EECS 127/227AT **Fall 2020**
Homework 10
Homework 10 is due on Gradescope by Friday 11/20 at 11.59 p.m.

1 An SDP with an infeasible dual

This question is an exercise in the structure of SDP duality. The relevant sections of the textbooks are Secs. 11.3.1 and 11.3.2 from the textbook of Calafiore and El Ghaoui and Sec. 4.6.2 from the textbook of Boyd and Vandenberghe.

Consider the following SDP in inequality form, which we call the primal SDP

$$\begin{aligned} p^* := \min_{x,y} \; & x \\ \text{s.t.} \; & \begin{bmatrix} 1 & x+y \\ x+y & y \end{bmatrix} \succeq 0. \end{aligned}$$

(a) Show that the primal SDP is strictly feasible.

(b) Find the conic dual of the primal SDP. This will be an SDP in standard form.

(c) Show that the dual SDP is infeasible.

(d) Since the primal SDP is strictly feasible we know that strong duality holds. However, the dual SDP is infeasible. What does this say about the primal value p^*? Can you directly justify this?

2 Lovász theta function of an undirected graph

This question is connected with the use of SDPs as relaxations for combinatorial optimization problems and is also connected with SDP duality. The relevant sections of the textbooks are Sec. 11.3.1 and 11.3.2 of the textbook of Calafiore and El Ghaoui and Sec. 4.6.2 of the textbook of Boyd and Vandenberghe.

The Lovász theta function of an undirected graph, introduced by Lázló Lovász, plays a role in computer science related to the theme of embedding graphs into Euclidean space in order to bring optimization theory techniques to bear on combinatorial questions.

Let G be an undirected graph with vertex set $V = \{1, \ldots, n\}$ and edge set E. Recall that E is a set of unordered pairs of vertices. We do not allow self loops. The Lovász theta function of G, denoted $\theta(G)$, is the optimal value of the following SDP in standard form (note that it is a maximization problem).

$$\begin{aligned} \theta(G) := \max_{X \in \mathbb{S}^n} \ & \text{trace}(\mathbb{1}\mathbb{1}^T X) \\ \text{s.t.} \ & \text{trace}(X) = 1, \\ & x_{ij} = 0, \ \text{ for all } (i,j) \in E, \\ & X \succeq 0. \end{aligned}$$

Here $\mathbb{1}$ denotes the column vector of all ones and X has entries x_{ij}. To recognize that this is actually an SDP in equality form, define the matrices $B_{ij} \in \mathbb{S}^n$ for $1 \leq i < j \leq n$ which are zero everywhere except for 1 in the (i,j) and (j,i) locations. Then the condition $x_{ij} = 0$ for $(i,j) \in E$ can be written as $\text{trace}(B_{ij}X) = 0$ (this is imposed only for $(i,j) \in E$) and the condition $\text{trace}(X) = 1$ is of course the same as $\text{trace}(I_n X) = 1$, where I_n is the $n \times n$ identity matrix.

(a) The *independence number* of the graph G, also called its *stability number*, and denoted $\alpha(G)$, is the cardinality of the largest *independent set*, also called *stable set*, in the graph. Here, a nonempty subset of vertices of the graph is called an independent set if there is no edge between any pair of vertices in that set. The decision problem of deciding whether the independence number of a graph is a given value is believed to be a hard computational problem.

Show that we have $\alpha(G) \leq \theta(G)$. Thus the Lovász theta function of the graph, which can be efficiently computed because it is the solution of an SDP, is an upper bound for the independence number of the graph.

Hint: Given an independent set S, consider the vector $x_S \in \mathbb{R}^n$ which equals $\frac{1}{\sqrt{|S|}}$ at the elements in S and zero elsewhere (here $|S|$ denotes the cardinality of S). Use this vector to find a feasible point for the SDP defining the Lovász theta function.

(b) Find the dual of the SDP that defines $\theta(G)$. This will be an SDP in inequality form. Show that it can be written as

$$\begin{aligned} \min_{t, (\nu_e, e \in E)} \ & t \\ \text{s.t.} \ & tI_n \succeq \mathbb{1}\mathbb{1}^T + \sum_{e \in E} \nu_e B_e. \end{aligned}$$

(c) Show that strong duality holds.

(d) Using strong duality, show that $\theta(G)$ can be characterized as the solution of the following problem

$$\begin{aligned} \min_{Y \in \mathbb{S}^n} \ & \lambda_{\max}(Y) \\ \text{s.t} \ & y_{ii} = 1, \ i = 1, \ldots, n, \\ & y_{ij} = 1, \ \text{ for all } (i,j) \notin E. \end{aligned}$$

Here the entries of Y are denoted y_{ij} and $\lambda_{\max}(Y)$ denotes the largest eigenvalue of Y (since $Y$is a symmetric matrix, all its eigenvalues are real numbers). Note that Y is not required to be positive semidefinite in this problem formulation. Nevertheless, since the trace of Y will have to be n, its largest eigenvalue will be positive (and in fact bigger than or equal to 1).

Hint: Consider dual feasible points, using the notation y_e for $1 + \nu_e$ for $e \in E$.

3 Markov Chain Mixing Time

This question aims to formulate a problem related to the convergence time of Markov chains as an SDP. The relevant sections of the textbooks are Sec. 11.3.1 of the textbook of Calaifore and El Ghaoui and Sec. 4.6.2 of the textbook of Boyd and Vandenberghe.

Consider a directed graph $G = (V, E)$, where $V = \{1, \ldots, n\}$ are the vertices and $E \subseteq V \times V$ are the directed edges. We allow for self loops. A *Markov chain* is a stochastic process (X_t) where if $(i, j) \in E$ then

$$\Pr(X_{t+1} = j \mid X_t = i) = P_{ij}$$

for some probability $P_{ij} \in [0, 1]$ associated to the directed edge (i, j). In order for these probabilities to make sense, the matrix $P \in \mathbb{R}^{n \times n}$ with value P_{ij} at the ith row and jth column must satisfy

$$\begin{aligned} P_{ij} &\geq 0 \qquad \forall i, j \in V, \\ P\mathbb{1} &= \mathbb{1}, \end{aligned}$$

where $\mathbb{1}$ denotes the all ones vector. We also enforce

$$P_{ij} = 0 \qquad \forall (i, j) \notin E,$$

so that the probability of moving to a non-adjacent vertex at any time step is zero, as well as

$$P^\top = P,$$

which will ensure that all the eigenvalue of P are real numbers.

The probability distribution for each X_t can be described by a vector $x_t \in \mathbb{R}^n$ such that $\mathbb{1}^\top x_t = 1$ and $x_t \geq 0$. That is, the value $(x_t)_i$ is the probability that $X_t = i$. From the definition of P, we have

$$x_t^\top = x_0^\top P^t,$$

for all $t \geq 0$. Here P^0 is the identity matrix I.

(a) Let $\lambda_1 \geq \ldots \geq \lambda_n$ be the eigenvalues of P. Show that

$$\lambda_1 = 1.$$

(b) We say that a nonnegative vector $x \in \mathbb{R}^n_+$ is a *stationary distribution* of the Markov chain if $x^\top P = x^\top$ and $x^\top \mathbb{1} = 1$. Show that $\frac{1}{n}\mathbb{1}$ is a stationary distribution for the Markov chain under our assumptions.

(c) Show that

$$\max\{\lambda_2, |\lambda_n|\} = \|P - (1/n)\mathbb{1}\mathbb{1}^\top\|_2.$$

(d) As justified in the optional addendum at the end of this question, $\max\{\lambda_2, |\lambda_n|\}$ is related to how fast the Markov chain converges to its stationary distribution from an arbitrary initial distribution, i.e. what one might roughly call its *mixing time*. The smaller $\max\{\lambda_2, |\lambda_n|\}$ is, the smaller the mixing time of Markov chain. Suppose we wish to minimize $\max\{\lambda_2, |\lambda_n|\}$ (and thus try to minimize the mixing time) over all Markov chains on the graph G. Show that this problem can be expressed as an SDP in standard form.

Hint: Introduce slack variables. Also, express the constraint $\|P - (1/n)\mathbb{1}\mathbb{1}^\top\|_2 \leq t$ via linear matrix inequalities.

Addendum (optional)

One way to define the *mixing time* of a Markov chain is as the first time τ at which we have

$$\|P^\tau x_0 - (1/n)\mathbb{1}\|_1 \leq 1/4$$

for all $x_0 \in \mathbb{R}^n$ such that $\mathbb{1}^\top x_0 = 1$ and $x_0 \geq 0$. Namely, irrespective of what the initial probability distribution is, by the time τ the distribution of the Markov chain will be within some fixed ℓ_1 distance (here $\frac{1}{4}$) of its stationary distribution (which we recall is the uniform distribution, under our assumptions).

From the Cauchy-Schwarz inequality, we have $\|v\|_1 \leq \sqrt{n}\|v\|_2$ for any vector v. Hence we can find an upper bound for the mixing time, as defined above, by finding an upper bound for the first τ such that

$$\|P^\tau x_0 - \frac{1}{n}\mathbb{1}\|_2^2 \leq \frac{1}{16n}.$$

Since P is a symmetric matrix under our assumptions, it has an orthonormal basis $\{v_1, \ldots, v_n\}$ of eigenvectors, one of which is $\frac{1}{\sqrt{n}}\mathbb{1}$, corresponding to the eigenvalue 1, and which we take to be v_1. Let us write the initial distribution x_0 in this eigenbasis of P as

$$x_0 = \sum_{i=1}^n a_i v_i$$

Since x_0 is a probability distribution, we have $(1/\sqrt{n})\mathbb{1}^\top x_0 = 1/\sqrt{n}$, and so $a_1 = 1/\sqrt{n}$. Also, since x_0 defines a probability distribution, we have

$$\|x_0\|_2 \leq 1,$$

and so

$$\sum_{i=1}^n a_i^2 \leq 1.$$

Thus, since $P\mathbb{1} = \mathbb{1}$, we have

$$\begin{aligned}
\|P^\tau x_0 - (1/n)\mathbb{1}\|_2^2 &= \|P^\tau(x_0 - (1/n)\mathbb{1})\|_2^2 \\
&= \|P^\tau(\sum_{i=2}^n a_i v_i)\|_2^2 \\
&= \sum_{i=2}^n (a_i \lambda_i^\tau)^2.
\end{aligned}$$

Since $|\lambda_i| \leq \max(\lambda_2, |\lambda_n|)$ for all $2 \leq i \leq n$, we can now write, using $\sum_{i=1}^n a_i^2 \leq 1$, that

$$\begin{aligned}
\|P^\tau x_0 - (1/n)\mathbb{1}\|_2^2 &= \sum_{i=2}^n (a_i \lambda_i^\tau)^2 \\
&\leq \sum_{i=2}^n a_i^2 (\max(\lambda_2, |\lambda_n|))^{2\tau} \\
&\leq (\max(\lambda_2, |\lambda_n|))^{2\tau},
\end{aligned}$$

which will be bounded above by $\frac{1}{16n}$ if we set

$$\tau \geq -\frac{\log 16n}{2 \log \max(\lambda_2, |\lambda_n|)}.$$

This explains why making $\max(\lambda_2, |\lambda_n|)$ small can be thought of as making the mixing time small.
Remark: Recall from HW 3 that λ_2 is related to *how connected* the graph is; graphs with small λ_2 are better connected and it is intuitively reasonable that mixing would happen faster in a well-connected graph. The eigenvalue λ_n turns out to be related to *how bipartite* a graph is. Under our assumptions the stationary distribution of the Markov chains is the uniform distribution. If the graph is bipartite then the probability distribution of each side alternates between the two sides, so τ could be unbounded if the initial distribution is sufficiently unbalanced between the two sides of the bipartite graph. This corresponds to the case where $\lambda_n = -1$.

4 A trust-region problem

In this question we look at an SDP relaxation for a nonconvex quadratic optimization problem that shows up as as subproblem in many numerical algorithms. The relevant sections of the textbooks are Sec. 11.3.3 of the textbook of Calafiore and El Ghaoui and Sec. 5.3 of the textbook of Boyd and Vandenberghe.

Consider the problem of minimizing a quadratic function over an Euclidean ball, i.e.

$$\begin{array}{rll} p^* := & \min\limits_{x\in\mathbb{R}^n} & \frac{1}{2}x^\top Hx + c^\top x + d \\ & \text{s.t.} & x^\top x \le r^2, \end{array}$$

where $H \in \mathbb{S}^n$ and $r > 0$ is the radius of the given ball. This problem is called the *trust-region* problem, and often arises in a context where the quadratic function being minimized is a second-order approximation over an Euclidean ball to some more complicated function. Solving this problem gives an idea of whether the approximation is being used over too large a ball.

(a) Suppose for this part only that $H \succ 0$. Prove that the optimal solution to the trust-region problem is unique and is given by

$$x(\lambda^*) = -(H + \lambda^* I)^{-1}c,$$

where $\lambda^* = 0$ if $\|H^{-1}c\|_2 \le r$, and is otherwise the unique value λ^* such that

$$\|(H + \lambda^* I)^{-1}c\|_2 = r.$$

(b) We will now assume that $H \not\succeq 0$. The trust-region problem will then be a nonconvex optimization problem. However, we can relax it to the SDP

$$\begin{array}{rll} q^* = & \max\limits_{X\in\mathbb{S}^n, x\in\mathbb{R}^n} & \frac{1}{2}\left\langle \begin{bmatrix} H & c \\ c^\top & 2d \end{bmatrix}, \begin{bmatrix} X & x \\ x^\top & 1 \end{bmatrix} \right\rangle \\ & \text{s.t.} & \operatorname{tr}(X) \le r^2, \\ & & \begin{bmatrix} X & x \\ x^\top & 1 \end{bmatrix} \succeq 0, \end{array} \tag{1}$$

where we have used the notation $\langle A, B\rangle$ as an alternate notation for $\operatorname{tr}(AB)$ for symmetric matrices A and B of the same dimension.

Show that this is indeed a relaxation, i.e., $q^* \le p^*$.

5 Gradient Descent and Pseudoinverse

This question studies a simple gradient descent algorithm with constant step size, attempting to find a minimizer for a quadratic function. The relevant sections of the textbooks are Secs. 12.1, 12.2.1 and 12.2.2 of the textbook of Calafiore and El Ghaoui and Secs. 9.1 and 9.2 of the textbook of Boyd and Vandenberghe.

Consider the problem

$$\arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2, \tag{2}$$

where $A \in R^{m\times n}$ is assumed to have full row rank and $m < n$.

We know that if we write an SVD for A it will have the form

$$A = U\Sigma V^T = U\begin{bmatrix}\Sigma_1 & 0\end{bmatrix}\begin{bmatrix}V^T_{\mathcal{R}(A^T)}\\ V^T_{\mathcal{N}(A)}\end{bmatrix},$$

where U is an orthogonal $m \times m$ matrix, Σ_1 is a diagonal $m \times m$ matrix with strictly positive diagonal entries, and V is an orthogonal $n \times n$ matrix. Further, the last $n - m$ columns of V, i.e. the columns of $V_{\mathcal{N}(A)}$, form an orthonormal basis for $\mathcal{N}(A)$ and the first m columns of V, i.e. the columns of $V_{\mathcal{R}(A^T)}$, form an orthonormal basis for $\mathcal{R}(A^T)$. Further, we know that the minimum ℓ_2 norm solution of the problem in (2) is given by

$$x^* = A^\dagger b = V_{\mathcal{R}(A^T)}\Sigma_1^{-1}U^T b,$$

where $A^\dagger = V_{\mathcal{R}(A^T)}\Sigma_1^{-1}U^T$ denotes the Moore-Penrose pseudoinverse of A.

Suppose that m, n are very large and computing the SVD and performing the matrix multiplications is quite costly. So you wonder, is there a better way for me to solve this problem?

One idea is to use gradient descent, assuming that you have a black box that can compute the gradient very efficiently. The gradient descent algorithm to minimize a function f starts at x_0 and iteratively computes x_k via the equation

$$x_{k+1} = x_k - \eta\nabla_x f(x_k),$$

where $\eta > 0$ is called the *step size*.

Your good friend Gireeja tells you that you can use gradient descent on the function: $f(x) := \frac{1}{2}\|Ax - b\|_2^2$ and that as long as your initial point x_0 is orthogonal to $\mathcal{N}(A)$ this algorithm will converge to the minimum norm solution if the step size $\eta > 0$ is sufficiently small.

Is Gireeja correct in making this claim? If so, provide a proof. Otherwise, provide a counterexample.

Hint: Gireeja is usually right.