EECS 127/227AT Optimization Models in Engineering UC Berkeley Fall 2023

Homework 13

This homework is optional, with no corresponding Gradescope portal. Solutions will be released on Wednesday, December 6.

1. **Spectrahedron**

   This question explores the structure of the feasibility set associated to a linear matrix inequality. The feasibility set of any semidefinite program is the feasibility set of a linear matrix inequality, so this question aimed at developing a better understanding of the feasibility sets of semidefinite programs. Given symmetric matrices $F_0, F_1, \ldots, F_m \in \mathbb{S}^n$, the set of symmetric matrices

   $$\{F_0 + x_1 F_1 + \ldots + x_m F_m : \vec{x} \in \mathbb{R}^m\},$$

   where $\vec{x}$ denotes $\begin{bmatrix} x_1 & \ldots & x_m \end{bmatrix}^\top$, is called a *linear matrix pencil*. It is an affine subspace of the vector space $\mathbb{S}^n$. We write $F(\vec{x})$ for $F_0 + \sum_{i=1}^m x_i F_i$. The intersection of a linear matrix pencil with the cone of symmetric positive semidefinite matrices is called a *spectrahedron*. The condition that needs to be satisfied for this, namely

   $$F(\vec{x}) \succeq 0,$$

   is called a *linear matrix inequality*. The term “spectrahedron" is also used to refer to

   $$\{\vec{x} \in \mathbb{R}^m : F(\vec{x}) \succeq 0\},$$

   in which case we think of it as a subset of $\mathbb{R}^m$. This set is also called the feasibility set of the LMI. In this question we will study the spectrahedron associated to the linear matrix pencil

   $$F(x, y) = \begin{bmatrix} 1 & 1-x & -x \\ 1-x & 1 & -y \\ -x & -y & 2y \end{bmatrix}.$$

   We will think of this spectrahedron as a subset of $\mathbb{R}^2$, with the vectors in $\mathbb{R}^2$ being written as $\begin{bmatrix} x & y \end{bmatrix}^\top$.

   (a) Find all the principal minors of $F(x, y)$. **Remark**: For a square $n \times n$ matrix A, for each $1 \leq k \leq n$ there are $\binom{n}{k}$ principal k-minors. These are found by picking a subset $J \subset \{1, \ldots, n\}$ of size k and considering the $k \times k$ matrix one gets from A by erasing all the rows with index not in J and all the columns with index not in J and then taking the determinant of this $k \times k$ matrix. Since $F(x, y)$ is a 3×3 matrix, there will be three principal 1-minors (which are just the diagonal entries), three principal 2-minors, and one principal 3-minor (which is just the determinant). You are asked to find these.

   (b) For a symmetric matrix $A \in \mathbb{S}^n$, it is known that A is symmetric positive semidefinite if and only if all its principal minors are nonnegative. Based on this, find a finite collection of polynomials in the variables (x, y), such that the spectrahedron can be defined as the set of (x, y) values where these polynomials are non-negative. A graph of the spectrahedron is shown in Figure 1.

   (c) Show that if $F(x, y)$ is symmetric positive semidefinite then $0 \leq x \leq 2$.

   (d) Suppose $x = 0$. Show that the only value of y for which $F(0, y)$ is symmetric positive semidefinite is $y = 0$. What is the rank of $F(0, 0)$?

   (e) Suppose $x = 2$. Show that there is no value of y for which $F(2, y)$ is symmetric positive semidefinite.

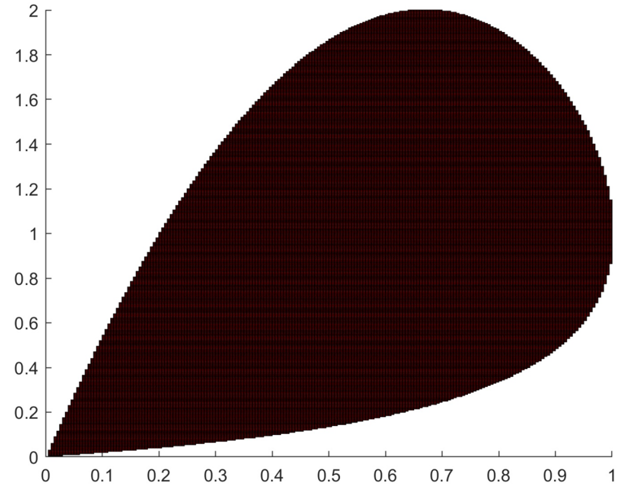

Figure 1: Spectrahedron. This is a convex set.

(f) What condition must the pair (x, y) satisfy for the rank of $F(x, y)$ to be strictly less than 3?

(g) Assume $0 < x < 2$. Verify that this makes the leading 2×2 block of $F(x, y)$ symmetric positive definite. Fixing x, apply the Schur complement criterion to show that $F(x, y)$ is symmetric positive semidefinite if and only if the determinant of $F(x, y)$ is nonnegative and symmetric positive definite if and only if the determinant of $F(x, y)$ is strictly positive.

Remark: From this part of the question we learn that the boundary of the spectrahedron in Figure 1 consists of pairs (x, y) where the rank of $F(x, y)$ is strictly less than 3.

(h) Show that for (x, y) in the spectrahedron, the rank of $F(x, y)$ is 1 precisely when $(x, y) = (0, 0)$.

Remark: From this part and the preceding part of the question we learn that the rank of $F(x, y)$ is 3 on the interior of the spectrahedron, is 2 at all points on the boundary of the spectrahedron except the point $(x, y) = (0, 0)$. where the rank of $F(0, 0)$ is 1.

2. Soft-Margin SVM

Consider the soft-margin SVM problem,

$$p^\star(C) = \min_{\vec{w}\in\mathbb{R}^m, b\in\mathbb{R}, \vec{\xi}\in\mathbb{R}^n} \quad \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i \tag{1}$$

$$\text{s.t.} \quad 1 - \xi_i - y_i(\vec{x}_i^\top \vec{w} - b) \leq 0, \quad i = 1, 2, \ldots, n \tag{2}$$

$$-\xi_i \leq 0, \quad i = 1, 2, \ldots, n, \tag{3}$$

where $\vec{x}_i \in \mathbb{R}^m$ refers to the i^{th} training data point, $y_i \in \{-1, 1\}$ is its label, and $C \in \mathbb{R}_+$ (i.e. $C > 0$) is a hyperparameter. Let α_i denote the dual variable corresponding to the inequality $1 - \xi_i - y_i(\vec{x}_i^\top \vec{w} - b) \leq 0$ and let β_i denote the dual variable corresponding to the inequality $-\xi_i \leq 0$. The Lagrangian is then given by

$$\mathcal{L}(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \beta) = \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - \xi_i - y_i(\vec{x}_i^\top \vec{w} - b)) - \sum_{i=1}^{n}\beta_i\xi_i. \tag{4}$$

Suppose $\vec{w}^\star, b^\star, \vec{\xi}^\star, \vec{\alpha}^\star, \beta^\star$ satisfy the KKT conditions. Classify the following statements as true or false and justify your answers mathematically.

(a) Suppose the optimal solution $\vec{w}^\star, b^\star$ changes when the training point $\vec{x}_i$ is removed. Then originally, we necessarily have $y_i(\vec{x}_i^\top \vec{w}^\star - b^\star) = 1 - \xi_i^\star$.

(b) Suppose the optimal solution $\vec{w}^\star, b^\star$ changes when the training point $\vec{x}_i$ is removed. Then originally, we necessarily have $\alpha_i^\star > 0$.

(c) Suppose the data points are strictly linearly separable, i.e. there exist $\tilde{\vec{w}}$ and $\tilde{b}$ such that for all i,

$$y_i(\vec{x}_i^\top \tilde{\vec{w}} - \tilde{b}) > 0. \tag{5}$$

Then $p^\star(C) \to \infty$ as $C \to \infty$.

3. Gradient Descent on a Graph

This question studies a gradient descent method to solve an optimization problem where the variables are thought of as being parametrized by the vertices of an undirected graph. We are given an undirected simple graph $G = (V, E)$ where $V = \{1, \cdots, n\}$ is the set of vertices and $E \subseteq V \times V$ is the set of edges. Note that the notation is such that if $(i, j) \in E$ then we will also have $(j, i) \in E$. In particular, E will be of even cardinality.

(a) Consider the problem of assigning weights x_i to each vertex $i \in V$ such that adjacent vertices get similar weights, and the sum of weights is close to 1. That is, we want the solution to the optimization problem

$$\vec{x}^\star := \arg\min_{\vec{x}\in\mathbb{R}^n} \sum_{(i,j)\in E} (x_i - x_j)^2 + 2\lambda \left(\sum_{i\in V} x_i - 1\right)^2$$

where $\lambda \geq 0$ is a constant. Show that this optimization problem is equivalent to

$$\vec{x}^\star = \arg\min_{\vec{x}\in\mathbb{R}^n} \frac{1}{2}\vec{x}^\top (L + \lambda\vec{\mathbf{1}}\vec{\mathbf{1}}^\top)\vec{x} - \lambda\vec{\mathbf{1}}^\top\vec{x}$$

where L is the Laplacian matrix for G and $\vec{\mathbf{1}}$ is the all-ones vector in $\mathbb{R}^n$.

Note: Let $G = (V, E)$ be a graph with $V = \{1, \cdots, n\}$. For each vertex $i \in V$, the *degree* of i, denoted deg(i), is defined as the number of edges incident to the vertex i. The Laplacian matrix L of G is the $n \times n$ matrix with i-th diagonal entry deg(i) and with (i, j)-th entry (with $i \neq j$) equal to -1 if $(i, j) \in E$ and equal to zero otherwise, for each $i, j \in V$.

Hint: For all $\vec{x} \in \mathbb{R}^n$ we have $\vec{x}^\top L\vec{x} = \frac{1}{2}\sum_{i,j\in V:(i,j)\in E}(x_i - x_j)^2$.

(b) What is the optimal $\vec{x}^\star$?

(c) Suppose we use gradient descent with step size $\eta > 0$ to find the optimal $\vec{x}^\star$. Write the gradient descent step; i.e., express $\vec{x}_{k+1}$, the $(k + 1)$th step of gradient descent, in terms of $\vec{x}_k$, L, η, and λ.

(d) Show that $\vec{x}_{k+1} - \vec{x}^\star = (I - \eta(L + \lambda\vec{\mathbf{1}}\vec{\mathbf{1}}^\top))(\vec{x}_k - \vec{x}^\star)$.

(e) We saw that for all $\vec{x} \in \mathbb{R}^n$ we have $\vec{x}^\top L\vec{x} = \frac{1}{2}\sum_{i,j\in V:(i,j)\in E}(x_i - x_j)^2$. Hence L is symmetric positive semidefinite. Also $L\vec{\mathbf{1}} = \vec{0}$, so the smallest eigenvalue of L is 0. Let $\lambda_1 \geq \cdots \geq \lambda_n = 0$ be the eigenvalues of L, and assume λ is given such that $\lambda_1 \geq n\lambda \geq \lambda_{n-1}$. Show that $\|\vec{x}_k - \vec{x}^\star\|_2 \leq \rho^k\|\vec{x}_0 - \vec{x}^\star\|_2$ for $\rho := \max\{|1 - \eta\lambda_{n-1}|, |1 - \eta\lambda_1|\}$, where $\vec{x}_0$ is the starting point of the gradient descent.

(f) Assuming that $\eta > 0$ is small enough that $0 < \rho < 1$, find the number of time steps needed to converge to some $\varepsilon > 0$ around $x^\star$ as a function of η, assuming $\|\vec{x}_0 - \vec{x}^\star\|_2 > \varepsilon$. That is, find $t(\eta)$ such that $\|\vec{x}_k - \vec{x}^\star\|_2 \leq \varepsilon$ for $k \geq t(\eta)$.

(g) Find the optimal step size, i.e. the solution to

$$\eta^\star = \arg\min_{\{\eta>0:0<\rho<1\}} t(\eta).$$

What is the corresponding $t(\eta^\star)$?

4. **Newton's Method, Coordinate Descent and Gradient Descent**

In this question, we will compare three different optimization methods: Newton's method, coordinate descent and gradient descent. We will consider the simple set-up of unconstrained convex quadratic optimization; i.e we will consider the following problem:

$$\min_{\vec{x}\in\mathbb{R}^d} \vec{x}^\top A\vec{x} - 2\vec{b}^\top \vec{x} + c \tag{6}$$

where $A \succ 0$ and $\vec{b} \in \mathbb{R}^d$.

(a) How many steps does Newton's method take to converge to the optimal solution? Recall that the update rule for Newton's method is given by the equation:

$$\vec{x}_{t+1} = \vec{x}_t - (\nabla^2 f(\vec{x}_t))^{-1}\nabla f(\vec{x}_t). \tag{7}$$

when optimizing a function f.

(b) Now, consider the simple two variable quadratic optimization problem for $\sigma > 0$:

$$\min_{\vec{x}\in\mathbb{R}^2} f(\vec{x}) = \sigma x_1^2 + x_2^2. \tag{8}$$

How many steps does coordinate descent take to converge on this problem? Assume that we start by updating the variable x_1 in the first step, x_2 in step two and so on; therefore, we will update x_1 and x_2 in odd and even iterations respectively:

$$(x_{t+1})_1 = \begin{cases} \operatorname{argmin}_{x_1} f(x_1, (x_t)_2) & \text{for odd t} \\ (x_t)_1 & \text{otherwise} \end{cases} \text{ and } (x_{t+1})_2 = \begin{cases} \operatorname{argmin}_{x_2} f((x_t)_1, x_2) & \text{for even t} \\ (x_t)_2. & \text{otherwise} \end{cases} \tag{9}$$

Here, $(x_t)_2$ represents x_2 at time t and so on.

(c) We will now analyze the performance of coordinate descent on another quadratic optimization problem:

$$\min_{\vec{x}\in\mathbb{R}^2} f(\vec{x}) = \sigma(x_1 + x_2)^2 + (x_1 - x_2)^2. \tag{10}$$

where we have, as before, $\sigma > 0$. Note that $(0, 0)$ is the optimal solution to this problem. Now, starting from the point $\vec{x}_0 = (1, 1)$, write how each coordinate of $(\vec{x}_{t+1})_i$ relates to $(\vec{x}_t)_i$ for $i = 1, 2$. Use this to show how the algorithm converges from the initial point $(1, 1)$ to $(0, 0)$. What happens when σ grows large? *HINT: First find the update rule for $(\vec{x}_t)_1$, i.e. keep $(\vec{x}_t)_2$ fixed and figure out how $(\vec{x}_t)_1$ changes when t is odd. Then do the same for $(\vec{x}_t)_2$ when $(\vec{x}_t)_1$ is fixed for even t.*

(d) Finally, for the objective function (10), write an equation relating $\vec{x}_t$ to $\vec{x}_0$ if $\vec{x}_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Assume for this part that $\sigma > 1$ and reason about how quickly gradient descent converges when σ grows large. *HINT: What is the optimal step size for gradient descent, using the previous part? HINT: Also note that f is given by:*

$$f(\vec{x}) = \vec{x}^\top A\vec{x} \textit{ where } A = 2\left(\sigma \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} + \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}\right). \tag{11}$$

5. Gradient Descent vs Newton Method

Run the Jupyter notebook `gradient_vs_newton.ipynb` which demonstrates differences between gradient descent and Newton's method.