LUND
UNIVERSITY

Department of
AUTOMATIC CONTROL

Exam in Optimization for Learning

2021-10-26

Grading and points

All answers must include a clear motivation. Answers should be given in English. Number all your solution sheets and indicate the total number of sheets, e.g., 1/12, 2/12 and so on.

The total number of points is 25. The maximum number of points is specified for each subproblem. Preliminary grading scales:

Grade 3:	12 points
4:	17 points
5:	22 points

Accepted aid

You are allowed to bring lecture slides. You may use the results in the slides unless the opposite is explicitly stated.

Results

Solutions will be posted on the course webpage, and results will be registered in LADOK. Date and location for display of corrected exams will be posted on the course webpage.

1. Determine if the following sets are convex or not:

a. $S_1 = \{x \in \mathbb{R} : x \text{ is integer and } x \geq 5\}$. (1 p)

b. $S_2 = \{x \in \mathbb{R}^m : \|x\|_2 \leq 1\}$. (1 p)

c. $S_3 = \{x \in \mathbb{R}^n : Ax + b \in D\}$ where

$$D = \{y \in \mathbb{R}^m : \|y\|_2 \leq 1\},$$

$A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. (1 p)

d. $S_4 = \{x \in \mathbb{R} : f(x) \leq 1\}$ where

$$f(x) = \begin{cases} \cos x & \text{if } 0 \leq x \leq 2\pi, \\ \infty & \text{otherwise} \end{cases}$$

for each $x \in \mathbb{R}$. (1 p)

Solution

a. S_1 is not convex. We have $5 \in S_1$ and $6 \in S_1$, but the convex combination $0.5 \cdot 5 + 0.5 \cdot 6 = 5.5 \notin S_1$.

b. S_2 is convex since it is a sublevel set of a convex function.

c. S_3 is convex since it is the inverse image of a convex set of an affine transformation.

d. S_4 is convex. The function f is less than or equal to 1 on $[0, 2\pi]$ and greater than 1 outside this interval. Therefore $S_4 = [0, 2\pi]$, which is convex.

2. Determine whether or not the functions below are convex.

a. $f_1 : \mathbb{R} \to \mathbb{R}$ such that

$$f_1(x) = \log\left(1 + \mathrm{e}^{-x^2}\right)$$

for each $x \in \mathbb{R}$. (1 p)

b. $f_2 : \mathbb{S}^n \to \mathbb{R}$ such that

$$f_2(X) = \lambda_{\max}(X)$$

for each $X \in \mathbb{S}^n$, where $\lambda_{\max}$ denotes the largest eigenvalue. (1 p)

c. $f_3 : \mathbb{R}^n \to \mathbb{R}$ such that

$$f_3(x) = \sum_{i=1}^{r} \left|x_{\langle i\rangle}\right|$$

for each $x \in \mathbb{R}^n$ where $1 \leq r \leq n$ is an integer and $x_{\langle i\rangle}$ is the component of x with the ith largest absolute value, meaning that

$$\left|x_{\langle 1\rangle}\right| \geq \ldots \geq \left|x_{\langle n\rangle}\right|.$$

(1 p)

d. $f_4 : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that

$$f_4 = \iota_S$$

where $S \subseteq \mathbb{R}^n$ is given by

$$S = \{x \in \mathbb{R}^n : \|x\|_0 = r\}$$

where $1 \leq r \leq n$ is a fixed integer and

$$\|x\|_0 = \text{ number of nonzero elements in the vector } x$$

for each $x \in \mathbb{R}^n$. (1 p)

Solution

a. Not convex. Note that

$$\underbrace{f_1(0.5 \cdot (-1) + (1 - 0.5) \cdot 1)}_{=\log 2} > \underbrace{0.5 f_1(-1) + (1 - 0.5) f_1(1)}_{=\log(1+\mathrm{e}^{-1})}$$

i.e. f_1 does not fulfill the definition of convexity.

b. Convex. Note that

$$f_2(X) = \min_{\substack{x \in \mathbb{R}^n \\ \text{s.t. } \|x\|_2 = 1}} \quad x^T X x$$

for each $X \in \mathbb{S}^n$, which shows that f_2 is a point-wise supremum of convex functions and therefore itself convex. Indeed, the mapping

$$X \mapsto x^T X x$$

in the maximum is linear and therefore convex.

c. Convex. Note that

$$f_3(x) = \max_{\substack{y \in \mathbb{R}^n \\ \text{s.t. } \|y\|_1 \leq r \\ \text{and } -\mathbf{1} \leq y \leq \mathbf{1}}} \quad y^T x$$

for each $x \in \mathbb{R}^n$, which shows that f_3 is a point-wise supremum of convex functions and therefore itself convex.

d. Not convex. The set S is not convex and therefore ι_s is not convex. Indeed, if $x \in S$, then $-x \in S$. However,

$$0.5x + (1 - 0.5)(-x) = 0 \notin S.$$

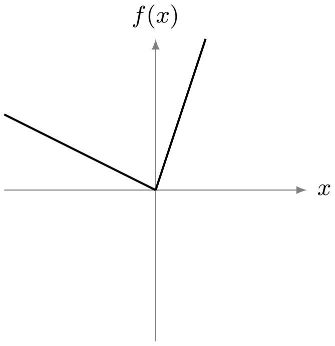

Figure 1 Function f in Problem 3.

3. Consider the function $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = \begin{cases} -0.5x & \text{if } x \leq 0, \\ 3x & \text{if } x \geq 0 \end{cases}$$

for each $x \in \mathbb{R}$.

a. Compute the subdifferential ∂f (1 p)

b. Compute prox_f (1 p)

c. Compute f^* (1 p)

d. Compute prox_{f^*} (1 p)

Solution

a. Note that f is finite-valued, closed and convex. Moreover,

$$\begin{aligned} \nabla f(x) &= -0.5, && \text{if } x < 0, \\ \nabla f(x) &= 3, && \text{if } x > 0. \end{aligned}$$

Thus,

$$\begin{aligned} \partial f(x) &= \{-0.5\}, && \text{if } x < 0, \\ \partial f(x) &= \{3\}, && \text{if } x > 0. \end{aligned}$$

Moreover, recall that f is maximally monotone. Thus, $\partial f(0) = [-0.5, 3]$. Therefore, we conclude that

$$\partial f(x) = \begin{cases} \{-0.5\} & \text{if } x < 0, \\ [-0.5, 3] & \text{if } x = 0, \\ \{3\} & \text{if } x > 0. \end{cases}$$

b. Suppose that

$$z = \mathrm{prox}_f(x) = \underset{\tilde{z} \in \mathbb{R}}{\mathrm{argmin}} \left(f(\tilde{z}) + \frac{1}{2} \|\tilde{z} - x\|_2^2 \right).$$

Fermat's rule implies that

$$0 \in \partial f(z) + z - x.$$

Plugging in the expression for ∂f gives

$$0 \in \begin{cases} \{-0.5 + z - x\} & \text{if } z < 0, \\ [-0.5, 3] - x & \text{if } z = 0, \\ \{3 + z - x\} & \text{if } z > 0. \end{cases}$$

Solving for z we get that

$$z = \operatorname{prox}_f(x) = \begin{cases} x + 0.5 & \text{if } x < -0.5, \\ 0 & \text{if } x \in [-0.5, 3], \\ x - 3 & \text{if } x > 3. \end{cases}$$

c. Recall that

$$f^*(s) = \sup_{x \in \mathbb{R}} (sx - f(x))$$

for each $s \in \mathbb{R}$.

- Case $s < -0.5$: Suppose that $x < 0$. Then

  $$f^*(s) \geq (s + 0.5)x \to \infty \quad \text{as} \quad x \to -\infty.$$

  Thus, $f^*(s) = \infty$.
- Case $s > 3$: Suppose that $x > 0$. Then

  $$f^*(s) \geq (s - 3)x \to \infty \quad \text{as} \quad x \to \infty.$$

  Thus, $f^*(s) = \infty$.
- Case $s \in [-0.5, 3]$: Fenchel-Youngs equality gives that $s \in \partial f(x)$ if and only if

  $$f^*(s) = sx - f(x).$$

  Since $s \in \partial f(0) = [-0.5, 3]$, we conclude that

  $$f^*(s) = s \cdot 0 - f(0) = 0.$$

In conclusion, we have that

$$f^* = \iota_{[-0.5,3]}.$$

d. Moreau decomposition gives that

$$\begin{aligned} \operatorname{prox}_{f^*}(x) &= x - \operatorname{prox}_f(x) \\ &= x - \begin{cases} x + 0.5 & \text{if } x < -0.5 \\ 0 & \text{if } x \in [-0.5, 3] \\ x - 3 & \text{if } z > 3 \end{cases} \\ &= \begin{cases} -0.5 & \text{if } x < -0.5 \\ x & \text{if } x \in [-0.5, 3] \\ 3 & \text{if } z > 3 \end{cases} \end{aligned}$$

for each $x \in \mathbb{R}$.

4. We will consider the problem of selecting an optimal portfolio of stocks using a mean-variance model. Suppose you wish to invest W SEK, for some $W > 0$, by picking among $n \in \mathbb{N}$ different stocks. Your portfolio of stocks is constructed at present time by purchasing x_i SEK worth of stock i, for each $i = 1, \dots, n$. The portfolio can be represented by the vector $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Naturally, there is the budget constraint that

$$\mathbf{1}^T x = W,$$

i.e., the sum of the investments equals the investment budget W. We denote the set of feasible portfolios by

$$B = \{x \in \mathbb{R}^n : \mathbf{1}^T x = W\}.$$

Note that we allow x to have negative components. A negative component x_i corresponds to short-selling stock i, i.e. borrowing the stock and immediately selling it. The portfolio of stocks is held constant until some predetermined time in the future when all investments are liquidated (sold). This corresponds to a one-period investment problem. Let r be n-dimensional, where r_i is the return of stock i over the period. In order to model our uncertainty of the future stock returns, we let r be a n-dimensional random variable with known expected value $\mathbb{E}[r] = \mu \in \mathbb{R}^n$ and known covariance matrix $\operatorname{Var}[r] = \Sigma \in \mathbb{S}^n_{++}$, i.e. Σ is a real-valued positive definite $n \times n$ matrix. The return of the portfolio, the expected return of the portfolio, and the variance of the return of the portfolio are given by

$$r^T x, \qquad \mathbb{E}\left[r^T x\right] = \mu^T x \quad \text{and} \quad \operatorname{Var}\left[r^T x\right] = x^T \Sigma x,$$

respectively. In the mean-variance model we seek the portfolio x that solves the optimization problem

$$\underset{x \in B}{\text{minimize}} \; -\mu^T x + \gamma x^T \Sigma x = \underset{x \in \mathbb{R}^n}{\text{minimize}} \; -\mu^T x + \gamma x^T \Sigma x + \iota_B(x) \qquad (1)$$

where $\gamma > 0$ is given. The variance of the return of the portfolio $x^T \Sigma x$ is a proxy for the risk inherent in the investment. Therefore, γ is usually called the risk aversion parameter and is an inverse measure of an investors risk appetite. For future reference, we define the function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x) = -\mu^T x + \gamma x^T \Sigma x$$

for each $x \in \mathbb{R}^n$.

a. Prove that f is strongly convex. (0.5 p)

b. Prove that B is convex. What does this imply for ι_B? (0.5 p)

c. Why does optimization problem (1) have a unique minimizer? (0.5 p)

d. Compute the subdifferential ∂f. (0.5 p)

e. Show that

$$\partial \iota_B(x) = \begin{cases} \{\alpha \mathbf{1} : \alpha \in \mathbb{R}\} & \text{if } x \in B, \\ \emptyset & \text{if } x \notin B \end{cases}$$

for each $x \in \mathbb{R}^n$. (1.5+0.5 p)

f. Using the subdifferentials in **d.** and **e.**, find the optimal portfolio according to the mean-variance model (1).
(You may assume that the expression for $\partial\iota_B$ in **e.** holds.) (2 p)

g. Show that the conjugate functions of f and ι_B satisfy

$$f^*(s) = \frac{1}{4\gamma}(s+\mu)^T\Sigma^{-1}(s+\mu)$$

for each $s \in \mathbb{R}^n$ and

$$\iota_B^*(s) = \begin{cases} \alpha W & \text{if } s = \alpha\mathbf{1} \text{ for some } \alpha \in \mathbb{R}, \\ \infty & \text{otherwise} \end{cases}$$

for each $s \in \mathbb{R}^n$, respectively. (1+1 p)

h. State the dual problem

$$\underset{s\in\mathbb{R}^n}{\text{minimize}}\, f^*(-s) + \iota_B^*(s) \tag{2}$$

to problem (1) and express it as an optimization problem over a single real variable. (You may assume that the expressions for f^* and ι_B^* in **g.** hold.)

Solve the dual problem (2) over that single variable and relate it to the optimal α in **f.** that comes from the subdifferential in **e.**. Give the dual optimal point $s^* \in \mathbb{R}^n$. (1 p)

i. Given the optimal point $s^* \in \mathbb{R}^n$ of the dual problem (2) in **h.**, show how to recover the primal solution, i.e. the solution to (1). You are allowed to directly use any one of the primal dual necessary and sufficient optimality conditions. Show that the recovered primal solution is the same as in **f.**. (1 p)

Solution

a. Note that

$$\nabla^2 f(x) = 2\gamma\Sigma \succeq 2\gamma\lambda_{\min}(\Sigma)I$$

for each $x \in \mathbb{R}^n$, where $\lambda_{\min}(\Sigma)$ is the smallest eigenvalue of the covariance matrix Σ. The second-order condition for strong convexity gives that f is $2\gamma\lambda_{\min}(\Sigma)$-strongly convex, since $\gamma, \lambda_{\min}(\Sigma) > 0$.

b. The set B is a hyperplane and therefore convex, which implies that ι_B is convex.

c. The objective function in optimization problem (1) can be written as

$$f(x) + \iota_B(x).$$

Since the sum of a strongly convex function and convex function is strongly convex, we conclude that the objective function of (1) is strongly convex. Optimization problem (1) has a unique minimizer since strongly convex functions always have an unique minimizer.

d. We proved above that f is convex. Moreover, f is differentiable with gradient

$$\nabla f(x) = -\mu + 2\gamma\Sigma x$$

for each $x \in \mathbb{R}^n$. The subdifferential of f is then given by

$$\partial f(x) = \{\nabla f(x)\}$$

for each $x \in \mathbb{R}^n$.

e. **Alternative 1:** By definition, the subdifferential of ι_B is given by

$$\partial\iota_B(x) = \Big\{s \in \mathbb{R}^n : \forall y \in \mathbb{R}^n,\ \iota_B(y) \geq \iota_B(x) + s^T(y-x)\Big\} \tag{3}$$

for each $x \in \mathbb{R}^n$.

Case $x \in B$: Suppose that $y \notin B$. For any choice of $s \in \mathbb{R}^n$, the inequality in (3) holds, since $\iota_B(y) = \infty$ and $\iota_B(x) = 0$. Therefore, we only need to further analyze the case when $y \in B$. The inequality in (3) becomes

$$0 \geq s^T(y-x) \tag{4}$$

since $\iota_B(y) = 0$ and $\iota_B(x) = 0$. Moreover, note that $2x - y \in B$, since $\mathbf{1}^T(2x - y) = W$. Therefore, (4) must hold when y is replaced by $2x - y$, i.e.

$$0 \geq s^T(x-y). \tag{5}$$

Combining (4) and (5), we get that $s \in \partial\iota_B(x)$ if and only if

$$0 = s^T(y-x), \tag{6}$$

for each $y \in B$.

First, suppose that $s \in \{\alpha\mathbf{1} : \alpha \in \mathbb{R}\} \subseteq \mathbb{R}^n$. Then (6) holds, since $\mathbf{1}^T(y-x) = 0$ for each $y \in B$, which implies that $\{\alpha\mathbf{1} : \alpha \in \mathbb{R}\} \subseteq \partial\iota_B(x)$.

Second, suppose that $s \in \partial\iota_B(x)$. This implies that (6) holds for this s. Note that (6) can be written as

$$0 = \left(\frac{\mathbf{1}^T s}{n}\mathbf{1} + \left(s - \frac{\mathbf{1}^T s}{n}\mathbf{1}\right)\right)^T (y-x) = \left(s - \frac{\mathbf{1}^T s}{n}\mathbf{1}\right)^T (y-x), \tag{7}$$

for each $y \in B$. Define the vector $u \in \mathbb{R}^n$ by $u = s - (\mathbf{1}^T s/n)\mathbf{1}$. Note that $x - u \in B$ since $\mathbf{1}^T u = 0$. Therefore, (7) must hold when y is replaced by $x - u$, which gives

$$0 = ||u||_2^2 \quad \Leftrightarrow \quad u = 0 \quad \Leftrightarrow \quad s = \frac{\mathbf{1}^T s}{n}\mathbf{1}.$$

This shows that $s \in \{\alpha\mathbf{1} : \alpha \in \mathbb{R}\}$. We conclude that $\{\alpha\mathbf{1} : \alpha \in \mathbb{R}\} = \partial\iota_B(x)$.

Case $x \notin B$: For any choice of $s \in \mathbb{R}^n$, the inequality in (3) fails for the choice $y = (W/n)\mathbf{1} \in B$, since $\iota_B(y) = 0$ and $\iota_B(x) = \infty$. Thus, $\partial\iota_B(x) = \emptyset$.

Summary: Summarizing, we get that

$$\partial\iota_B(x) = \begin{cases} \{\alpha\mathbf{1} : \alpha \in \mathbb{R}\} & \text{if } x \in B, \\ \emptyset & \text{if } x \notin B \end{cases}$$

as desired.

Alternative 2: Note that

$$\begin{aligned} \iota_B(x) &= \iota_{\{W\}}(\mathbf{1}^T x) \\ &= (\iota_{\{W\}} \circ \mathbf{1}^T)(x) \end{aligned}$$

for each $x \in \mathbb{R}^n$. Also, we have that

$$\partial\iota_{\{W\}}(a) = \begin{cases} \mathbb{R} & \text{if } a = W, \\ \emptyset & \text{if } a \neq W \end{cases}$$

for each $a \in \mathbb{R}$. Since

$$\begin{aligned} \operatorname{relint}\operatorname{dom}\iota_{\{W\}} \circ \mathbf{1}^T &= \operatorname{relint} B \\ &= B \\ &\neq \emptyset \end{aligned}$$

the subdifferential calculus rules give that

$$\begin{aligned} \partial\iota_B(x) &= \partial(\iota_{\{W\}} \circ \mathbf{1}^T)(x) \\ &= \mathbf{1}\partial\iota_{\{W\}}(\mathbf{1}^T x) \\ &= \begin{cases} \{\alpha\mathbf{1} : \alpha \in \mathbb{R}\} & \text{if } x \in B, \\ \emptyset & \text{if } x \notin B \end{cases} \end{aligned}$$

for each $x \in \mathbb{R}^n$, as desired.

f. By Fermat's rule, the point $x \in \mathbb{R}^n$ is the minimizer of optimization problem (1) if and only if

$$0 \in \partial\left(f + \iota_B\right)(x) = \partial f(x) + \partial\iota_B(x).$$

The last equality holds since f and ι_B are (closed) convex and constraint qualification holds since f has full effective domain and B has nonempty relative interior. Using **d.** and **e.**, we get that $x \in \mathbb{R}^n$ is the minimizer of optimization problem (1) if and only if

$$0 = -\mu + 2\gamma\Sigma x + \alpha\mathbf{1} \quad \Leftrightarrow \quad x = \frac{1}{2\gamma}\Sigma^{-1}\left(\mu - \alpha\mathbf{1}\right),$$

for some $\alpha \in \mathbb{R}$ and $x \in B$. Left multiplying with $\mathbf{1}^T$ gives that

$$W = \mathbf{1}^T x = \frac{1}{2\gamma}\left(\mathbf{1}^T\Sigma^{-1}\mu - \alpha\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right) \quad \Leftrightarrow \quad \alpha = \frac{\mathbf{1}^T\Sigma^{-1}\mu - 2\gamma W}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}.$$

Thus, the optimal portfolio is given by

$$x = \frac{1}{2\gamma}\left(\Sigma^{-1}\mu + \frac{2\gamma W - \mathbf{1}^T\Sigma^{-1}\mu}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}\Sigma^{-1}\mathbf{1}\right).$$

g. Let $s \in \mathbb{R}^n$. We have that

$$\begin{aligned} f^*(s) &= \sup_{x\in\mathbb{R}^n} \left(s^T x - f(x)\right) \\ &= \sup_{x\in\mathbb{R}^n} \left(s^T x + \mu^T x - \gamma x^T \Sigma x\right) \\ &= -\inf_{x\in\mathbb{R}^n} \underbrace{\left(-s^T x - \mu^T x + \gamma x^T \Sigma x\right)}_{=g(x)} \\ &= -\inf_{x\in\mathbb{R}^n} g(x). \end{aligned}$$

Note that g has gradient

$$\nabla g(x) = -s - \mu + 2\gamma\Sigma x$$

for each $x \in \mathbb{R}^n$ and is convex, since

$$\nabla^2 g(x) = 2\gamma\Sigma \succ 0$$

for each $x \in \mathbb{R}^n$. By Fermat's rule, $x^* \in \mathbb{R}^n$ is a minimizer of g if and only if

$$\begin{gathered} 0 \in \partial g(x^*) = \{\nabla g(x^*)\} \\ \Leftrightarrow \\ 0 = -s - \mu + 2\gamma\Sigma x^* \\ \Leftrightarrow \\ x^* = \frac{1}{2\gamma}\Sigma^{-1}(s+\mu). \end{gathered}$$

We get that

$$g(x^*) = -\frac{1}{4\gamma}\left(s+\mu\right)^T \Sigma^{-1}\left(s+\mu\right)$$

and therefore

$$\begin{aligned} f^*(s) &= -g(x^*) \\ &= \frac{1}{4\gamma}\left(s+\mu\right)^T \Sigma^{-1}\left(s+\mu\right) \end{aligned}$$

as desired.

We also have that

$$\begin{aligned} \iota_B^*(s) &= \sup_{x\in\mathbb{R}^n} \left(s^T x - \iota_B(x)\right) \\ &= \sup_{x\in B} s^T x \end{aligned}$$

for each $s \in \mathbb{R}^n$. First, suppose that $s = \alpha\mathbf{1} \in \mathbb{R}^n$, for some $\alpha \in \mathbb{R}$. Then, $\iota_B^*(s) = \alpha W$. Next, suppose that $s \in \mathbb{R}^n \setminus \{\alpha\mathbf{1} : \alpha \in \mathbb{R}\}$. Let $x = t(s - (\mathbf{1}^T s/n)\mathbf{1}) + (W/n)\mathbf{1} \in B$ and $t > 0$. Note that

$$s^T x = t\underbrace{\left(\|s\|_2^2 - \frac{(\mathbf{1}^T s)^2}{n}\right)}_{>0} + \frac{W}{n}\mathbf{1}^T s \to \infty \quad \text{as} \quad t \to \infty,$$

by the Cauchy-Schwarz inequality and the assumption that $s \in \mathbb{R}^n \setminus \{\alpha\mathbf{1} : \alpha \in \mathbb{R}\}$. I.e. $\iota_B^*(s) = \infty$. We summarize the cases as

$$\iota_B^*(s) = \begin{cases} \alpha W & \text{if } s = \alpha\mathbf{1} \text{ for some } \alpha \in \mathbb{R}, \\ \infty & \text{otherwise} \end{cases}$$

for each $s \in \mathbb{R}^n$, as desired.

h. The dual problem is

$$\underset{s\in\mathbb{R}^n}{\text{minimize}}\, f^*(-s) + \iota_B^*(s) = \underset{\alpha\in\mathbb{R}}{\text{minimize}}\, \underbrace{\frac{1}{4\gamma}\left(\alpha\mathbf{1} - \mu\right)^T \Sigma^{-1}\left(\alpha\mathbf{1} - \mu\right) + \alpha W}_{h(\alpha)}.$$

The function h is a convex quadratic in α. By Fermat's rule and that the subdifferential of a convex differentiable function contains only the gradient, $\alpha^\star \in \mathbb{R}$ minimizes h if and only if

$$\begin{gathered} 0 = \nabla h(\alpha^\star) \\ \Leftrightarrow \\ 0 = \frac{1}{2\gamma}\mathbf{1}^T\Sigma^{-1}(\alpha^\star\mathbf{1} - \mu) + W \\ \Leftrightarrow \\ \alpha^\star = \frac{\mathbf{1}^T\Sigma^{-1}\mu - 2\gamma W}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}} \end{gathered}$$

which is the same as the optimal α in **f.**. The dual optimal variable $s^* \in \mathbb{R}^n$ is then

$$s^* = \frac{\mathbf{1}^T\Sigma^{-1}\mu - 2\gamma W}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}\mathbf{1}.$$

i. We use the primal dual necessary and sufficient optimality condition

$$\begin{cases} x^* \in \partial f^*(-s^*) \\ x^* \in \partial \iota_B^*(s^*). \end{cases}$$

We will use the first condition in this pair to extract the optimal primal variable $x^* \in \mathbb{R}^n$. Recall that the conjugate f^* always is convex. Moreover, f^* is differentiable with gradient

$$\nabla f^*(s) = \frac{1}{2\gamma}\Sigma^{-1}\left(s + \mu\right)$$

for each $s \in \mathbb{R}^n$. The subdifferential of f^* is then given by

$$\partial f^*(s) = \{\nabla f^*(s)\}$$

for each $s \in \mathbb{R}^n$. The first condition then gives that the primal solution is

$$\begin{aligned} x^* &= \nabla f^*(-s^*) \\ &= \frac{1}{2\gamma}\left(\Sigma^{-1}\mu + \frac{2\gamma W - \mathbf{1}^T\Sigma^{-1}\mu}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}\Sigma^{-1}\mathbf{1}\right) \end{aligned}$$

i.e. the same as in **f.**

5. Consider the 1-norm regularized SVM problem

$$\underset{w\in\mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{N} \underbrace{\max\left(0, 1 - y_i w^T x_i\right)}_{=f_i(w)} + \lambda \|w\|_1 \tag{8}$$

given the labeled training data set $\{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$ are training data and labels, respectively.

a. Find the smallest nonnegative constant $\lambda_0 \in \mathbb{R}$ such that if $\lambda \geq \lambda_0$, then

$$w = 0$$

is an optimal point for (8). (2 p)

b. Is the proximal gradient method applicable to find a solution of problem (8)? Is it applicable to solve a corresponding Fenchel dual problem? (1 p)

Solution

a. Since each function involved in the objective function of (8) is convex and CQ holds (all functions involved have full domain), Fermat's rule implies that $w \in \mathbb{R}^n$ is an optimal point for (8) if and only if

$$0 \in \sum_{i=1}^{N} \partial f_i(w) + \lambda \partial(\|\cdot\|_1)(w).$$

Since we are only interested in the case for which $w = 0$ is a solution, we evaluate the subdifferentials only at this point. Note that

$$\|w\|_1 = \sum_{i=1}^{n} |w_i|$$

for each $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$. Since

$$\partial(|\cdot|)(0) = [-1, 1]$$

the subdifferential of $\|\cdot\|_1$ at 0 is

$$\begin{aligned}\partial(\|\cdot\|_1)(0) &= \begin{bmatrix} \partial(|\cdot|)(0) \\ \vdots \\ \partial(|\cdot|)(0) \end{bmatrix} \\ &= \begin{bmatrix} [-1, 1] \\ \vdots \\ [-1, 1] \end{bmatrix} \\ &= [-1, 1]^n.\end{aligned}$$

Let $h : \mathbb{R} \to \mathbb{R}$ such that

$$h(v) = \max(0, 1 - v)$$

for each $v \in \mathbb{R}$. Then

$$f_i(w) = h\left(y_i x_i^T w\right)$$

for each $w \in \mathbb{R}^n$ and for each $i = 1, \ldots, N$. Since h is convex and the CQ holds, we have

$$\partial f_i(w) = y_i x_i \partial h_i \left(y_i x_i^T w\right)$$

for each $w \in \mathbb{R}^n$ and for each $i = 1, \ldots, N$. This implies that

$$\begin{aligned}\partial f_i(0) &= y_i x_i \partial h_i(0) \\ &= \{-y_i x_i\}\end{aligned}$$

for each $i = 1, \ldots, N$ and we get

$$\sum_{i=1}^{N} \partial f_i(0) = \{-Xy\}$$

where

$$y = (y_1, \ldots, y_N) \quad \text{and} \quad X = \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix}.$$

Therefore, according to the optimality condition, $w = 0$ is an optimal point for (8) if and only if

$$Xy \in \lambda[-1, 1]^n.$$

This holds if and only if

$$\lambda \geq \max_{i=1,\ldots,n} |(Xy)_i| = \|Xy\|_\infty = \lambda_0.$$

b. Neither of the functions in the objective are smooth. Thus, the proximal gradient method is not applicable. Since neither of the functions in the objective are strongly convex, their associated conjugate functions are not smooth. Therefore, the proximal gradient is not applicable to the dual problem either.