ij} = \frac{w_i u_{ij}}{\boldsymbol{p}^\top \boldsymbol{u}_i}$. Therefore, we can write down the equilibrium conditions as follows:

$$\begin{cases} z_i = \frac{w_i}{\boldsymbol{p}^\top \boldsymbol{u}_i} & \forall i \\ x_{ij} = \frac{w_i u_{ij}}{\boldsymbol{p}^\top \boldsymbol{u}_i} & \forall i, j \\ \sum_{i=1}^n x_{ij} \le \bar{s}_j & \forall j \\ p_j \left(\bar{s}_j - \sum_{i=1}^n x_{ij}\right) = 0 & \forall j \end{cases}$$

or equivalently (removing x_{ij})

$$\begin{cases} z_i = \dfrac{w_i}{\boldsymbol{p}^\top \boldsymbol{u}_i} & \forall i \\ \displaystyle\sum_{i=1}^n z_i u_{ij} \le \bar{s}_j & \forall j \\ p_j \left(\bar{s}_j - \displaystyle\sum_{i=1}^n z_i u_{ij}\right) = 0 & \forall j \end{cases}$$

Solution 2 (Simplest): Substitute $x_{ij} = z_i u_{ij}$ for all i and j, and then derive the equilibrium conditions using just variables z_i and p_j:

$$\begin{cases} z_i = \frac{w_i}{\boldsymbol{p}^\top \boldsymbol{u}_i} & \forall i \\ \sum_{i=1}^n z_i u_{ij} \le \bar{s}_j & \forall j \\ p_j \left(\bar{s}_j - \sum_{i=1}^n z_i u_{ij}\right) = 0 & \forall j \end{cases}$$

(d) (5') Derive a social optimization problem to represent these conditions.

Solution The social optimization problem can be written as follows:

$$\begin{aligned} \max_{z_i, \boldsymbol{x}_i} \quad & \sum_{j=1}^n w_i \log z_i \\ \text{s.t.} \quad & \sum_{i=1}^n x_{ij} \le \bar{s}_j \quad \forall j \\ & z_i u_{ij} \le x_{ij} \quad \forall i, j \\ & \boldsymbol{x}_i \ge \mathbf{0} \quad \forall i \end{aligned}$$

or the simplest (removing all $\boldsymbol{x}_i$)

$$\begin{aligned} \max_{\boldsymbol{z}} \quad & \sum_{i=1}^n w_i \log z_i \\ \text{s.t.} \quad & \sum_{i=1}^n u_{ij} z_i \le \bar{s}_j \quad \forall i, j \end{aligned} \tag{20}$$

This is a convex optimization problem, and hence the first-order KKT conditions are sufficient and necessary for optimality (to be rigorous, the constraint qualification is satisfied since the feasible region has an interior, but we won't require students to write down such details like checking existence of strict interior points).

Writing down the KKT conditions, we see that it's the same as what we have deduced in (c) with the simplest representation (where $\boldsymbol{p}$ is the multiplier of the constraints in social optimization), and hence we can solve the above the social optimization problem to get the equilibrium price and the optimal allocations altogether.

7. **Question 7** (Bonus 5')

Consider the unconstrained quadratic minimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x},$$

where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is a PSD matrix and a minimizer is $\boldsymbol{x}^* = \boldsymbol{0}$. Assume $\boldsymbol{Q}$ has K distinct positive eigenvalues and denote them by $\lambda_1, ..., \lambda_K$.

Let $\boldsymbol{x}_0$ be any initial solution in $\mathbb{R}^n$, and

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - \frac{1}{\lambda_k}\nabla f(\boldsymbol{x}_{k-1}),$$

for $k = 1, ..., K$. Show that $\boldsymbol{x}_K$ is one optimal solution, that is, the process stops at most K steps. (Hint: you may write $\boldsymbol{Q} = \sum_{k=1}^{K} \lambda_k \boldsymbol{v}_k \boldsymbol{v}_k^\top$, where $\boldsymbol{v}_k$ is the normalized eigenvector corresponding to eigenvalue λ_k.)

Solution Note that $\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - \frac{1}{\lambda_k}\boldsymbol{Q}\boldsymbol{x}_{k-1} = (\boldsymbol{I} - \frac{\boldsymbol{Q}}{\lambda_k})\boldsymbol{x}_{k-1}$ implies that

$$\boldsymbol{x}_K = \prod_{k=1}^{K}(\boldsymbol{I} - \frac{\boldsymbol{Q}}{\lambda_k})\boldsymbol{x}_0.$$

Given the optimality condition $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{0}$, it suffices to prove $\boldsymbol{Q}\prod_{k=1}^{K}(\lambda_k \boldsymbol{I} - \boldsymbol{Q}) = \boldsymbol{0}$, which is true since $f(t) = t\prod_{k=1}^{K}(\lambda_k - t)$ can be divided by the minimal polynomial of $\boldsymbol{Q}$.

This statement also holds for general unconstrained quadratic minimization problem, *i.e.*, $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top\boldsymbol{x}$. Moreover, $\boldsymbol{Q}$ can have same eigenvalues.

In this case, if $\boldsymbol{b}$ is not in the column space of $\boldsymbol{Q}$, the problem is unbounded. Otherwise, we can show that $\boldsymbol{x}_K$ is one optimal solution.

Since $\boldsymbol{Q}$ is a PSD matrix, there exists an orthogonal matrix $\boldsymbol{A}$ such that

$$\boldsymbol{A}\boldsymbol{Q}\boldsymbol{A}^\top = \boldsymbol{\Lambda},\ \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I},$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_K, 0, ..., 0)$. Since $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b}$ is feasible, we have $\boldsymbol{b}$ is in the linear span of first K columns of $\boldsymbol{A}^\top$ and perpendicular to other columns.

Let $\boldsymbol{b}' = \boldsymbol{A}\boldsymbol{b}$, which is the coordinates of $\boldsymbol{b}$ in the basis constructed by the column space of $\boldsymbol{A}^\top$. Then, for any given $\boldsymbol{x}_0$, we have

$$\begin{aligned}\boldsymbol{A}\boldsymbol{x}_1 &= \boldsymbol{A}\boldsymbol{x}_0 - \frac{1}{\lambda_1}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x}_0 + \frac{1}{\lambda_1}\boldsymbol{A}\boldsymbol{b} \\ &= \boldsymbol{A}\boldsymbol{x}_0 - \frac{1}{\lambda_1}\boldsymbol{\Lambda}\boldsymbol{A}\boldsymbol{x}_0 + \frac{1}{\lambda_1}\boldsymbol{A}\boldsymbol{b}.\end{aligned}$$

It implies that $(\boldsymbol{A}\boldsymbol{x}_1)_1 = \frac{1}{\lambda_1}b'_1$. Moreover, by induction, we can show that for any $k \geq 1$, if $(\boldsymbol{A}\boldsymbol{x}_k)_1 = \frac{1}{\lambda_1}b'_1$, it also holds for $k+1$. Thus, we have $\lambda_1(\boldsymbol{A}\boldsymbol{x}_k) = b'_1$ for all k.

Similarly, we can show that at step k, the gradient descent actually find one solution such that $\lambda_k(\boldsymbol{A}\boldsymbol{x}_k)_k = b'_k$, *i.e.*, identifying the coordinates of $\boldsymbol{b}$ in the basis constructed by the column space of $\boldsymbol{A}^\top$. After K steps, we have

$$\boldsymbol{A}\boldsymbol{x}_K = \boldsymbol{\Lambda}^{-1}\boldsymbol{A}\boldsymbol{b} \Rightarrow \boldsymbol{Q}\boldsymbol{x}_K = \boldsymbol{b}.$$

5 Assignment 3

Reading. Read selected sections in [LY21] Chapters 3, 5, 6, 8, 9, 10 and 14.

1. (10') In most real applications, the (first-order) Lipschitz constant β is unknown. Furthermore, we hope to use a localized Lipschitz constant β^k at iteration k such that

$$f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k) - f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^k)^\top (\alpha \boldsymbol{d}^k) \leq \frac{\beta^k}{2} \|\alpha \boldsymbol{d}^k\|^2, \tag{21}$$

where $\boldsymbol{d}^k$ is the steepest descent direction $-\nabla f(\boldsymbol{x}^k)$. The goal is to decide a step-size $\alpha \approx \frac{1}{\beta^k}$.

Consider the following *forward-backward tracking method*. In the following, assume that $\beta^k \geq 1$ and $\alpha_{\max} \geq 1/\beta^k$. Notice that if $\beta^k < 1$, we can enforce it to satisfy our assumption by replacing it with $\max\{1, \beta^k\}$.

Now start at a initial guess $\alpha > 0$,

   (i) If $\alpha \leq \frac{2(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k))}{\|\boldsymbol{d}^k\|^2}$, then doubling the step-size: $\alpha \leftarrow 2\alpha$, stop as soon as the inequality is reversed or $\alpha > \alpha_{\max} (> 0)$, and select the latest α such that the inequality ($\alpha \leq \frac{2(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k))}{\|\boldsymbol{d}^k\|^2}$) holds and $\alpha \leq \alpha_{\max}$.

   (ii) Otherwise halving the step-size: $\alpha \leftarrow \alpha/2$; stop as soon as $\alpha \leq \frac{2(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k))}{\|\boldsymbol{d}^k\|^2}$ and return it.

   (a) (4') Let $\bar{\alpha}$ be a step-size generated by the scheme. Show that $\bar{\alpha} \geq \frac{1}{2\beta^k}$.

   (b) (3') Prove that the above scheme will terminate in finite steps.

   (c) (3') Show that $f(\boldsymbol{x}^k + \bar{\alpha} \boldsymbol{d}^k) \leq f(\boldsymbol{x}^k) - \frac{1}{4\beta^k} \|\boldsymbol{d}^k\|_2^2$.

Solution

   (a) Based on the above definition, the procedure will only terminate in the following two cases:

   - Case I: $2\bar{\alpha} > \alpha_{\max}$. In this case we clearly have $\bar{\alpha} > \frac{1}{2}\alpha_{\max}$ and thus $\bar{\alpha} \geq \frac{1}{2\beta^k}$ since $\alpha_{\max} \geq \frac{1}{\beta^k}$ as it is assumed.
   - Case II: $2\bar{\alpha} > \frac{2(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^k + 2\bar{\alpha} \boldsymbol{d}^k))}{\|\boldsymbol{d}^k\|^2}$. According to (21) it is the case that

$$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^k + 2\bar{\alpha} \boldsymbol{d}^k) \geq 2\bar{\alpha} \|\boldsymbol{d}^k\|^2 - \frac{1}{2}(2\bar{\alpha})^2 \beta^k \|\boldsymbol{d}^k\|^2 = 2(\bar{\alpha} - \beta^k \bar{\alpha}^2) \|\boldsymbol{d}^k\|^2$$

Combining above two inequalities gives

$$2\bar{\alpha} > 4(\bar{\alpha} - \beta^k \bar{\alpha}^2),$$

which implies $\bar{\alpha} \geq \frac{1}{2\beta^k}$ since $\bar{\alpha}, \beta^k > 0$.

(b) Since the scheme is safeguarded above by $\alpha_{\max}$, it suffices to show that the inequality

$$\alpha \leq \frac{2(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k))}{\|\boldsymbol{d}^k\|^2} \tag{22}$$

holds for sufficiently small α. We claim that this inequality holds for any $\alpha \leq \frac{1}{\beta^k}$. To see this, recall (21), which gives for any $\alpha > 0$,

$$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k) \geq \left(\alpha - \frac{\beta^k}{2}\alpha^2\right) \|\boldsymbol{d}^k\|^2.$$

For $\alpha \leq \frac{1}{\beta^k}$ we have $\beta^k \leq \frac{1}{\alpha}$, and thus

$$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k) \geq \left(\alpha - \frac{1}{2\alpha}\alpha^2\right) \|\boldsymbol{d}^k\|^2 = \frac{1}{2}\alpha \|\boldsymbol{d}^k\|^2,$$

which immediately implies the inequality $\alpha \leq \frac{2(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k))}{\|\boldsymbol{d}^k\|^2}$.

(c) Based on the above definition, the $\bar{\alpha}$ generated by the scheme always satisfies the inequality (22). Combining (22) and the fact that $\bar{\alpha} \geq \frac{1}{2\beta^k}$ immediately tells that

$$f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k) - f(\boldsymbol{x}^k) \leq -\frac{1}{2}\bar{\alpha}\|\boldsymbol{d}^k\|^2 \leq -\frac{1}{4\beta^k}\|\boldsymbol{d}^k\|^2.$$

2. (10') (L_2 Regularization and Logarithmic Barrier) Consider the optimization problem

$$\begin{aligned} \min_{x_1, x_2} \quad & (x_1 - x_2 + 1)^2 \\ \text{s.t.} \quad & x_1 \geq 0 \quad x_2 \text{ "free"}. \end{aligned}$$

Then we may combine the L_2-regularization and barrier together, that is, for any $\mu > 0$, consider

$$\min_{x_1, x_2} \quad (x_1 - x_2 + 1)^2 + \frac{\mu}{2}(x_1^2 + x_2^2) - \mu \log(x_1)$$

(a) (4') Develop explicit path formula in terms of μ. What is the limit solution as $\mu \to 0$?

(b) (3') Using $\mu = 1$ and $\boldsymbol{x}^0 = (1,\ 0)$, apply one step of SDM (Steepest Descent Method) with step-size $1/5$ to compute the next iterate.

(c) (3') Using $\mu = 1$ and $\boldsymbol{x}^0 = (1,\ 0)$, apply one step of Newton's Method to compute the next iterate.

Solution

(a) Let $f(\boldsymbol{x};\mu) := (x_1 - x_2 + 1)^2 + \frac{\mu}{2}(x_1^2 + x_2^2) - \mu\log(x_1)$. Then the gradient of f is

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x};\mu) = \begin{bmatrix} 2(x_1 - x_2 + 1) + \mu\boldsymbol{x}_1 - \mu/x_1 \\ 2(x_2 - x_1 - 1) + \mu\boldsymbol{x}_2 \end{bmatrix}.$$

Setting the gradient to zero (KKT conditions), we obtain that $x_2 = \frac{2x_1+2}{2+\mu}$, and $x_1 = \frac{\mu+2}{\mu+4}$. And as a result, we obtain that the path formula of $\boldsymbol{x}(\mu)$ is

$$\boldsymbol{x}(\mu) = \begin{bmatrix} (\mu+2)/(\mu+4) \\ (4\mu+12)/(\mu^2+6\mu+8) \end{bmatrix}.$$

Taking $\mu \to 0$, we obtain that the limit solution is $x_1 = 1/2,\ x_2 = 3/2$.

(b) By definition of SDM, we have

$$\boldsymbol{x}^1 = \boldsymbol{x}^0 - \frac{1}{5}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^0;1) = \begin{bmatrix} 1/5 \\ 4/5 \end{bmatrix}.$$

(c) By simple calculation, we have

$$\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x};\mu) = \begin{bmatrix} 2+\mu+\mu/x_1^2 & -2 \\ -2 & 2+\mu \end{bmatrix}.$$

And so with $\boldsymbol{x}^0 = (1,0)$ and $\mu = 1$, we have

$$\nabla^2_{\boldsymbol{x}} f((1,0);1) = \begin{bmatrix} 4 & -2 \\ -2 & 3 \end{bmatrix}.$$

By definition of Newton's method, we have

$$\boldsymbol{x}^1 = \boldsymbol{x}^0 - \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}^0;1)^{-1}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^0;1) = \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}.$$

3. (20') (L_2 Path-Following) Consider a convex function $f : \mathbb{R}^n \to \mathbb{R}$ in C^2 that is twice continuously differentiable. Assume that its value is bounded from below and that it has a minimizer. For any given positive parameter $\mu > 0$, consider the regulated minimization problem

$$\min\ f(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}\|^2. \tag{23}$$

Prove the following claims:

(a) (2') Write down the first-order optimality condition of Problem (23). Is it sufficient for $\boldsymbol{x}$ to be a minimizer?

(b) (5') The minimizer, denoted by $\boldsymbol{x}(\mu)$, of Problem (23) is unique in μ.

(c) (5') $f(\boldsymbol{x}(\mu))$ is an increasing function of μ (*i.e.*, $f(\boldsymbol{x}(\mu)) \geq f(\boldsymbol{x}(\mu'))$ if $\mu \geq \mu' > 0$), and $\|\boldsymbol{x}(\mu)\|$ is a decreasing function of μ.

(d) (5') As $\mu \to 0^+$ (*i.e.*, μ decreases to 0), $\boldsymbol{x}(\mu)$ converges to the minimizer of $f(\boldsymbol{x})$ with the minimal Euclidean norm.

(e) (3') Consider the specific example

$$\min_{x_1,x_2} (x_1 - x_2 - 1)^2,$$

where the optimal solution set is unbounded. Write out the explicit path formula of $\boldsymbol{x}(\mu) = (x_1(\mu), \ldots, x_n(\mu))$ in terms of μ. What is the limit solution as $\mu \to 0$?

Solution

(a) The first-order optimality condition is $\nabla f(\boldsymbol{x}) + \mu \boldsymbol{x} = 0$. Yes, because the problem is convex, and hence KKT conditions are sufficient (with no other requirements).

(b) In the following, apart from proving the uniqueness (we only require you to prove this in the homework), we also prove that $\boldsymbol{x}(\mu)$ is continuous in μ, which is similar to part of the proof in (d) below.

The uniqueness is simply due to strict (actually strong) convexity. Suppose $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ are both the optimal solutions of Problem (23), then we have the optimality condition

$$\nabla f(\boldsymbol{x}_1) = -\mu \boldsymbol{x}_1, \nabla f(\boldsymbol{x}_2) = -\mu \boldsymbol{x}_2.$$

Hence, we have

$$(\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2))^\top (\boldsymbol{x}_1 - \boldsymbol{x}_2) = -\mu \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 < 0$$

However, the gradient mapping must be monotone, *i.e.*,

$$(\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2))^\top (\boldsymbol{x}_1 - \boldsymbol{x}_2) \geq 0, \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2$$

since f is convex, which leads to a contradiction. To see the monotonicity of $\nabla f(\cdot)$, we simply add two gradient inequalities together:

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^\top (\boldsymbol{x}_2 - \boldsymbol{x}_1), \; f(\boldsymbol{x}_1) \geq f(\boldsymbol{x}_2) + \nabla f(\boldsymbol{x}_2)^\top (\boldsymbol{x}_1 - \boldsymbol{x}_2).$$

To show the continuity, notice that $\nabla f(\boldsymbol{x}(\mu))+\mu\boldsymbol{x}(\mu) = 0$ for any $\mu > 0$. Consider any $\mu, \mu' > 0$, we subtract their optimality conditions and obtain

$$-\mu\boldsymbol{x}(\mu) + \mu'\boldsymbol{x}(\mu') = \nabla f(\boldsymbol{x}(\mu)) - \nabla f(\boldsymbol{x}(\mu')).$$

Multiplying $\boldsymbol{x}(\mu) - \boldsymbol{x}(\mu')$ on both sides, we get

$$(\boldsymbol{x}(\mu)-\boldsymbol{x}(\mu'))^\top(-\mu\boldsymbol{x}(\mu)+\mu'\boldsymbol{x}(\mu')) \geq (\boldsymbol{x}(\mu)-\boldsymbol{x}(\mu'))^\top(\nabla f(\boldsymbol{x}(\mu))-\nabla f(\boldsymbol{x}(\mu'))) \geq 0.$$

Hence we have

$$\mu\|\boldsymbol{x}(\mu) - \boldsymbol{x}(\mu')\|^2 \leq (\mu' - \mu)(\boldsymbol{x}(\mu) - \boldsymbol{x}(\mu'))^\top\boldsymbol{x}(\mu').$$

Now we make use of the monotonicity *to be established* in (c) (which does not depend on the result in (b) here), and observe that if we fix $\mu > 0$, and assume that $|\mu'-\mu| \leq C$ for some $C > 0$, then the monotonicity of $\|\boldsymbol{x}(\mu)\|$ ensures that $\|\boldsymbol{x}(\mu')\|$ is bounded by some constant that may depend on μ (which is fixed now). Hence by taking the limit $\mu' \to \mu$, we obtain that the RHS above converges to 0 (using in addition that $\boldsymbol{x}^\top\boldsymbol{y} \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$). Hence we conclude that $\|\boldsymbol{x}(\mu) - \boldsymbol{x}(\mu')\| \to 0$ as $\mu' \to \mu > 0$. Since μ is arbitrary, we have proved that $\boldsymbol{x}(\mu)$ is continuous in μ.

(c) Let $0 < \mu' < \mu$. Then by the optimality of the solutions, we have

$$f(\boldsymbol{x}(\mu')) + \mu'/2\|\boldsymbol{x}(\mu')\|^2 \leq f(\boldsymbol{x}(\mu)) + \mu'/2\|\boldsymbol{x}(\mu)\|^2$$

and

$$f(\boldsymbol{x}(\mu)) + \mu/2\|\boldsymbol{x}(\mu)\|^2 \leq f(\boldsymbol{x}(\mu')) + \mu/2\|\boldsymbol{x}(\mu')\|^2.$$

Adding the above inequalities, we see that

$$\frac{\mu - \mu'}{2}\|\boldsymbol{x}(\mu')\|^2 \geq \frac{\mu - \mu'}{2}\|\boldsymbol{x}(\mu)\|^2.$$

Since $\mu-\mu' > 0$, we have $\|\boldsymbol{x}(\mu')\|^2 \geq \|\boldsymbol{x}(\mu)\|^2$, showing that $\|\boldsymbol{x}(\mu)\|$ is a decreasing function of μ. Finally, using any one of the first inequality above, we see that $f(\boldsymbol{x}(\mu')) \leq f(\boldsymbol{x}(\mu)) + \mu'/2(\|\boldsymbol{x}(\mu)\|^2 - \|\boldsymbol{x}(\mu')\|^2) \leq f(\boldsymbol{x}(\mu))$, and hence $f(\boldsymbol{x}(\mu))$ is an increasing function of μ.

(d) Let $\boldsymbol{x}^*$ be an arbitrary minimizer of $f(\boldsymbol{x})$. Then we have $\nabla f(\boldsymbol{x}^*) = 0$, and together with the fact that $\nabla f(\boldsymbol{x}(\mu)) + \mu\boldsymbol{x}(\mu) = 0$, we obtain that

$$\nabla f(\boldsymbol{x}(\mu)) - \nabla f(\boldsymbol{x}^*) + \mu\boldsymbol{x}(\mu) = 0.$$

Similar to (b), multiplying both sides with $\boldsymbol{x}(\mu) - \boldsymbol{x}^*$, we have by convexity of f that

$$-\mu(\boldsymbol{x}(\mu) - \boldsymbol{x}^*)^\top \boldsymbol{x}(\mu) = (\boldsymbol{x}(\mu) - \boldsymbol{x}^*)^\top \nabla f(\boldsymbol{x}(\mu)) \geq 0.$$

Hence we have that $\|\boldsymbol{x}(\mu)\|^2 \leq \boldsymbol{x}(\mu)^\top \boldsymbol{x}^* \leq \|\boldsymbol{x}^*\| \|\boldsymbol{x}(\mu)\|$, which proves that $\|\boldsymbol{x}(\mu)\| \leq \|\boldsymbol{x}^*\|$ (for any $\mu > 0$).

Now notice that to prove the claim in the question, we only need to show that for any sequence $\mu_k \to 0^+$, $\boldsymbol{x}(\mu_k)$ converges to a minimizer $\bar{\boldsymbol{x}}$ of $f(\boldsymbol{x})$ with minimal Euclidean norm.

To this end, we first notice that $\{\boldsymbol{x}(\mu_k)\}_{k=1}^\infty$ is bounded, and hence it has a limit point, *i.e.*, there exists a convergent subsequence $\boldsymbol{x}(\mu_{k_i})$ with the limit point $\bar{\boldsymbol{x}}$. Moreover, for any accumulation point $\bar{\boldsymbol{x}}'$ of $\boldsymbol{x}(\mu_k)$, since f is C^2, $\nabla f(\cdot)$ is continuous, and hence we have $\nabla f(\bar{\boldsymbol{x}}') = \lim_{i\to\infty} \nabla f(\boldsymbol{x}(\mu_{l_i})) = -\lim_{i\to\infty} \mu_{l_i} \boldsymbol{x}(\mu_{l_i}) = \boldsymbol{0}$ for some subsequence l_i where the last equality comes from the fact that $\boldsymbol{x}(\mu_{l_i})$ is bounded and $\mu_{l_i} \to 0^+$. Hence we see that any accumulation point of $\boldsymbol{x}(\mu_k)$ is an optimal solution.

In addition, it also comes with minimum Euclidean norm since $\|\boldsymbol{x}(\mu)\| \leq \|\boldsymbol{x}^*\|$ for any minimizer $\boldsymbol{x}^*$, which implies that $\|\bar{\boldsymbol{x}}'\| \leq \|\boldsymbol{x}^*\|$ also holds and hence $\bar{\boldsymbol{x}}'$ is a minimal Euclidean norm minimizer of $f(\boldsymbol{x})$.

Finally, it's easy to show that the minimal norm solution is unique by contradiction. Suppose there are two minimal norm solutions $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, then $(\boldsymbol{x}_1 + \boldsymbol{x}_2)/2$ is also a minimizer of $f(\boldsymbol{x})$ but has a smaller norm (unless $\boldsymbol{x}_1 = \boldsymbol{x}_2$), which is a contradiction. Hence we see that any accumulation point of $\boldsymbol{x}(\mu_k)$ is the same minimal norm minimizer $\bar{\boldsymbol{x}}$. Since μ_k is an arbitrary sequence with $\mu_k \to 0^+$, we conclude that as $\mu \to 0^+$, $\boldsymbol{x}(\mu)$ converges to the unique minimizer of $f(\boldsymbol{x})$ with the minimal Euclidean norm.

(e) By the first-order optimality conditions, we have $x_1 = -x_2 = \frac{2}{\mu+4}$. As $\mu \to 0$, the limit solution is $x_1 = -x_2 = \frac{1}{2}$.

4. (15') (Affine-Scaling Interior-Point SD) Consider the conic constrained optimization problem

$$\min_x \quad f(\boldsymbol{x}) \qquad \text{s.t.} \quad \boldsymbol{x} \geq \boldsymbol{0} \tag{24}$$

where we assume the objective function f is first-order β-Lipschitz. Starting from $\boldsymbol{x}^0 = \boldsymbol{e} > \boldsymbol{0}$, consider the affine-scaling interior-point method as follows: at iterate

$\boldsymbol{x}^k > \mathbf{0}$ let diagonal scaling matrix $\boldsymbol{D}$ be

$$D_{ii} = \min\{1,\ x_i^k\}$$

and

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha^k \boldsymbol{D}^2 \nabla f(\boldsymbol{x}^k),$$

with step-size

$$\alpha^k = \min\left\{\frac{1}{\beta},\ \frac{1}{2\|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty}\right\}. \tag{25}$$

(a) (3') Show that $\boldsymbol{D}^2\nabla f(\boldsymbol{x}^k)$ is a descent direction.

(b) (3') Show that $\boldsymbol{x}^{k+1} > \mathbf{0}$ for all $k = 0, 1, ...$

(c) (6') Show that

$$f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^k) \le \min\left\{-\frac{1}{2\beta}\|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty^2, -\frac{1}{4}\|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty\right\}$$

(d) (3') Derive a iterative complexity bound for $\|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty \le \epsilon$.

Solution

(a) Denote $\boldsymbol{d}^k = -\boldsymbol{D}^2\nabla f(\boldsymbol{x}^k)$. By definition of $\boldsymbol{D}$ and the fact that $\boldsymbol{x}^k > \mathbf{0}$ one clearly have $\boldsymbol{D} \succ \mathbf{0}$, and thus $\boldsymbol{D}^2 \succ \mathbf{0}$. Hence $\boldsymbol{d}^{k\top}\nabla f(\boldsymbol{x}^k) = -\nabla f(\boldsymbol{x}^k)\boldsymbol{D}^2\nabla f(\boldsymbol{x}^k) < 0$ if $\nabla f(\boldsymbol{x}^k) \neq \mathbf{0}$. Hence $\boldsymbol{d}^k$ is a descent direction if $\nabla f(\boldsymbol{x}^k) \neq \mathbf{0}$.

(b) We will prove the fact that $\boldsymbol{x}^k > \mathbf{0}$ for all $k = 0, 1, \ldots$ by induction. Clearly for $k = 0$ one have $\boldsymbol{x}^0 = \boldsymbol{e} > 0$ by assumption. Assume this fact holds for k, *i.e.*, $\boldsymbol{x}^k > 0$, consider the $(k+1)$-th iteration. It follows that the i-th component of $\boldsymbol{x}^{k+1}$ satisfies

$$x_i^{k+1} = x_i^k - \alpha^k(\boldsymbol{D}^2\nabla f(\boldsymbol{x}^k))_i \ge x_i^k - \alpha^k|(\boldsymbol{D}^2\nabla f(\boldsymbol{x}^k))_i| = x_i^k - \alpha^k(D_{ii})|(\boldsymbol{D}\nabla f(\boldsymbol{x}^k))_i|$$

Since $D_{ii} = \min\{1, x_i^k\}$ one have $D_{ii} \le x_i^k$, and thus

$$x_i^k - \alpha^k(D_{ii})|(\boldsymbol{D}\nabla f(\boldsymbol{x}^k))_i| \ge x_i^k - \alpha^k x_i^k|(\boldsymbol{D}\nabla f(\boldsymbol{x}^k))_i| = x_i^k(1 - \alpha^k(\boldsymbol{D}\nabla f(\boldsymbol{x}^k))_i|)$$

Note that $D_{ii}|\nabla f(\boldsymbol{x}^k)|_i \le \|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty$, by definition of α^k we obtain $\alpha^k D_{ii}|\nabla f(\boldsymbol{x}^k)|_i \le \frac{1}{2}$. Combining these results above yields $x_i^{k+1} \ge \frac{1}{2}x_i^k > 0$. This concludes the induction.

(c) Since the function is first-order β-Lipschitz, for each step,

$$\begin{aligned} f(\boldsymbol{x}^{k+1}) &\leq f(\boldsymbol{x}^k) - (\alpha^k \boldsymbol{D}^2 \nabla f(\boldsymbol{x}^k))^\top \nabla f(\boldsymbol{x}^k) + \frac{\beta}{2}(\alpha^k)^2 \|\boldsymbol{D}^2 \nabla f(\boldsymbol{x}^k)\|_2^2 \\ &= f(\boldsymbol{x}^k) - \alpha^k \|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_2^2 + \frac{\beta}{2}(\alpha^k)^2 \|\boldsymbol{D}^2 \nabla f(\boldsymbol{x}^k)\|_2^2 \end{aligned}$$

Since $D_{ii} \leq 1$ we have $\|\boldsymbol{D}^2 \nabla f(\boldsymbol{x}^k)\|_2^2 \leq \|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_2^2$, and therefore

$$\begin{aligned} f(\boldsymbol{x}^{k+1}) &\leq f(\boldsymbol{x}^k) - \alpha^k \|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_2^2 + \frac{\beta}{2}(\alpha^k)^2 \|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_2^2 \\ &= f(\boldsymbol{x}^k) - \left(\alpha^k - \frac{\beta}{2}(\alpha^k)^2\right) \|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_2^2 \end{aligned} \tag{26}$$

Note that $\|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_2^2 \geq \|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty^2$. According to the scheme, the inequality $0 \leq \alpha^k \leq \frac{1}{\beta}$ always holds, which implies $\alpha^k - \frac{\beta}{2}(\alpha^k)^2 \in [0, \frac{1}{2\beta}]$. Therefore

$$f(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{x}^k) - \left(\alpha^k - \frac{\beta}{2}(\alpha^k)^2\right) \|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty^2 \tag{27}$$

According to the step-size scheme (25), there are two cases:

- Case I: $\alpha^k = \frac{1}{\beta} \leq \frac{1}{2\|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty}$. In this case (according to (27)),

$$f(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{x}^k) - \frac{1}{2\beta}\|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty^2 \tag{28}$$

- Case II: $\alpha^k = \frac{1}{2\|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty} \leq \frac{1}{\beta}$. In this case (according to (27)),

$$\begin{aligned} f(\boldsymbol{x}^{k+1}) &\leq f(\boldsymbol{x}^k) - \left(1 - \frac{\beta}{2}\alpha^k\right) \frac{1}{2\|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty} \|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty^2 \\ &= f(\boldsymbol{x}^k) - \frac{1}{2}\left(1 - \frac{\beta}{2}\alpha^k\right) \|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty \\ &\leq f(\boldsymbol{x}^k) - \frac{1}{4}\|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty \end{aligned} \tag{29}$$

  where in the last inequality we used the fact that $\alpha^k \leq \frac{1}{\beta}$.

(d) Combining the two cases immediately tells that the method will identify an $\boldsymbol{x}^k$ such that $\|\boldsymbol{D}\nabla f(\boldsymbol{x}^k)\|_\infty \leq \varepsilon$ within $\max\left\{\frac{4(f(\boldsymbol{x}^0)-f^*)}{\varepsilon}, \frac{2\beta(f(\boldsymbol{x}^0)-f^*)}{\varepsilon^2}\right\}$ steps.

Computational Homework:

5. (10') There is a simple nonlinear least squares approach for Sensor Network Localization:

$$\min \sum_{(ij)\in N_x} \left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 - d_{ij}^2\right)^2 + \sum_{(kj)\in N_a} \left(\|\boldsymbol{a}_k - \boldsymbol{x}_j\|^2 - d_{kj}^2\right)^2 \tag{30}$$

which is an unconstrained nonlinear minimization problem.

   (a) (5') Apply the Steepest Descent Method, starting with either the origin or a random solution as the initial solution for model (30), to solve few selected SNL instances you created in Assignment 1. Does it work?

   (b) (5') Apply the same Steepest Descent Method, starting from the SOCP or SDP solution (which may not have errors) as the initial solution for model (30), to solve the same instances in (a). Does it work? Does the SOCP or SDP initial solution make a difference?

Solution The Julia code is online here where we try SDM on Problem 9(b) of Assignment 1. We will accept any reasonable solutions showing sufficient efforts, and the code is just for references.

   (a) Although the SDM typically converges under a reasonable error tolerance (*e.g.*, the tolerance of the gradient norm is 10^{-6}), whether it can recover the sensor location exactly depends on the initialization point.

   (b) We randomly generate 1000 sensor locations in $[-1, 1]^2$ and run SDM with the origin as the initial point, SOCP relaxation, SDP relaxation, SDM with the SOCP solution as the initial point, SDM with the SDP solution as the initial point. We observe that

      – SDM can help improve the solutions from SOCP and SDP relaxations.

      – SDM with the SDP solution as the initial point performs the best in terms of the chance of exact recovery.

6. (30') (Multi-Block ADMM)

Part I Implement the ADMM to solve the divergence example:

$$\begin{aligned} \text{minimize} \quad & 0 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 \\ \text{subjectto} \quad & \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \mathbf{0} \end{aligned}$$

(a) (5') Try $\beta = 0.1$, $\beta = 1$, and $\beta = 10$, respectively. Does the choice of β make a difference?

(b) (5') Add the objective function to minimize

$$0.5(x_1^2 + x_2^2 + x_3^2)$$

to the problem, and retry $\beta = 0.1$, $\beta = 1$, and $\beta = 10$, respectively. Does the choice of β make a difference?

(c) (5') Set $\beta = 1$ and apply the randomly permuted updating-order of $\boldsymbol{x}$ (discussed in class) to solving each of the two problems in (a) and (b). Does the iterate converge?

Part II Generate some (feasible) convex QP problems with linear equality constraints, say 30 variables and 10 constraints (*i.e.*, $\boldsymbol{A} \in \mathbb{R}^{10\times 30}$),

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned}$$

(d) (5') Divide the variables of $\boldsymbol{x}$ into 5 blocks and apply the ADMM with $\beta = 1$. Does it converge? (You may construct 5 different blocks and conduct the experiments.)

(e) (5') Apply the randomly permuted updating-order of the 5 blocks in each iteration of the ADMM. Does it converge? Convergence performance?

(f) (5') Consider the following scheme – random-sample-without-replacement: in each iteration of ADMM, randomly sample 6 variables for update, and then randomly select 6 variables from the remaining 24 variable for update, and... , till all 30 variables are updated; then update the multipliers as usual. Does it converge? Convergence performance?

Solution The Julia code is online here.

First of all, we would like to comment that since for each sub-step, the minimization problem is a convex QP, and hence we can simply solve it by taking the gradient and setting it to 0, which reduces to the problem to solving a linear system. Hence in this problem, either you apply a gradient step to the sub-problem or solve it directly, you never need to use some other general optimization solvers.

(a) As shown in Figure 10, the choice of β doesn't really make a difference. For all three choices of β, the procedure diverges, especially they diverge in a very similar geometric rate. The MATLAB code below is based on the algorithmic mapping in Lecture Note 11.

```
% Part I (a)
rng(1); A = [1,1,1;1,1,2;1,2,2]; LA = tril(A^2); UA = LA-A^2;
figure();
for beta = [0.1,1,10]
    M_lhs = [LA*beta, zeros(3,3);A*beta,eye(3)];
    M_rhs = [UA*beta, A; zeros(3,3),eye(3)];
    M = M_lhs\M_rhs;
    x = rand(3,1); y = rand(3,1); xnorm = [];
    for time = 1:1000
        xynew = M*[x;y];
        x = xynew(1:3); y = xynew(4:6);
        xnorm = [xnorm, norm(x)];
    end
    semilogy(xnorm); hold on
end
legend('\beta = 0.1','\beta = 1','\beta = 10', 'Location','best');
xlabel('Iterations'); ylabel('x Norm');
t = title('3-block ADMM: divergence example');
```

(b) After we add $0.5(x_1^2 + x_2^2 + x_3^2)$ to the objective function, the choice of β makes a difference. In Figure 11, it appears that the procedure diverges for $\beta = 10$, and converges for $\beta = 0.1$ and $\beta = 1$. This time the procedure can converge for some β and hence, to a certain extent, better than the previous one in (a). It is because the objective function becomes strictly convex. The step size corresponds to $\beta = 10$ is too large so that the procedure does not converge. At the same time, $\beta = 0.1$ corresponds to a much smaller step size, making the convergence rate not

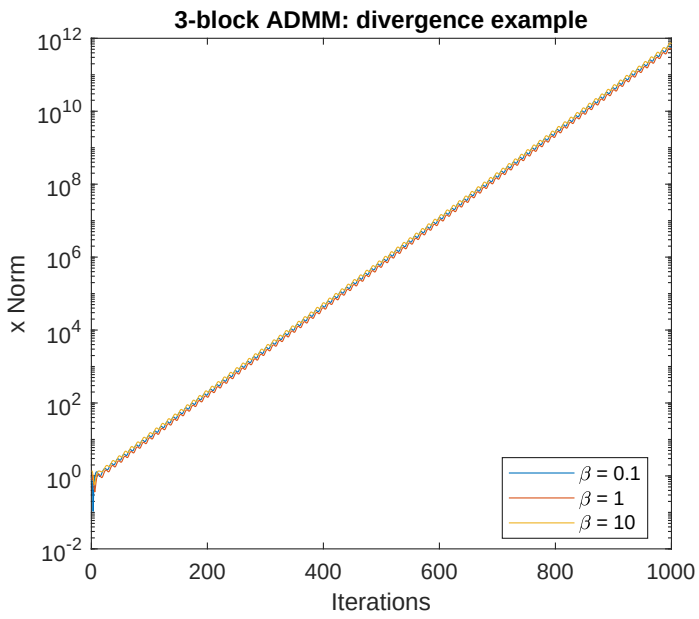


Figure 10: (a): the divergence example

as good as the case of $\beta = 1$. Still, both cases converge geometrically.

```
% Part I (b)
rng(1); A = [1,1,1;1,1,2;1,2,2]; LA = tril(A^2); UA = LA-A^2;
figure();
for beta = [0.1,1,10]
    M_lhs = [LA*beta + eye(3), zeros(3,3);A*beta,eye(3)];
    M_rhs = [UA*beta, A; zeros(3,3),eye(3)];
    M = M_lhs\M_rhs;
    x = rand(3,1); y = rand(3,1); xnorm = [];
    for time = 1:1000
        xynew = M*[x;y];
        x = xynew(1:3); y = xynew(4:6);
        xnorm = [xnorm, norm(x)];
    end
    semilogy(xnorm); hold on
end
legend('\beta = 0.1','\beta = 1','\beta = 10', 'Location','best');
xlabel('Iterations'); ylabel('x Norm');
t = title('3-block ADMM: Adding 0.5(x_1^2 + x_2^2 + x_3^2) to ...
    objective');
```

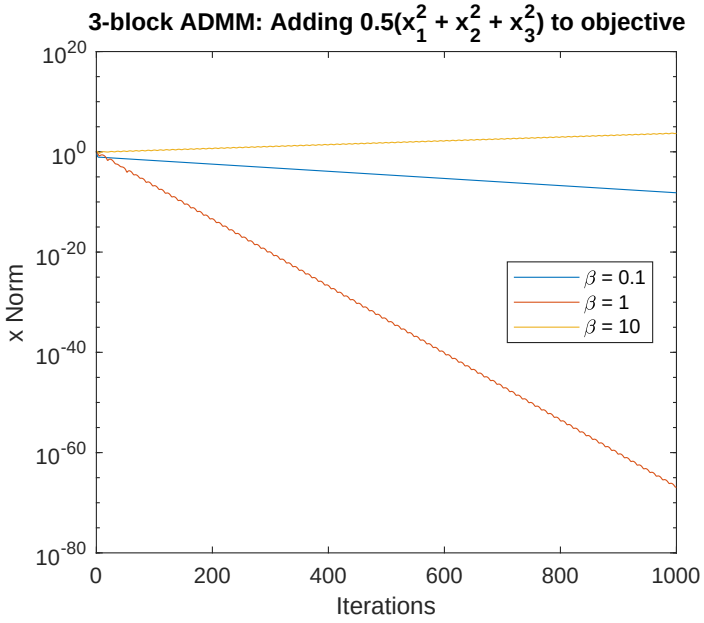


Figure 11: (b): regularized 3-block ADMM

(c) See Figure 12. After randomly permuting the updating-order of $\boldsymbol{x}$, for both problems in (a) and (b), the iterates converge. The one corresponds to (a) converges slower than the that to (b), as the problem is more "convex" for (b). Both converge in roughly a geometric rate.

```
% Part I (c)
A = [1,1,1;1,1,2;1,2,2]; beta = 1;
x = rand(3,1); y = rand(3,1); xnorm = [];

for time = 1:1000
    r_index = randperm(3);
    A_2 = A^2; LA = tril(A_2(r_index,r_index));
    [t,r_rank] = sort(r_index);
    LA = LA(r_rank, r_rank); UA = LA-A^2;
    M_lhs = [LA*beta, zeros(3,3);A*beta,eye(3)];
    M_rhs = [UA*beta, A; zeros(3,3),eye(3)];
    M = M_lhs\M_rhs;
    xynew = M*[x;y]; x = xynew(1:3); y = xynew(4:6);
    xnorm = [xnorm, norm(x)];
end
figure();
```

```
semilogy(xnorm); legend('\beta = 1'); xlabel('Iterations'); ...
    ylabel('x Norm');
t = title('3-block ADMM: randomly permuted updating-order - (a)');

x = rand(3,1); y = rand(3,1); xnorm = [];
for time = 1:1000
    r_index = randperm(3);
    A_2 = A^2; LA = tril(A_2(r_index,r_index));
    [t,r_rank] = sort(r_index);
    LA = LA(r_rank, r_rank); UA = LA-A^2; LA = LA+eye(3);
    M_lhs = [LA*beta, zeros(3,3);A*beta,eye(3)];
    M_rhs = [UA*beta, A; zeros(3,3),eye(3)];
    M = inv(M_lhs)*M_rhs;
    xynew = M*[x;y]; x = xynew(1:3); y = xynew(4:6);
    xnorm = [xnorm, norm(x)];
end
figure();
semilogy(xnorm); legend('\beta = 1'); xlabel('Iterations'); ...
    ylabel('x Norm');
t = title('3-block ADMM: randomly permuted updating-order - (b)');
```

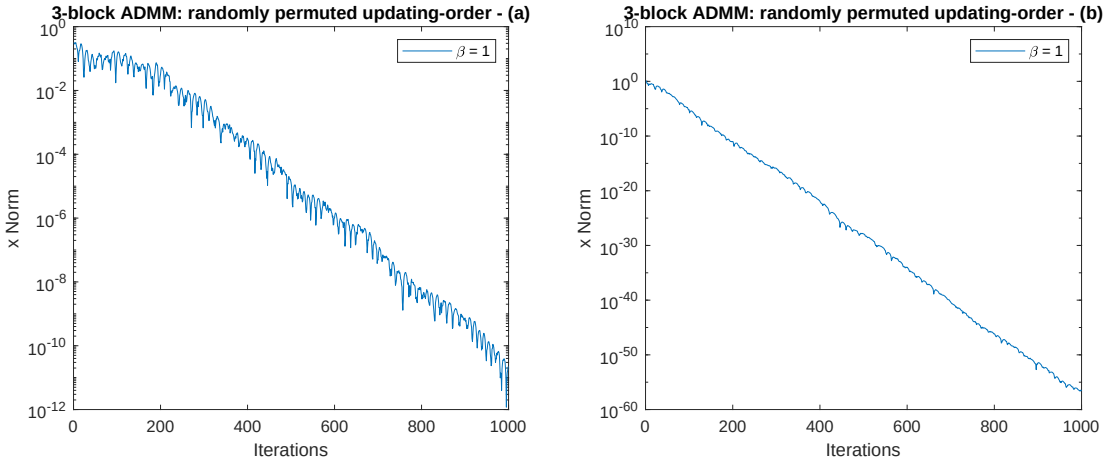


Figure 12: (c): randomly permuting the updating-order

(d)-(f) In Figure 13, we consider three procedures. We'll refer to them as procedure 1,2,3 respectively.

1. Divide the variables of $\boldsymbol{x}$ into 5 blocks and apply the ADMM with $\beta = 1$. The procedure does converge.

2. Apply the randomly permuted updating-order of the 5 blocks in each iteration of the ADMM. The procedure also does converge.

3. The new scheme using random-sample-without-replacement: in each iteration of ADMM, randomly sample 6 variables for update, and then randomly select 6 variables from the remaining 24 variable for update, and... , till all 30 variables are updated; then update the multipliers as usual. The procedure also converges.

All of the three converge in geometric rate. The new scheme-random-sample-without-replacement appears to converge the fastest, followed by no permutation. Note that the yellow line doesn't appear to move after 300 iterations, this may be due to that the "true value" of $\boldsymbol{x}$ we used is not perfectly accurate.

```
%%construct the problem
rng(1);
A = rand(10,30); Q_half = rand(30,30);
x_0 = rand(30,1); b = A*x_0; Q = Q_half'*Q_half; LU = Q + A'*A;
%%use cvx to find the true value
cvx_begin quiet
    variable x_true(30)
    minimize( x_true'*Q*x_true )
    subject to
        A*x_true == b
cvx_end

a=ones(6,6); n=5;
AA=repmat(a,n,1); BB=mat2cell(AA,6*ones(1,n),6);
Diagonal_B=blkdiag(BB{:}); lower_B = tril(ones(30,30)) + ...
   triu(Diagonal_B,1);
run_num = 2000;
%%Without random permutation
beta = 1; x = rand(30,1); y = rand(10,1); xnorm1 = [];
for time = 1:run_num
    LA = lower_B.*LU; UA = LA-LU;
    M_lhs = [LA, zeros(30,10);A,eye(10)];
```

```
    M_rhs = [UA, A'; zeros(10,30),eye(10)];
    M_b = [A'*b; b];
    xynew = M_lhs\(M_rhs*[x;y]+M_b); x = xynew(1:30); y = ...
        xynew(31:40);
    xnorm1 = [xnorm1, norm(x-x_true)];
end

%%With random permutation of block
x = rand(30,1); y = rand(10,1); xnorm2 = [];
for time = 1:run_num
    group_index = randperm(5);
    r_index = reshape(6*repmat(group_index,6,1) + ...
        repmat((0:5)',1,5)-5,1,30);

    LA = lower_B.*(LU(r_index,r_index)); [t,r_rank] = ...
        sort(r_index);
    LA = LA(r_rank, r_rank); UA = LA-LU;

    M_lhs = [LA, zeros(30,10);A,eye(10)];
    M_rhs = [UA, A'; zeros(10,30),eye(10)];
    M_b = [A'*b; b];

    xynew = M_lhs\(M_rhs*[x;y]+M_b); x = xynew(1:30); y = ...
        xynew(31:40);
    xnorm2 = [xnorm2, norm(x-x_true)];
end

%%new scheme random-sample-without-replacement

x = rand(30,1);y = rand(10,1);xnorm3 = [];

for time = 1:run_num
    r_index = randperm(30);

    LA = lower_B.*(LU(r_index,r_index));
    [t,r_rank] = sort(r_index); LA = LA(r_rank, r_rank); UA = ...
        LA-LU;

    M_lhs = [LA, zeros(30,10);A,eye(10)];
    M_rhs = [UA, A'; zeros(10,30),eye(10)];
    M_b = [A'*b; b];
```

```
58
59      xynew = M_lhs\(M_rhs*[x;y]+M_b); x = xynew(1:30); y = ...
            xynew(31:40);
60      xnorm3 = [xnorm3, norm(x-x_true)];
61  end
62
63  %%plot
64  figure(); semilogy(xnorm1); hold on
65  semilogy(xnorm2); hold on
66  semilogy(xnorm3);
67
68  legend('No random permutation', 'Random permutation of ...
        blocks', 'New scheme: random-sample-without-replacement', ...
        'Location','northeast');
69  xlabel('Iterations'); ylabel('||x-x_0||');
70  t = title('3-block ADMM: With and without random permutations');
```

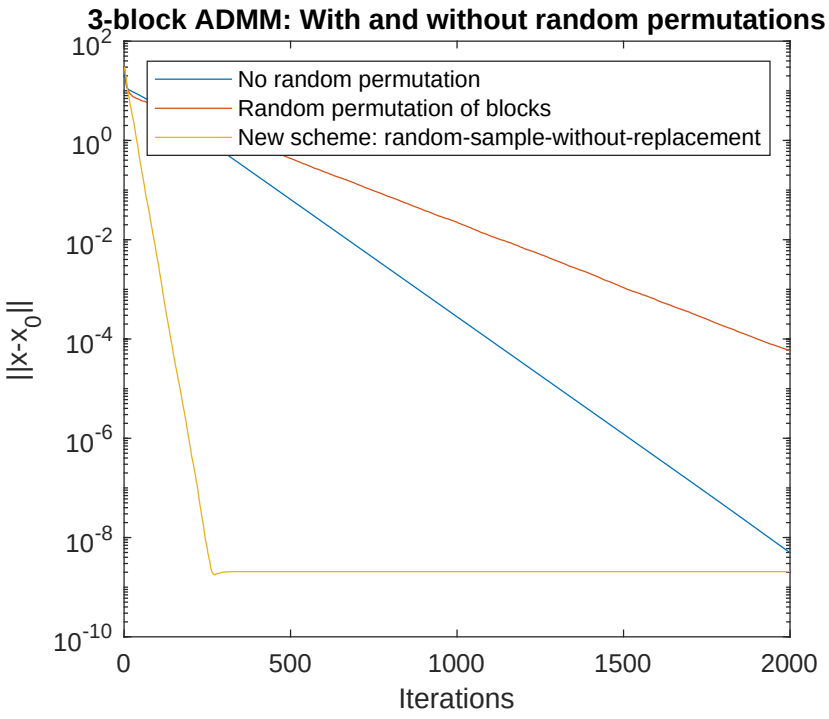


Figure 13: ADMM: Comparison of convergence speed

6 Assignment 4

Reading. Read selected sections in [LY21] Chapters 5, 6, 8, 10 and 14.

1. Recall that the (local) second-order (SO), concordant second-order (CSO) and scaled concordant second-order (SCSO) Lipschitz conditions (LC) are defined as follows:

   SOLC : $\|\nabla f(\boldsymbol{x}+\boldsymbol{d}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})\boldsymbol{d}\| \le \beta\|\boldsymbol{d}\|^2$, where $\|\boldsymbol{d}\| \le C$ for some $C > 0$

   CSOLC : $\|\nabla f(\boldsymbol{x}+\boldsymbol{d}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})\boldsymbol{d}\| \le \beta|\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x})\boldsymbol{d}|$, where $\|\boldsymbol{d}\| \le C$ for some $C > 0$,

   and

   $$\text{SCSOLC : } \|\boldsymbol{x}(\nabla f(\boldsymbol{x}+\boldsymbol{d}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})\boldsymbol{d})\| \le \beta|\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x})\boldsymbol{d}|, \quad \text{where } \|\boldsymbol{X}^{-1}\boldsymbol{d}\| \le C \text{ for some } C > 0,$$

   and $\boldsymbol{X} = \operatorname{diag}(\boldsymbol{x}) > \boldsymbol{0}$. Here we have implicitly assumed/required that $\boldsymbol{x}$ and $\boldsymbol{x}+\boldsymbol{d}$ are in the domain of f. Here the constant C should be independent of $\boldsymbol{x}$.

   For each of the following scalar functions, find the Lipschitz parameter β value of (SOLC), (CSOLC) and (SCSOLC). You can provide an upper bound on β or state that it doesn't exist.

   (a) $f(x) = \frac{1}{3}x^3 + x$, $x > 0$

   (b) $f(x) = -\log(x)$, $x > 0$.

   (c) $f(x) = x\log(x)$, $x > 0$

   Solution Basic comments:

   - The (local) here actually only means for a bounded region of $\boldsymbol{d}$ instead of arbitrary $\boldsymbol{d}$. But it's global in terms of $\boldsymbol{x}$. But we are accepting solutions that talks about local constants for $\boldsymbol{x}$.
   - By saying that you can provide an upper bound on β, we just mean that you don't need to provide the tightest β.

   The solution below is talking about **global** constants for $\boldsymbol{x}$.

   (a) $f(x) = \frac{1}{3}x^3 + x$, $x > 0$.
   Note that $f'(x) = x^2 + 1$, $f''(x) = 2x$.

The SOLC condition holds for $\beta = 1$. To see this, we observe that for all $x > 0$, and d such that $x + d > 0$,

$$|f'(x+d) - f'(x) - f''(x) \cdot d| = d^2$$

Hence $f(x)$ is 1-SOLC.

The CSOLC does not hold for any β. To see this, simply notice that the LHS is still d^2, while the RHS becomes $2|x|\beta d^2$. By taking $x \to 0$, we see that no β will satisfy the CSOLC.

The SCSOLC holds for $\beta = 1/2$. For all $x > 0$, and d such that $x + d > 0$, we have that

$$|x(f'(x+d) - f'(x) - f''(x) \cdot d)| = xd^2 = \frac{1}{2}|d^2 f''(x)|$$

Hence $f(x)$ is 1/2-SCSOLC.

(b) $f(x) = -\log(x)$, $x > 0$.

Note that $f'(x) = -x^{-1}$, $f''(x) = x^{-2}$, and that

$$|f'(x+d) - f'(x) - f''(x)d| = \frac{d^2}{x^2(x+d)}$$

The SOLC does not hold for any $\beta > 0$. To see this, simply notice that for any $d > 0$ (no matter how small it is), by taking $x \to 0^+$, the LHS goes to $+\infty$ while the RHS βd^2 remains finite, and hence no β satisfies this inequality.

The CSOLC does not hold for any $\beta > 0$. To see this, simply notice that the RHS is $\beta d^2/x^2$, and hence LHS $\leq$ RHS $\Rightarrow 1/(x+d) \leq \beta$. By taking both x and d going to 0, we see that β can not be finite.

The SCSOLC holds for $\beta = 2$ if $|x^{-1}d| \leq \frac{1}{2}$. To see that, for all $x > 0$ and d such that $|x^{-1}d| \leq 1/2$, we have $1 + \frac{d}{x} \geq \frac{1}{2}$. It follows that

$$|x(f'(x+d) - f'(x) - f''(x)d)| = \frac{d^2}{x(x+d)} = \frac{d^2}{x^2\left(1+\frac{d}{x}\right)} \leq 2\frac{d^2}{x^2} = 2|d^2 f''(x)|.$$

Hence f is 2-SCSOLC provided $|x^{-1}d| \leq \frac{1}{2}$.

(c) $f(x) = x\log(x)$, $x > 0$.

Note that $f'(x) = 1 + \log x$, $f''(x) = 1/x$, and that for any d such that $x + d > 0$,

$$|f'(x+d) - f'(x) - f''(x)d| = \frac{d}{x} - \log\left(1 + \frac{d}{x}\right).$$

Recall that $\frac{x}{1+x} \leq \log(1+x) \leq x$ for all $x > -1$.

The SOLC does not hold for any $\beta > 0$. To see this, notice that by the L'Hospital rule, we have for any fixed $x > 0$,

$$\lim_{d \to 0} \frac{|f'(x+d) - f'(x) - f''(x)d|}{d^2} = \frac{1}{2x^2},$$

which is unbounded as x goes to 0.

The CSOLC does not hold for any $\beta > 0$. To see this, again notice that by the L'Hospital rule, we have for any fixed $x > 0$,

$$\lim_{d \to 0} \frac{|f'(x+d) - f'(x) - f''(x)d|}{d^2/x} = \frac{1}{2x},$$

which is again unbounded as x goes to 0.

The SCSOLC holds for $\beta = 2$ if $|x^{-1}d| \leq \frac{1}{2}$. To see this, notice that when $|x^{-1}d| \leq 1/2$, we have

$$\frac{|x||f'(x+d) - f'(x) - f''(x)d|}{d^2/x} = \frac{|d/x - \log(1+d/x)|}{d^2/x^2} \leq 2.$$

2. Consider the following questions:

   (a) Let $\phi(\boldsymbol{y})$, where $\boldsymbol{y} \in \mathbb{R}^m$, be (regular) β-second-order (SO) Lipschitz and be δ-strongly convex, that is, for all $\boldsymbol{y}$ in the domain of ϕ, the smallest eigenvalue of $\nabla^2 \phi(\boldsymbol{y})$ is bounded below by $\delta > 0$. Prove that the function

   $$f(\boldsymbol{x}) = \phi(\boldsymbol{A}\boldsymbol{x}),$$

   where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $n \geq m$, is a constant coefficient matrix with rank m, is concordant second-order Lipschitz for all $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ is in the domain of ϕ.

   (b) Find the concordant Lipschitz bounds α for the following three functions (or show that a global constant doesn't exist):

   - $f(\boldsymbol{x}) = \frac{1}{2}(x_1 + x_2)^2$
   - $f(\boldsymbol{x}) = e^{x_1 + x_2}$
   - $f(\boldsymbol{x}) = (x_1 + x_2)\log(x_1 + x_2)$ where $x_1 + x_2 > 0$.

Solution

(a) The key is to notice that $\nabla f(\boldsymbol{x}) = \boldsymbol{A}^\top \nabla \phi(\boldsymbol{A}\boldsymbol{x})$ and $\nabla^2 f(\boldsymbol{x}) = \boldsymbol{A}^\top \nabla^2 \phi(\boldsymbol{A}\boldsymbol{x})\boldsymbol{A}$. Then since $\phi(x)$ is second-order Lipschitz, we have that for all $\boldsymbol{x}, \boldsymbol{d}$ such that $\boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}(\boldsymbol{x}+\boldsymbol{d})$ in the domain of ϕ,

$$\|\nabla \phi(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}\boldsymbol{d}) - \nabla \phi(\boldsymbol{A}\boldsymbol{x}) - \nabla^2 \phi(\boldsymbol{A}\boldsymbol{x})\boldsymbol{A}\boldsymbol{d}\| \leq \beta \|\boldsymbol{A}\boldsymbol{d}\|^2, \quad \text{where } \|\boldsymbol{A}\boldsymbol{d}\| \leq O(1)$$

Hence

$$\begin{aligned} & \|\nabla f(\boldsymbol{x}+\boldsymbol{d}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})\boldsymbol{d}\| \\ = \; & \|\boldsymbol{A}^\top (\nabla \phi(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}\boldsymbol{d}) - \nabla \phi(\boldsymbol{A}\boldsymbol{x}) - \nabla^2 \phi(\boldsymbol{A}\boldsymbol{x})\boldsymbol{A}\boldsymbol{d})\| \\ \leq \; & \|\boldsymbol{A}^\top\| \cdot \beta \|\boldsymbol{A}\boldsymbol{d}\|^2. \end{aligned}$$

Because ϕ is strongly convex, we have that for all $\boldsymbol{x}$,

$$|\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x})\boldsymbol{d}| = |(\boldsymbol{A}\boldsymbol{d})^\top \nabla^2 \phi(\boldsymbol{A}\boldsymbol{x})(\boldsymbol{A}\boldsymbol{d})| \geq \delta \|\boldsymbol{A}\boldsymbol{d}\|^2$$

It follows that

$$\|\nabla f(\boldsymbol{x}+\boldsymbol{d}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})\boldsymbol{d}\| \leq \|\boldsymbol{A}^\top\| \frac{\beta}{\delta} (\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x})\boldsymbol{d}), \quad \text{where } \|\boldsymbol{A}\boldsymbol{d}\| \leq O(1)$$

Because $\boldsymbol{A}$ is of full row rank, it is equivalent to say $\|\boldsymbol{d}\| \leq O(1)$. Hence f is concordant second-order Lipchitz.

(b) Although it's not that difficult to talk about global constants in terms of $\boldsymbol{x}$ as in Problem 1, we show how to make use of part (a) to obtain local constants.

- $f(\boldsymbol{x}) = (x_1 + x_2)^2/2$. In this case, $\boldsymbol{A}^\top = [1, 1]$, and hence $\|\boldsymbol{A}^\top\| = \sqrt{2}$, Furthermore, $\delta = 1$ and $\beta = 0$. Hence we can set $\alpha = 0$.
- $f(\boldsymbol{x}) = e^{x_1 + x_2}$. In this case, again $\boldsymbol{A}^\top = [1, 1]$ and $\|\boldsymbol{A}^\top\| = \sqrt{2}$. Furthermore, $\delta(y) = e^y$ and $\beta(y) = O(e^y)$. Hence we can set $\alpha = O(1)$. Notice that here we used the local version of (a) (see the comment above at the beginning of (b)) to obtain a global constant α.
- $f(\boldsymbol{x}) = (x_1 + x_2)\log(x_1 + x_2)$, where $x_1 + x_2 > 0$. Once again, $\boldsymbol{A}^\top = [1, 1]$ and hence $\|\boldsymbol{A}\| = \sqrt{2}$. Furthermore, $\delta = 1/y$ and $\beta(y) = O(1/y^2)$, and hence we can choose $\alpha = O(1/(x_1 + x_2))$.
  Remark: Globally, by computing the LHS and RHS exactly, we can easily see that it's not CSOLC by taking $x_1 + x_2 \to \infty$.

3. Prove the logarithmic approximation lemma for SDP. Let $\boldsymbol{D} \in \mathbb{S}^n$ and $\|\boldsymbol{D}\|_2 < 1$. Then,

$$\mathrm{Tr}(\boldsymbol{D}) \geq \log\det(\boldsymbol{I}+\boldsymbol{D}) \geq \mathrm{Tr}(\boldsymbol{D}) - \frac{\|\boldsymbol{D}\|_F^2}{2(1-\|\boldsymbol{D}\|_2)}$$

where for any given symmetric matrix $\boldsymbol{D}$, $\|\boldsymbol{D}\|_F^2$ is the sum of all its squared eigenvalues, and $\|\boldsymbol{D}\|_2$ is its largest absolute eigenvalue.

Hint: $\det(\boldsymbol{I}+\boldsymbol{D})$ equals the product of the eigenvalues of $\boldsymbol{I}+\boldsymbol{D}$. Then the proof follows from Taylor's expansion.

Solution Suppose that the eigenvalues of $\boldsymbol{D}$ are λ_j, $j = 1, \ldots, n$. Then we have

$$\log\det(\boldsymbol{I}+\boldsymbol{D}) = \sum_{j=1}^{n} \log(1+\lambda_j) \leqslant \sum_{j=1}^{n} \lambda_j = \mathrm{Tr}(\boldsymbol{D}) \tag{31}$$

and

$$\mathrm{Tr}(\boldsymbol{D}) - \frac{\|\boldsymbol{D}\|_F^2}{2(1-\|\boldsymbol{D}\|_\infty)} = \sum_{j=1}^{n} \lambda_j - \frac{\sum_{j=1}^{n}\lambda_j^2}{2(1-\max_j|\lambda_j|)} \leqslant \sum_{j=1}^{n}\lambda_j - \sum_{j=1}^{n}\frac{\lambda_j^2}{2(1-|\lambda_j|)} \tag{32}$$

Hence it suffices to prove $\log(1+x) \geqslant x - \frac{x^2}{2(1-|x|)}$ for all $|x| < 1$.

To this end, note that by Taylor's expansion, we have $\log(1+x) = x - x^2/2 + x^3/3 - x^4/4 + x^5/5 - \ldots$. On the other hand, we also have $\frac{x^2}{2(1-|x|)} = \frac{x^2}{2}(1+|x|+|x|^2+\ldots) = x^2/2 + |x|^3/2 + |x|^4/2 + \cdots \geq x^2/2 - x^3/3 + x^4/4 - \ldots$. Comparing term by term, we immediately see that $\log(1+x) \geq x - \frac{x^2}{2(1-|x|)}$, which completes our proof.

References

[DW17] Dmitriy Drusvyatskiy and Henry Wolkowicz. The many faces of degeneracy in conic optimization. *Foundations and Trends® in Optimization*, 3(2):77–170, 2017.

[GVL13] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

[LY21] David G Luenberger and Yinyu Ye. *Linear and nonlinear programming*, volume 228 of *International Series in Operations Research & Management Science*. Springer Cham, 5th edition, 2021.