EECS 127 - EECS 227AT 5/10/2018
L. El Ghaoui

Final: Solutions

1. (*10 points*) Consider the set in $\mathbb{R}^{2n}$

$$\mathcal{E} := \left\{ z = (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \ : \ F(x, y) := \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}^T \begin{pmatrix} A_{xx} & A_{xy} & a_x \\ A_{xy}^T & A_{yy} & a_y \\ a_x^T & a_y^T & \alpha \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \leq 0 \right\},$$

where $A_{xx}, A_{xy}, A_{yy} \in \mathbb{R}^{n \times n}$, $a_x, a_y \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, and the $2n \times 2n$ matrix

$$A := \begin{pmatrix} A_{xx} & A_{xy} \\ A_{xy}^T & A_{yy} \end{pmatrix}$$

is symmetric and positive-definite. We define $a = (a_x, a_y) \in \mathbb{R}^{2n}$.

(a) *(3 points)* Express the function F, evaluated at a point $z = (x, y)$, as

$$F(x, y) = (z - z_0)^T Q (z - z_0) - \gamma$$

for some appropriate matrix Q, vector z_0, and scalar γ, which you will determine.

(b) *(3 points)* What is the shape of $\mathcal{E}$ in $\mathbb{R}^{2n}$? Express your answer in geometric terms, assuming that some relevant eigenvalue decomposition is available. In particular, explain why the condition $\alpha > a^T A^{-1} a$ ensures that $\mathcal{E}$ is empty.

(c) *(4 points)* Likewise, determine the projection of $\mathcal{E}$ onto the space of x-variables. *Hint:* a point $x \in \mathbb{R}^n$ belongs to the projection if and only if $\min_y \ F(x, y) \leq 0$.

Solution:

(a) We have

$$F(z) = z^T A z + 2 a^T z + \alpha = (z - z_0)^T A (z - z_0) + \alpha - z_0^T A z_0$$

where $z_0 := A^{-1} a$. We thus have $Q = A$ and $\gamma = z_0^T A z_0 - \alpha = a^T A^{-1} a - \alpha$.

(b) In the space of $z = (x, y) \in \mathbb{R}^{2n}$, the set $\mathcal{E}$ has the form

$$\mathcal{E} = \left\{ z \ : \ (z - z_0)^T A (z - z_0) \leq \gamma \right\}.$$

Thus, if $\gamma < 0$, that is, $\alpha > a^T A^{-1} a$, the above set is empty; if $\alpha = a^T A^{-1} a$, the set is the singleton $\{z_0\}$. Otherwise, the set is an ellipse having center at z_0, and with semi-axis lengths given by $\sqrt{\gamma / \lambda_i}$, $i = 1, \ldots, n$, with λ_i the inverse eigenvalues of A, and principal directions given by the eigenvectors of A.

(c) The projection $\mathcal{P}$ is the set of points $x \in \mathbb{R}^n$ such that

$$\begin{aligned} 0 &\geq \min_y F(x,y) && (1)\\ &= \min_y x^T A_x x x + y^T A_{yy} y + 2x^T A_{xy} y + 2a_x^T x + 2a_y^T y + \alpha && (2) \end{aligned}$$

The above problem is an unconstrained minimization problem with a convex quadratic objective. Solving for the minimum simply entails taking the derivative with respect to y and setting it to zero. This leads to the optimality condition

$$A_{yy} y + A_{xy}^T x + a_y = 0.$$

Solving for y, we obtain the optimal solution

$$y^*(x) = -A_{yy}^{-1}(A_{xy}^T x + a_y).$$

Note that A_{yy} is invertible since A is positive-definite.

Replacing y by its optimal value in (2), we conclude that the condition $x \in \mathcal{P}$ is equivalent to

$$x^T(A_{xx} - A_{xy}A_{yy}^{-1}A_{xy}^T)x + 2(a_x - A_{yy}^{-1}a_y)^T x + \beta \leq 0,$$

where

$$\beta := \alpha - a_y^T A_{yy}^{-1} a_y.$$

The above condition can be analyzed as before.

2. (*10 points*) We consider regularized supervised learning problems of the form

$$P(\lambda) \ : \ \min_w \ f_\lambda(w) := \mathcal{L}(X^T w) + \frac{\lambda}{2}\|w\|_2^2,$$

where $\lambda > 0$ is a regularization parameter, $X \in \mathbb{R}^{n\times m}$ is a matrix of data points, and $\mathcal{L}$ is a differentiable, convex loss function, which we assume is decomposable: for any given $z \in \mathbb{R}^m$:

$$\mathcal{L}(z) = \sum_{i=1}^m l_i(z_i),$$

where l_i are given (differentiable, convex) functions.

We consider a gradient method to solve the problem, which, for a single value of λ, involves iterates $w_\lambda(k)$, $k = 0, 1, 2, \ldots$, obtained via the gradient recursion

$$w_\lambda(k+1) = w_\lambda(k) - t\nabla f_\lambda(w(k)), \quad k = 0, 1, 2, 3, \ldots \tag{3}$$

where $t > 0$ is a (fixed) step size. In this exercise, we would like to solve the problems $P(\lambda)$ for a sequence of values $\lambda \in \lambda_1, \ldots, \lambda_p$. We explore the use of matrix-matrix products to run the different gradient algorithms faster than running the recursions (3) one after the other in sequence. (Matrix-matrix products are highly optimized memory-wise, and usually are much faster than a corresponding sequence of matrix-vector products, for example.)

(a) *(2 points)* Show that forming the gradient of $\mathcal{L}$ at a point $z \in \mathbb{R}^m$ is a component-wise operation, namely

$$\nabla\mathcal{L}(z) = \begin{pmatrix} \nabla l_1(z_1) \\ \vdots \\ \nabla l_m(z_m) \end{pmatrix}.$$

Hint: approximate $\mathcal{L}(z+\delta) - \mathcal{L}(z)$ for any $z, \delta \in \mathbb{R}^m$, with δ small.

(b) *(2 points)* Show that for any given w, and $\lambda > 0$, $\nabla f_\lambda(w) = X\nabla\mathcal{L}(X^T w) + \lambda w$. *Hint:* again use a first-order expansion of f_λ.

(c) *(3 points)* Define the matrices $W(k) = [w_{\lambda_1}(k), \ldots, w_{\lambda_p}(k)] \in \mathbb{R}^{n\times p}$, $k = 0, 1, 2, \ldots$. Show that the recursion (3) can be written

$$W(k+1) = W(k) - t\left(X\nabla\mathcal{L}(X^T W(k)) + W(k)D\right), \quad k = 0, 1, 2, 3, \ldots$$

for an appropriate diagonal matrix D, which you will determine, and with the convention that $\nabla\mathcal{L}$ applies in column-wise fashion on a matrix input: for a matrix $Z = [z_1, \ldots, z_p] \in \mathbb{R}^{m\times p}$, we set $\nabla\mathcal{L}(Z) = [\nabla\mathcal{L}(z_1), \ldots, \nabla\mathcal{L}(z_p)] \in \mathbb{R}^{m\times p}$.

(d) *(3 points)* In some variants of the gradient method, the step size depends on the regularization parameter λ and/or on the iteration count k, which we denote by $t_\lambda(k)$. How would the matrix recursion above be modified in that case?

Solution:

(a) We have, for any $z, \delta \in \mathbb{R}^m$ and $i \in \{1, \ldots, m\}$:

$$l_i(z_i + \delta_i) \approx l_i(z_i) + \delta_i \nabla l_i(z_i),$$

Summing we obtain

$$\mathcal{L}(z + \delta) - \mathcal{L}(z) \approx \delta^T \nabla \mathcal{L}(z),$$

provided we define $\nabla \mathcal{L}(z)$ as indicated.

(b) Again, we use a first-order expansion; let $w, u \in \mathbb{R}^n$. With $\delta = X^T u$:

$$\begin{aligned} f_\lambda(w + u) &= \mathcal{L}(X^T w + \delta) + \frac{\lambda}{2}(w + u)^T (w + u) \\ &\approx \mathcal{L}(X^T w) + \delta^T \nabla \mathcal{L}(X^T w) + \frac{\lambda}{2}(w^T w + 2w^T u + u^T u) \\ &\approx f_\lambda(X^T w) + u^T (X \nabla (X^T w) + \lambda w), \end{aligned}$$

which proves the result.

(c) We express the update equations at step k as

$$w_{\lambda_i}(k+1) = w_{\lambda_i}(k) - t \left(X \nabla \mathcal{L}(X^T w_{\lambda_i}(k)) + \lambda_i w_{\lambda_i}(k) \right), \quad 1 \le i \le p,$$

which is the same as

$$W(k+1) = W(k) - t \left(X \nabla \mathcal{L}(X^T W(k)) + W D_\lambda \right),$$

with the suggested column-wise convention on $\nabla \mathcal{L}$, and with

$$D_\lambda := \mathbf{diag}(\lambda_1, \ldots, \lambda_p).$$

(d) When the step size $t = t_\lambda(k)$ depends on the regularization parameter λ and the iteration count k, we simply replace the iterations above by

$$W(k+1) = W(k) - \left(X \nabla \mathcal{L}(X^T W(k)) + W D_\lambda \right) D_\lambda(k),$$

where D_λ is defined as before, and

$$D_\lambda(k) := \mathbf{diag}(t_{\lambda_1}(k), \ldots, t_{\lambda_p}(k)).$$

3. (10 points) *Risk Budgeting.* We are given a symmetric, $n \times n$ positive-definite matrix C, and a vector $\theta \in \mathbb{R}^n_{++}$, with $\mathbf{1}^T\theta = 1$, where $\mathbf{1}$ is the n-vector of ones. We assume that a factorization of the form $C = R^TR$, with $R \in \mathbb{R}^{n\times n}$, is available.

We consider a *risk budgeting* problem arising in financial optimization, which consists in finding a vector $x \in \mathbb{R}^n$ such that

$$x > 0, \quad x_i(Cx)_i = \theta_i(x^TCx), \quad i = 1, \ldots, n. \tag{4}$$

(In case you are curious: the term "risk budgeting" refers to the fact that each so-called "partial risk" $x_i(Cx)_i$ is assigned a fixed proportion θ_i of the total risk, defined as the variance x^TCx, itself the sum of the partial risks.)

(a) *(1 point)* Is the problem of finding a risk budgeting portfolio, as defined above, or determine there is no such portfolio, convex? Justify your answer carefully.

(b) *(3 points)* Consider a generic constraint on a triple (z, u, v), with z a vector and u, v scalars, of the form

$$u \geq 0, \quad v \geq 0, \quad uv \geq z^Tz.$$

Prove that the above so-called "rotated second-order cone" constraint can be written as

$$\left\| \begin{pmatrix} 2z \\ u - v \end{pmatrix} \right\|_2 \leq u + v.$$

Make sure to carefully handle the signs of u, v.

(c) *(2 points)* Consider the constraints

$$x > 0, \quad x_i(Cx)_i \geq \theta_i(x^TCx), \quad i = 1, \ldots, n. \tag{5}$$

Show how they can be written as second-order cone constraints on x.

(d) *(3 points)* Show that any point $x \in \mathbb{R}^n$ that satisfies (5) also satisfies constraints (4). *Hint:* proceed by contradiction, and think about summing constraints (5).

(e) *(1 point)* Consider the problem of maximizing an expected return $\hat{r}^Tx$, subject to the risk budget constraints. Express the problem as an SOCP.

Solution:

(a) The risk budgeting problem is not convex as given, since it involves quadratic *equality* constraints.

(b) Assume that the rotated cone constraint holds. Then,

$$\left\| \begin{pmatrix} 2z \\ u - v \end{pmatrix} \right\|_2^2 = 4z^Tz + (u - v)^2 = 4z^Tz + (u + v)^2 - 4uv \leq (u + v)^2.$$

Since $u + v \geq 0$, we obtain that the constraint

$$\left\| \begin{pmatrix} 2z \\ u - v \end{pmatrix} \right\|_2 \leq u + v.$$

holds. Conversely, if the above holds, then $u + v \geq 0$, and after squaring the above, we also get

$$uv \geq z^T z.$$

Thus $uv \geq 0$, which, together with $u + v \geq 0$, implies $u \geq 0$, $v \geq 0$.

(c) The constraints can be written as

$$\left\| \begin{pmatrix} 2\sqrt{\theta_i} Rx \\ x_i - (Cx)_i \end{pmatrix} \right\|_2 \leq x_i + (Cx)_i, \quad i = 1, \ldots, n.$$

(d) Assume that x satisfies (5). Assume that one of these inequalities is actually strict. Summing over i, we obtain

$$x^T Cx = \sum_{i=1}^{n} x_i (Cx)_i > \left(\sum_{i=1}^{n} \theta_i \right) (x^T Cx) = x^T Cx,$$

a contradiction.

(e) The problem writes

$$\max_x \ \hat{r}^T x \ : \ x \geq 0, \ \left\| \begin{pmatrix} 2Rx \\ x_i - (Cx)_i \end{pmatrix} \right\|_2 \leq x_i + (Cx)_i, \quad i = 1, \ldots, n.$$

4. (*10 points*) Let $A \in \mathbb{R}^{m\times n}$, $y \in \mathbb{R}^m$ and $\mu > 0$. Consider the problem

$$p^* = \min_x \ \|Ax - y\|_1 + \mu\|x\|_\infty.$$

For $j \in \{1, \ldots, n\}$, we denote by a_j the j-th column of A, so that $A = [a_1, \ldots, a_n]$, and define

$$\|A\|_1 := \sum_{j=1}^n \|a_j\|_1.$$

(a) *(2 points)* Express the problem as an LP.

(b) *(4 points)* Show that a dual to the problem can be written as

$$d^* = \max_u \ -u^T y \ : \ \|u\|_\infty \leq 1, \ \ \|A^T u\|_1 \leq \mu.$$

Hint: use the fact that, for any vector z:

$$\max_{u\,:\,\|u\|_1\leq 1} u^T z = \|z\|_\infty, \quad \max_{u\,:\,\|u\|_\infty\leq 1} u^T z = \|z\|_1.$$

In the next questions, you may assume that strong duality holds.

(c) *(2 points)* Show that the condition "$\|A^T u\|_1 < \mu$ for every u with $\|u\|_\infty \leq 1$" ensures that $x = 0$ is optimal.

(d) *(2 points)* Show that the condition in the previous part holds if $\mu > \|A\|_1$.

Solution:

(a) The problem writes

$$\min_{x,z,t} \ z^T\mathbf{1} + \mu t \ : \ t \geq \|x\|_\infty, \ \ z_i \geq |(Ax - y)_i|, \ \ i = 1, \ldots, m.$$

which is an LP:

$$\begin{array}{lll} \min\limits_{x,z,t} \ z^T\mathbf{1} + \mu t \ : & t \geq x_j, \ \ t \geq -x_j, j = 1, \ldots, n & \\ & z_i \geq (Ax - y)_i, \ \ z_i \geq -(Ax - y)_i, \ \ i = 1, \ldots, m. & \end{array} \tag{6}$$

(b) Based on the hint, we use the Lagrangian

$$\mathcal{L}(x, u, v) = u^T(Ax - y) + v^T x,$$

which is such that

$$p^* = \min_x \ \max_{u,v} \ \{\mathcal{L}(x, u, v) \ : \ \|u\|_\infty \leq 1, \ \ \|v\|_1 \leq \mu\}. \tag{7}$$

Exchanging min and max leads to the dual; assuming strong duality[1]

$$p^* \geq d^* = \max_{u,v} \; g(u,v),$$

with g the dual function

$$g(u,v) = \min_x \; \mathcal{L}(x,u,v) = \begin{cases} -u^T y & \text{if } A^T u + v = 0, \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem writes

$$d^* = \max_u \; -u^T y \;\; : \;\; A^T u + v = 0, \;\; \|u\|_\infty \leq 1, \;\; \|v\|_1 \leq \mu.$$

We can eliminate v:

$$d^* = \max_u \; -u^T y \;\; : \;\; \|u\|_\infty \leq 1, \;\; \|A^T u\|_1 \leq \mu.$$

(c) If

$$\|A^T u\|_1 < \mu \text{ for every } u, \;\; \|u\|_\infty \leq 1, \tag{8}$$

then the constraint $\|A^T u\|_1 \leq \mu$ cannot active in the dual. Therefore, the dual problem has the same value as one where that constraint is removed, meaning that

$$p^* = d^* = \max_{u \,:\, \|u\|_\infty \leq 1} \; -u^T y = \|y\|_1.$$

We observe that the value $p^* = \|y\|_1$ is attained for $x = 0$, which shows that $x = 0$ is optimal.

(d) The condition (8) is satisfied iff

$$\mu > \max_{u \,:\, \|u\|_\infty \leq 1} \; \sum_{j=1}^n |a_j^T u|.$$

Since

$$\max_{u \,:\, \|u\|_\infty \leq 1} \; \sum_{j=1}^n |a_j^T u| \leq \sum_{j=1}^n \max_{u \,:\, \|u\|_\infty \leq 1} \; |a_j^T u| = \|A\|_1,$$

the desired result follows.

[1] In fact, we can use Sion's theorem here to show that strong duality holds.