# Homework 5

## Convex Optimization 10-725/36-725

**Due Friday December 4 at 4:00pm**
**submitted to Mallory Deptola in GHC 8001**
(Remember to submit each problem on a separate sheet of paper, with your name on at the top)

## 1 Nonconvex, but still strong [Shashank]

*This question is from B&V Additional Exercises 4.6.* Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) = x^T A x + 2b^T x \\ \text{subject to} & x^T x \leq 1 \end{array} \tag{1}$$

with variable $x \in \mathbb{R}^n$, and data $A \in \mathbb{S}^n$, $b \in \mathbb{R}^n$. We do not assume that $A$ is positive semidefinite, and therefore the problem is not necessarily convex. In this exercise we show that $x$ is (globally) optimal if and only if there exists a $\lambda$ such that

$$\|x\|_2 \leq 1, \;\; \lambda \geq 0, \;\; A + \lambda I \succeq 0, \;\; (A + \lambda I)x = -b, \;\; \lambda(1 - \|x\|_2^2) = 0 \tag{2}$$

From this we will develop an efficient method for finding the global solution. The conditions (2) are the KKT conditions for (1) with the inequality $A + \lambda I \succeq 0$ added.

(a) Show that if $x$ and $\lambda$ satisfy (2), then $f(x) = \inf_{\widetilde{x}} L(\widetilde{x}, \lambda) = g(\lambda)$, where $L$ is the Lagrangian of the problem and $g$ is the dual function. Therefore strong duality holds, and $x$ is globally optimal.

(b) Next show that the conditions (2) are also necessary. Assume that $x$ is globally optimal for (1). Distinguish the two cases:

(i) $\|x\|_2 < 1$. Show that (2) holds with $\lambda = 0$. *(Hint: If the constraint is inactive at the solution, the function has a local minima at $x$.)*

(ii) $\|x\|_2 = 1$. First prove that $(A + \lambda I)x = -b$ for some $\lambda \geq 0$. It then remains to show that $A + \lambda I \succeq 0$. If not, argue that there exists a $w$ with $w^T(A + \lambda I)w < 0$ such that $w^T x \neq 0$. Show that for such a $w$, the point $y = x - 2\frac{w^T x}{w^T w} w$ satisfies $\|y\|_2 = 1$ and $f(y) < f(x)$.

(c) The optimality conditions (2) can be used to derive a simple algorithm for (1). Using the eigenvalue decomposition $A = \sum_{i=1}^n \alpha_i q_i q_i^T$, of $A$, we make a change of variables $y_i = q_i^T x$, and write (1) as

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{i=1}^n \alpha_i y_i^2 + 2\sum_{i=1}^n \beta_i y_i \\ \text{subject to} & y^T y \leq 1 \end{array} \tag{3}$$

where $\beta_i = q_i^T b$. The transformed optimality conditions (2) are

$$\|y\|_2 \leq 1, \quad \lambda \geq -\alpha_n, \quad (\alpha_i + \lambda) y_i = -\beta_i, \; i = 1, \ldots, n, \quad \lambda(1 - \|y\|_2) = 0$$

if we assume that $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_n$. Give an algorithm for computing the solution $y$ and $\lambda$.

## 2 Safe screening rules for the lasso [Hanzhang]

In this problem we are will derive the SAFE screening rule for Lasso.

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

A screen rule enables us to perform some efficient computation to know that some dimensions $\beta_j$ has to be 0 in the optimal solution, without actually solving the optimization.

(a) Let $f$ be a convex function, $X \in \mathbb{R}^{n\times p}$ be the feature matrix, and $\lambda > 0$. Show that the dual of

$$\min_{\beta \in \mathbb{R}^p} f(X\beta) + \lambda\|\beta\|_1, \tag{4}$$

is as follows.

$$\max_{u \in \mathbb{R}^n} -f^*(-u) \quad \text{subject to} \quad \|X^T u\|_\infty \leq \lambda. \tag{5}$$

(Hint: You did something similar in in exam question 5.)

(b) Show the KKT stationary condition of the primal problem (4) is as follows.

$$X^T u \in \lambda\partial\|\beta\|_1 = \lambda \begin{cases} \{\text{sign}(\beta_j)\} & \text{if } \beta_j \neq 0 \\ [-1, 1] & \text{if } \beta_j = 0 \end{cases}, j = 1, \dots p. \tag{6}$$

**The idea of the screening rule is as follows.** If we are given that the dual maximiztion has a lower bound $\gamma$, i.e., there exists $u$ feasible such that $-f^*(-u) \geq \gamma$, then for each $k = 1, 2, ..., p$, we can perform the following optimization:

$$T_k := \max_{u \in \mathbb{R}^n} |X_k^T u| \quad \text{subject to} \quad -f^*(-u) \geq \gamma. \tag{7}$$

If we have $T_k < \lambda$, then at the optimal dual solution, we will also have $|X_k^T u^*| < \lambda$, since $-f^*(-u) \geq \gamma$ contains the optimal solution $u^*$. This is critical because by the stationary KKT condition of the primal in (6) we have $\beta_k = 0$. So by strong duality the primal solution also has $\beta_k = 0$, hence giving us a screening rule that eliminates $\beta_k$ from the problem.

The remainder of this problem has two tasks: solving the optimization for $T_k$ and finding the lower bound $\gamma$.

(c) **(Solving $T_k$)** We break problem (7) with the sign of $X_k^T u$. Derive that the dual of the positive part

$$T_{k,+} := \max_{u \in \mathbb{R}^n} X_k^T u \quad \text{subject to} \quad -f^*(-u) \geq \gamma \tag{8}$$

is:

$$T_{k,+} = \min_{\mu > 0} -\mu\gamma + \mu f(-\frac{X_k}{\mu}). \tag{9}$$

and show that for Lasso, where $f(Z) = \frac{1}{2}\|Z - Y\|_2^2$, we have

$$T_{k,+} = \sqrt{Y^T Y - 2\gamma} + Y^T X_k,$$

assuming that feature is normalized so that $X_k^T X_k = 1$.

Similarly you can do the same for $T_{k,-}$ (you don't have to show this) and finally you will have

$$T_k = \max(T_{k,+}, T_{k,-}) = \sqrt{Y^T Y - 2\gamma} + |Y^T X_k|. \tag{10}$$

(d) **(Finding $\gamma$)** It's enough to find an appropriate $\gamma$ that lower bounds the dual objective. One way to do so is first find a $u_0$, such that $\|X^T u_0\|_\infty = \lambda_0 \geq \lambda$, and then scale $u_0$ so that the dual constraint is met, i.e., we set $u = su_0$. To find the optimal scaling factor $s$, we solve the following:

$$\gamma(u_0) = \max_s -f^*(-su_0) \quad \text{subject to} \quad |s| \leq \frac{\lambda}{\lambda_0}.$$

Assume $\lambda_{max} = \max_k |X_k^T Y| \geq \lambda$. Solve this optimization for Lasso by setting $u_0 = -Y$. What's the optimal $s$ and the resulting $\gamma$? What's the final form of $T_k$ in terms of $X_k$, $Y$, $\lambda$ and $\lambda_{max}$?

(e) (Bonus) What happens if $\lambda_{max} < \lambda$?

# 3 ADMM to the rescue [Dallas]

For **any two of the folloiwng three** problems, reparametrize in a such a way that allows you to apply ADMM, and describe the ADMM steps, with the dual variable in scaled form. Note: if the ADMM subproblems require further optimization, i.e., they do not admit a closed form solution, then you must explain how to solve them. (Bonus points for introducing the fewest auxiliary vairables as possible, and getting each subproblem to be solveable closed form.)

(a) For a given matrix $S \in \mathbb{S}_+^n$,

$$\min_{\Theta \in \mathbb{S}_n^+} \ \operatorname{tr}(S\Theta) - \log\det\Theta + \lambda \|M(\Theta)\|_1$$

where $M$ is a linear operator that excludes the diagonal of the input matrix, i.e.,

$$[M(\Theta)]_{ij} = \begin{cases} \Theta_{ij} & i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

(b) For given vectors $a, b \in \mathbb{R}^n$,

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & \|x - a\|_2^2 + \lambda \|x - b\|_\infty \\ \text{subject to} & 1^T x = 1, \ x \geq 0. \end{array}$$

(c) For a given matrix $X \in \mathbb{R}^{n \times p}$,

$$\begin{array}{ll} \min\limits_{P \in \mathbb{S}_+^n} & \|X - PX\|_F^2 + \lambda \sum\limits_{i=1}^{p} \sum\limits_{j=1}^{p-1} |P_{i,j} - P_{i,j+1}| \\ \text{subject to} & \operatorname{tr}(P) = k, \ P \succeq 0. \end{array}$$

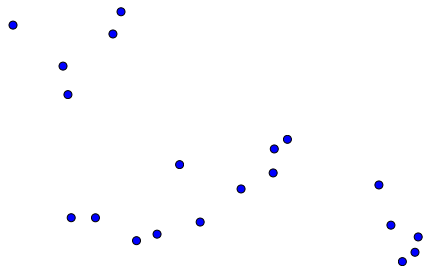

Figure 1: 20 points in $\mathbb{R}^2$ to be clustered.

# 4 Convex clustering via ADMM [Matt]

In this problem we will implement convex clustering using ADMM. To keep the implementation simple, we will deal with a small dataset of $n = 20$ points in $\mathbb{R}^2$, shown in Figure 1, and available as `points.txt` on the course website. Arranging the points as the rows of the matrix $Y \in \mathbb{R}^{20\times 2}$, we'll perform clustering by solving the optimization problem

$$\min_{X\in\mathbb{R}^{20\times 2}} \ (1/2)\|X - Y\|_F^2 + \sum_{i,j} w_{ij}\|x_i - x_j\|_2$$

where $x_i^T$ is the $i$th row of the matrix matrix $X$ and the weight $w_{ij}$ is defined as

$$w_{ij} = \exp\left(-\gamma\|y_i - y_j\|_2^2\right)$$

Conceptually, even though $X$ is the same size as $Y$, the group fused lasso penalty will encourage many of these points to be equal, leaving us with a smaller number of cluster centers at the solution (of course, depending on the parameter $\gamma$).

Next, as in class, define $D$ to be the $|E| \times n$ differencing operator over a graph: if $e_\ell = (i, j)$, then $D$ has $\ell$th row

$$D_\ell = (0, \ldots \underset{\uparrow}{-1}_{i}, \ldots \underset{\uparrow}{1}_{j}, \ldots 0).$$

In this problem we use the fully connected graph and so we have $|E| = (20 \cdot 19)/2$. Now, we apply ADMM to the modified problem

$$\begin{aligned} \min_{X,Z} \ & (1/2)\|X - Y\|_F^2 + \sum_{i,j\in E} w_{ij}\|z_{ij}\|_2, \\ \text{subject to } \ & Z = DX \end{aligned}$$

where $z_{ij}^T$ is the row of $Z$ corresponding to the difference between nodes $i$ and $j$. By modifying the problem in this way, the objective is now separable *and* each of the terms will yield a subproblem with simple closed-form solution (as we will see shortly). The ADMM iterations for this problem are given by

$$\begin{aligned} X^{k+1} &:= \operatorname*{argmin}_X \ (1/2)\|X - Y\|_F^2 + (\rho/2)\|DX - Z^k + U^k\|_F^2 \\ Z^{k+1} &:= \operatorname*{argmin}_X \ \sum_{i,j\in E} w_{ij}\|z_{ij}\|_2 + (\rho/2)\|DX^{k+1} - Z + U^k\|_F^2 \\ U^{k+1} &:= U^k + DX^{k+1} - Z^{k+1} \end{aligned}$$

where $U$ is a scaled version of the dual variable for the equality constraint.

(a) Derive the closed form solutions for the $X$- and $Z$-updates.

(b) Implement these iterations and run the algorithm with $\gamma = 1.5$ and $\rho = 0.1$, use the stopping criterion

$$\|DX^k - Z^k\|_F \leq 10^{-3}$$
$$\|\rho D^T(Z^k - Z^{k-1})\|_F \leq 10^{-3}.$$

Plot the cluster centers along with the original points.

(c) (bonus) Vary $\gamma$ and plot the solution path (e.g. in the style of lecture 19, slide 47).