Optimization Methods **Fall 2019**

Homework 3

Instructor: Lijun Zhang *Name:* Li Weikang, *StudentId:* 181220031

Notice

- The submission email is: **njuoptfall2019@163.com**.
- Please use the provided LaTeX file as a template. If you are not familiar with LaTeX, you can also use Word to generate a **PDF** file.

Problem 1: Equality Constrained Least-squares
Consider the equality constrained least-squares problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|Ax-b\|_2^2 \\ \text{subject to} & Gx = h \end{array}$$

where $A \in \mathbf{R}^{m\times n}$ with $\mathbf{rank}\ A = n$, and $G \in \mathbf{R}^{p\times n}$ with $\mathbf{rank}\ G = p$.

a) Derive the Lagrange dual problem with Lagrange multiplier vector v.

b) Derive expressions for the primal solution $x^\star$ and the dual solution $v^\star$.

Solution.

a) It is easy to see that the Lagrangian is

$$\begin{aligned} \mathcal{L}(x,v) &= \frac{1}{2}\|Ax-b\|_2^2 + v^T(Gx-h) \\ &= \frac{1}{2}x^TA^TAx + (G^Tv - A^Tb)^Tx - v^Th + \frac{1}{2}b^Tb, \end{aligned}$$

with minimizer $x = -(A^TA)^{-1}(G^Tv - A^Tb)$. Accordingly the dual function is

$$g(v) = -\frac{1}{2}(G^Tv - 2A^Tb)^T(A^TA)^{-1}(G^Tv - 2A^Tb) - v^Th + \frac{1}{2}b^Tb$$

Therefore, the Lagrange dual problem can be described as

$$\text{maximize} \quad -\tfrac{1}{2}(G^Tv - 2A^Tb)^T(A^TA)^{-1}(G^Tv - 2A^Tb) - v^Th + \tfrac{1}{2}b^Tb$$

b) The KKT optimality conditions are

$$A^T(Ax^* - b) + G^Tv^* = 0, \quad Gx^* = h.$$

From the first equation,

$$x^* = (A^TA)^{-1}(A^Tb - G^Tv^*).$$

Plugging this expression for x^* into the second equation gives

$$G(A^TA)^{-1}A^Tb - G(A^TA)^{-1}G^Tv^* = h$$

i.e.,

$$v^* = -(G(A^TA)^{-1}G^T)^{-1}(h - G(A^TA)^{-1}A^Tb).$$

Substituting in the first expression gives an analytical expression for x^*.

□

Problem 2: Support Vector Machines
Consider the following optimization problem

$$\text{minimize} \quad \sum_{i=1}^{n} \max\left(0, 1 - y_i(w^T x_i + b)\right) + \frac{\lambda}{2}\|w\|_2^2$$

where $x_i \in \mathbf{R}^d, y_i \in \mathbf{R}, i = 1, \cdots, n$ are given, and $w \in \mathbf{R}^d, b \in \mathbf{R}$ are the variables.

a) Derive an equivalent problem by introducing new variables $u_i, i = 1, \cdots, n$ and equality constraints

$$u_i = y_i(w^T x_i + b), i = 1, \cdots, n.$$

b) Derive the Lagrange dual problem of the above equivalent problem.

c) Give the Karush-Kuhn-Tucker conditions.

Hint: Let $\ell(x) = \max(0, 1 - x)$*. Its conjugate function* $\ell^*(y) = \sup_x (yx - \ell(x)) = \begin{cases} y, & -1 \le y \le 0 \\ \infty, & \textit{otherwise} \end{cases}$

Solution.

a) We plug the equality constraints into the original problem to derive the equivalent problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \max(0, 1 - u_i) + \frac{\lambda}{2}\|w\|_2^2 \\ \text{subject to} & u_i = y_i(w^T x_i + b), \quad i = 1, \cdots, n. \end{array}$$

b) From the equivalent problem above we derive its Lagrangian

$$\mathcal{L}(w, b, u, v) = \sum_{i=1}^{n} \max(0, 1 - u_i) + \frac{\lambda}{2}\|w\|_2^2 + \sum_{i=1}^{n} v_i(u_i - y_i(w^T x_i + b))$$

Thus, the Lagrangian dual function is

$$\begin{aligned} g(v) &= \inf_{w,b,u} \sum_{i=1}^{n} \max(0, 1 - u_i) + \frac{\lambda}{2}\|w\|_2^2 + \sum_{i=1}^{n} v_i(u_i - y_i(w^T x_i + b)) \\ &= \inf_u \sum_{i=1}^{n} [\max(0, 1 - u_i) + v_i u_i] + \frac{1}{2\lambda}\|\sum_{i=1}^{n} v_i x_i y_i\|^2 - \sum_{i=1}^{n} \frac{v_i y_i}{\lambda} \sum_{j=1}^{n} v_j x_j^T x_i 2 y_j \end{aligned}$$

Here we introduce $\ell(x) = \max(0, 1 - x)$.

And its conjugate function $\ell^*(y) = \sup_x (yx - \ell(x)) = \begin{cases} y, & -1 \le y \le 0 \\ \infty, & \text{otherwise} \end{cases}$

Therefore,

$$\begin{aligned} g(v) &= \inf_u \sum_{i=1}^{n} [\ell(u_i) + v_i u_i] + \frac{1}{2\lambda}\|\sum_{i=1}^{n} v_i x_i y_i\|^2 - \sum_{i=1}^{n} \frac{v_i y_i}{\lambda} \sum_{j=1}^{n} v_j x_j^T x_i y_j \\ &= -\sup_u \sum_{i=1}^{n} [-v_i u_i - \ell(u_i)] + \frac{1}{2\lambda}\|\sum_{i=1}^{n} v_i x_i y_i\|^2 - \sum_{i=1}^{n} \frac{v_i y_i}{\lambda} \sum_{j=1}^{n} v_j x_j^T x_i y_j \\ &= -\sum_{i=1}^{n} \sup_u [-v_i u_i - \ell(u_i)] + \frac{1}{2\lambda}\|\sum_{i=1}^{n} v_i x_i y_i\|^2 - \sum_{i=1}^{n} \frac{v_i y_i}{\lambda} \sum_{j=1}^{n} v_j x_j^T x_i y_j \\ &= \sum_{i=1}^{n} v_i + \frac{1}{2\lambda}\|\sum_{i=1}^{n} v_i x_i y_i\|^2 - \sum_{i=1}^{n} \frac{v_i y_i}{\lambda} \sum_{j=1}^{n} v_j x_j^T x_i y_j \end{aligned}$$

where $0 \le v_i \le 1$.
Finally, the Lagrangian dual problem is

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{n} v_i + \frac{1}{2\lambda}\|\sum_{i=1}^{n} v_i x_i y_i\|^2 - \sum_{i=1}^{n} \frac{v_i y_i}{\lambda} \sum_{j=1}^{n} v_j x_j^T x_i y_j \\ \text{subject to} & \sum_{i=1}^{n} v_i y_i = 0, \\ & 0 \le v_i \le 1. \end{array}$$

c) The KKT optimality conditions are:

$$u_i^* = y_i({w^*}^T x_i + b^*), i = 1, \cdots, n.$$

$$\nabla \sum_{i=1}^{n} \max(0, 1 - u_i^*) + \frac{\lambda}{2}\|w^*\|_2^2 + \nabla \sum_{i=1}^{n} v_i(u_i^* - y_i({w^*}^T x_i + b^*)) = 0.$$

□

Problem 3: Euclidean Projection onto the Simplex
Consider the following optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|y - x\|_2^2 \\ \text{subject to} & \mathbf{1}^T y = r \\ & y \succeq 0 \end{array}$$

where $r > 0$, $x \in \mathbb{R}^n$ is given, and $y \in \mathbf{R}^n$ is the variable. Give an algorithm to solve this problem and prove the correctness of your algorithm.
Hint: Derive the Lagrangian of this problem and apply the Karush-Kuhn-Tucker conditions. If you need more hints, please read the following paper [1]

Solution.

To begin with, we put forward an efficient algorithm to solve the problem.

Algorithm 1 Euclidean projection of a vector onto the probability simplex.

Input: $\mathbf{x} \in \mathbb{R}^n$

Sort $\mathbf{x}$ into $\mathbf{u}$: $u_1 \geq u_2 \geq \cdots \geq u_n$

Find $\rho = \max\{1 \leq j \leq n : u_j + \frac{1}{j}(r - \sum_{i=1}^{j} u_i) > 0\}$

Define $\lambda = \frac{1}{\rho}(r - \sum_{i=1}^{\rho} u_i)$

Output: $\mathbf{y}$ s.t. $y_i = \max\{x_i + \lambda, 0\}, i = 1, \ldots, n$.

Proof. The Lagrangian of the problem is

$$\mathcal{L}(\mathbf{y}, \lambda, \beta) = \frac{1}{2}\|y - x\|_2^2 - \lambda(\mathbf{1}^T y - r) - \beta^T y$$

where λ and $\beta = [\beta_1, \ldots, \beta_n]^T$ are the Lagrange multipliers for the constraints. At the optimal solution $\mathbf{y}$ the following KKT conditions hold:

$$y_i - x_i - \lambda - \beta_i = 0, \qquad i = 1, \ldots, n \tag{1}$$

$$y_i \geq 0, \qquad i = 1, \ldots, n \tag{2}$$

$$\beta_i \geq 0, \qquad i = 1, \ldots, n \tag{3}$$

$$y_i \beta_i = 0, \qquad i = 1, \ldots, n \tag{4}$$

$$\sum_{i=1}^{n} y_i = r. \tag{5}$$

From the condition (4), it is trivial that if $y_i > 0$, we must have $\beta_i = 0$ and $y_i = x_i + \lambda$; if $y_i = 0$, we must have $\beta_i \geq 0$ and $y_i = x_i + \lambda + \beta_i = 0$, thus $x_i + \lambda = -\beta_i \leq 0$. It is obvious that the components of the optimal solution $\mathbf{y}$ that are zeros correspond to the smaller components of $\mathbf{x}$. Without loss of generality, we assume the components of $\mathbf{x}$ are sorted and $\mathbf{y}$ uses the same ordering, i.e.,

$$x_1 \geq \cdots \geq x_\rho \geq x_\rho + 1 \geq \cdots \geq x_n,$$
$$y_1 \geq \cdots \geq y_\rho > y_\rho + 1 = \cdots = y_n,$$

and that $y_1 \geq \cdots \geq y_\rho > 0$, $y_\rho + 1 = \cdots = y_n = 0$ In other words, ρ is the number of positive components in the solution $\mathbf{y}$. Now we apply the last condition and have

$$r = \sum_{i=1}^{n} y_i = \sum_{i=1}^{\rho} y_i = \sum_{i=1}^{\rho} (x_i + \lambda)$$

which gives $\lambda = \frac{1}{\rho}\left(r - \sum_{i=1}^{\rho} x_i\right)$. Hence ρ is the key to the solution. And once we know ρ(there are only n possible values of it), we can compute λ and the optimal solution is obtained by just adding λ to each component of $\mathbf{x}$ and thresholding as in the end of Algorithm 1.(It is easy to check that this solution indeed satisfies all KKT conditions.) In the algorithm, we carry out the tests for $j = 1, \dots, n$ if $t_j = x_j + \frac{1}{j}\left(r - \sum_{i=1}^{j} x_i\right) > 0$ We now prove that the number of times this test turns out positive is exactly ρ. The following theorem is essentially Lemma 3 of Shalev-Shwartz and Singer (2006).

Theorem 1. Let ρ be the number of positive components in the solution $\mathbf{y}$, then

$$\rho = \max\{1 \le j \le n : x_j + \frac{1}{j}(r - \sum_{i=1}^{j} x_i) > 0\}.$$

Proof. From the KKT condition (2) we have that $\lambda\rho = (r - \sum_{i=1}^{\rho} x_i)$, $x_i + \lambda > 0$ for $i = 1, \dots, \rho$ and $x_i + \lambda \le 0$ for $i = \rho+1, \dots, n$. In the sequel, we know that for $j = 1, \dots, n$, the test will continue to be positive until $j = \rho$ and then stay non-positive afterwards, i.e., $x_j + \frac{1}{j}\left(r - \sum_{i=1}^{j} x_i\right) > 0$ for $j \le \rho$, and $x_j + \frac{1}{j}\left(r - \sum_{i=1}^{j} x_i\right) \le 0$ for $j > \rho$.

(i) For $j = \rho$, we have

$$x_\rho + \frac{1}{\rho}\left(r - \sum_{i=1}^{\rho} x_i\right) = x_\rho + \lambda = y_\rho > 0$$

(ii) For $j < \rho$, we have

$$x_j + \frac{1}{j}\left(r - \sum_{i=1}^{j} x_i\right) = \frac{1}{j}(jx_j + r - \sum_{i=1}^{j} x_i) = \frac{1}{j}\left(jx_j + \sum_{i=j+1}^{\rho} x_i + r - \sum_{i=1}^{\rho} x_i\right) = \frac{1}{j}\left(jx_j + \sum_{i=j+1}^{\rho} x_i + \rho\lambda\right)$$
$$= \frac{1}{j}\left(j(x_j + \lambda) + \sum_{i=j+1}^{\rho}(x_i + \lambda)\right).$$

Since $x_i + \lambda > 0$ for $i = j, \dots, \rho$, we have $x_j + \frac{1}{j}(r - \sum_{i=1}^{j} x_i) > 0$.

(iii) For $j > \rho$, we have

$$x_j + \frac{1}{j}\left(r - \sum_{i=1}^{j} x_i\right) = \frac{1}{j}(jx_j + r - \sum_{i=1}^{j} x_i) = \frac{1}{j}\left(jx_j + r - \sum_{i=1}^{\rho} x_i - \sum_{i=\rho+1}^{j} x_i\right) = \frac{1}{j}\left(jx_j + \rho\lambda - \sum_{i=\rho+1}^{j} x_i\right)$$
$$= \frac{1}{j}\left(\rho(x_j + \lambda) + \sum_{i=\rho+1}^{j}(x_j - x_i)\right).$$

Notice that $x_j + \lambda \le 0$ for $j > \rho$, and $x_j \le x_i$ for $j \ge i$ since $\mathbf{x}$ is sorted, therefore $x_j + \frac{1}{j}(1 - \sum_{i=1}^{j} x_i) < 0$. $\square$

Problem 4: Optimality Conditions

Consider the problem

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(2X) - \log\det(3X) \\ \text{subject to} & 2Xs = y \end{array}$$

with variable $X \in \mathbf{S}^n$ and domain $\mathbf{S}^n_{++}$. Here, $y \in \mathbf{R}^n$ and $s \in \mathbf{R}^n$ are given, with $s^T y = 1$.

a) Give the Lagrange and then derive the Karush-Kuhn-Tucker conditions.

b) Verify that the optimal solution is given by

$$X^\star = \frac{1}{2}\left(I + yy^T - \frac{ss^T}{s^T s}\right).$$

Solution.

We introduce a Lagrange multiplier $z \in \mathbf{R}^n$ for the equality constraint.
According to the properties of **trace**, $\nabla_A \text{tr}(AB) = \nabla_A \text{tr}(BA) = B^T$.
Thus, we have:

$$\nabla_X \text{tr}(nX) = n\nabla_X \text{tr}(IX) = nI$$

Refer to the proof in **section A.4.1** of the book Stephen Boyd, Lieven Vandenberghe, Convex Optimization, we have:

$$\nabla_X \log\, \det X = X^{-1}$$

The KKT optimality conditions are:

$$X \succ 0, \quad 2Xs = y, \quad X^{-1} = 2I + zs^T + sz^T. \tag{1}$$

We first determine z from the condition $2Xs = y$. Multiplying the gradient equation on the right with y gives

$$s = \frac{1}{2}X^{-1}y = y + \frac{1}{2}(z + (z^T y)s). \tag{2}$$

By taking the inner product with y on both sides and simplifying, we get $z^T y = 1 - y^T y$.
Substituting in (2) we get

$$z = -2y + (1 + y^T y)s,$$

and substitute this expression for z in (1) gives

$$X^{-1} \;=\; 2(I - ys^T - sy^T + (1 + y^T y)ss^T)$$

Finally we verify that this inverse of the matrix X^* given above:

$$\begin{aligned} 2(I - ys^T - sy^T + (1 + y^T y)ss^T)X^* &= (I + yy^T - (1/s^T s)ss^T) + (1 + y^T y)(ss^T + sy^T - ss^T) \\ &\quad - (ys^T + yy^T - ys^T) - (sy^T + (y^T y)sy^T - (1/s^T s)ss^T) \\ &= I \end{aligned}$$

To complete the solution, we prove that $X^* \succ 0$. An easy way to see this is to note that

$$X^\star = \frac{1}{2}\left(I + yy^T - \frac{ss^T}{s^T s}\right) = \frac{1}{2}\left(I + \frac{ys^T}{\|s\|_2} - \frac{ss^T}{s^T s}\right)\left(I + \frac{ys^T}{\|s\|_2} - \frac{ss^T}{s^T s}\right)^T.$$

□

References

[1] Weiran Wang, and Miguel Á. Carreira-Peroiñán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv:1309.1541*, 2013.