LUND UNIVERSITY

Department of
AUTOMATIC CONTROL

Exam in optimization for learning

2022-10-25

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.
- Write your anonymous code and personal identifier on each solution sheet.

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

- All lecture slides
- Mathematical prerequisites document

You may use the results in the lecture slides and the mathematical prerequisites document unless the opposite is explicitly stated.

Submission

Scan all your solution sheets using a scanner app on your phone and save as a single PDF file. Upload this PDF file to the Canvas course webpage. You will have 15 minutes to complete this part after the official end of the exam. You must also hand in your physical solution sheets at the end of the exam.

Results

Suggested solutions will be posted on the Canvas course webpage after the exam. The exam will be graded anonymously via Canvas, and your graded submission will be made visible in Canvas once we are done with the grading. Results will be registered in LADOK.

Remark: In this exam you are allowed to assume that norms are convex and that the function $h_p : \mathbb{R} \to \mathbb{R}$ such that

$$h_p(z) = \max(0, z)^p \tag{1}$$

for each $z \in \mathbb{R}$ is convex and nondecreasing, for any $p \geq 1$.

1. Show that the following sets are convex:

a. $S_1 = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 : (x_1 - x_2)^2 + (x_3 - x_4)^2 \leq \pi\}$. (1 p)

b. $S_2 = \{(x, y) \in \mathbb{R}^2 : x \geq y^2\}$. (1 p)

Show that the following sets are nonconvex:

c. $S_3 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. (1 p)

d. $S_4 = \{(x, t) \in \mathbb{R}^2 : \exists y \in \mathbb{R} \text{ such that } |t| \leq y \text{ and } y^2 = x^3 - x\}$. (1 p)

e. $S_5 = \{x \in \mathbb{R}^2 : 1 \leq \|x\|_\infty \leq 10\}$. (1 p)

2. Show that the following functions are convex:

a. $f_1 : \mathbb{R}^2 \to \mathbb{R}$ such that

$$f_1(x) = x^T \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} x$$

for each $x \in \mathbb{R}^2$. (1 p)

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

$$f_2(x) = \int_0^x e^{e^t} \mathrm{d}t$$

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

Hint: Recall that the fundamental theorem of calculus gives that if a function $g : \mathbb{R} \to \mathbb{R}$ is continuous then

$$\frac{\mathrm{d}}{\mathrm{d}x} \int_0^x g(t) \mathrm{d}t = g(x)$$

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

c. $f_3 : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ such that

$$f_3(x) = \sup_{y \in \mathbb{R}} (xy - \cos y)$$

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

Show that the following functions are nonconvex:

d. $f_4 : \mathbb{S}^2 \to \mathbb{R}$ such that

$$f_4(X) = \lambda_{\min}(X)$$

for each $X \in \mathbb{S}^2$, where $\lambda_{\min}$ denotes the smallest eigenvalue. (1 p)

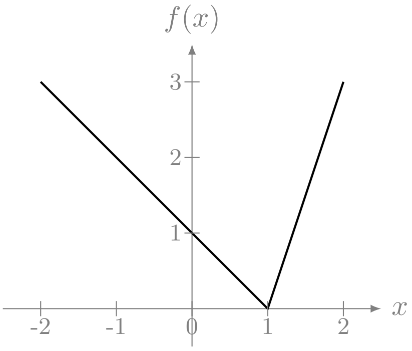

Figure 1 The function f in Problem 3

e. $f_5 : \mathbb{R} \to \mathbb{R}$ such that

$$f_5(x) = \min(x^2 + 10, -x^2)$$

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

3. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined as

$$f(x) = \begin{cases} -x + 1 & \text{if } x < 1, \\ 3x - 3 & \text{if } x \geq 1. \end{cases}$$

See Figure 3.

a. Compute the subdifferential ∂f. (1 p)

b. Compute prox_f. (1 p)

c. Compute the conjugate function f^*. (1 p)

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

e. Find a function $g : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ that is not equal to f^* but that satisfies $g^* = f$. You are allowed to used graphical arguments in this subproblem. (1 p)

4. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper, closed and convex.

a. Show that

$$f^*(0) < \infty$$

implies that the infimum of f is finite. (1 p)

b. Let $n = 1$. Suppose that

$$f^*(0) < \infty$$

and that there exists no $x \in \mathbb{R}$ such that

$$0 \in \partial f(x).$$

Find an example of a function f that satisfies this. (1 p)

5. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \frac{1 - \cos(2\pi x)}{2} x^2 + x^2$$

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

a. Let $g : \mathbb{R} \to \mathbb{R}$ be the mapping $x \mapsto x^2$. Show that

$$g \leq \text{env } f.$$

Hint: Show that $g \leq f$ and that g is convex. (1 p)

b. Compute $(\text{env } f)(10)$.
Hint: Use $\text{env } f \leq f$. (1 p)

6. Consider a two-layer feed-forward neural network without bias terms, $m(\cdot\,;\theta) : \mathbb{R} \to \mathbb{R}$, defined as

$$m(x;\theta) = W_2\sigma_1(W_1 x)$$

for each $x \in \mathbb{R}$, where $\theta = (W_1, W_2) \in \mathbb{R}^2$ contains the two parameters of the neural network and $\sigma_1 : \mathbb{R} \to \mathbb{R}$ is some activation function. Moreover, consider the training problem

$$\underset{\theta \in \mathbb{R}^2}{\text{minimize}} \sum_{i=1}^{N} L(m(x_i;\theta), y_i) \tag{2}$$

over some training set $\{(x_i, y_i)\}_{i=1}^N$ where $(x_i, y_i) \in \mathbb{R}^2$ for each $i = 1, \ldots, N$ and $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is some loss function.

a. Assume that $\sigma_1 : \mathbb{R} \to \mathbb{R}$ is the ReLU activation function, i.e.,

$$\sigma_1(x) = \max(0, x)$$

for each $x \in \mathbb{R}$, that W_1 is fixed, that $L(\cdot, \cdot)$ is convex in the first argument, and that we optimize over W_2 only. Define the feature vector

$$\phi(x_i) = \sigma_1(W_1 x_i)$$

for each training point $i = 1, \ldots, N$. The training problem (2) can then be written as

$$\underset{W_2 \in \mathbb{R}}{\text{minimize}} \sum_{i=1}^{N} L(W_2\phi(x_i), y_i). \tag{3}$$

That is, we optimize only over the weights in the final layer. Is (3) a convex optimization problem? (1 p)

b. Assume instead that $\sigma_1 : \mathbb{R} \to \mathbb{R}$ is the identity mapping, i.e., $\sigma_1 = \mathrm{Id}$, that $N = 1$, that $(x_1, y_1) = (1, 0)$, and that

$$L(u, y) = \frac{1}{2}(u - y)^2$$

for each $(u, y) \in \mathbb{R}^2$. The training problem (2) can then be written as

$$\underset{(W_1,W_2)\in\mathbb{R}^2}{\text{minimize}}\ \frac{1}{2}(W_2 W_1)^2. \tag{4}$$

Let $f : \mathbb{R}^2 \to \mathbb{R}$ denote the objective function in (4), i.e.,

$$f(\theta) = \frac{1}{2}(W_2 W_1)^2 \tag{5}$$

for each $\theta = (W_1, W_2) \in \mathbb{R}^2$. Show that f is nonconvex. (1 p)

c. What is the optimal value of the training problem defined in (4)? (1 p)

d. Let S_s denote the set of all stationary points (that includes all local minima) to (4), i.e.,

$$S_s = \{\theta \in \mathbb{R}^2 : \nabla f(\theta) = 0\},$$

where f is defined in (5). Let S_g denote the set of all globally optimal points to (4), i.e.,

$$S_g = \{\theta \in \mathbb{R}^2 : f(\theta) = f^*\},$$

where f^* is the optimal value of (4) computed in subproblem **c.**.

Show that S_s and S_g are equal with the common value

$$\{\, (W_1, W_2) \in \mathbb{R}^2 : W_1 = 0 \text{ or } W_2 = 0 \,\}.$$

I.e., all stationary points for the nonconvex training problem are global minima.

Hint: The gradient of the function f in (5) satisfies

$$\nabla f(\theta) = \begin{bmatrix} W_1 W_2^2 \\ W_1^2 W_2 \end{bmatrix} \tag{6}$$

for each $\theta = (W_1, W_2) \in \mathbb{R}^2$. (1 p)

e. Let $\gamma \in \mathbb{R}$. Consider a gradient update

$$\theta^{(+)} = \theta - \gamma \nabla f(\theta),$$

applied to f in (5), where we start from a point $\theta = (W_1, W_2) \in \mathbb{R}^2$ and go to the point $\theta^{(+)} = (W_1^{(+)}, W_2^{(+)}) \in \mathbb{R}^2$.

Suppose that

$$(W_1, W_2) \in \mathbb{R}^2_{++},$$

i.e., $W_1 > 0$ and $W_2 > 0$. Provide bounds on γ such that

$$0 < W_1^{(+)} < W_1 \quad \text{and} \quad 0 < W_2^{(+)} < W_2.$$

(1 p)

f. Consider f in (5). It can be shown that

$$\begin{cases} f(\theta_1) \leq f(\theta_2) + \nabla f(\theta_2)^T(\theta_1 - \theta_2) + \dfrac{3}{2}\left\|\theta_1 - \theta_2\right\|_2^2, \\ f(\theta_1) \geq f(\theta_2) + \nabla f(\theta_2)^T(\theta_1 - \theta_2) - \dfrac{3}{2}\left\|\theta_1 - \theta_2\right\|_2^2 \end{cases} \tag{7}$$

holds for each $\theta_1, \theta_2 \in \mathbb{R}^2$ such that

$$\left\|\theta_1\right\|_\infty \leq 1 \quad \text{and} \quad \left\|\theta_2\right\|_\infty \leq 1.$$

Remark: Such a function is said to be *locally 3-smooth* on the set

$$\{\theta \in \mathbb{R}^2 : \left\|\theta\right\|_\infty \leq 1\}.$$

Now, let $\theta^{(0)} \in \mathbb{R}^2_{++}$ such that $\left\|\theta^{(0)}\right\|_\infty \leq 1$ and $\gamma \in (0, 2/3)$. The gradient method then defines the update

$$\theta^{(k+1)} = \theta^{(k)} - \gamma \nabla f\left(\theta^{(k)}\right)$$

for each integer $k \geq 0$. Show that

$$\left\|\nabla f\left(\theta^{(k)}\right)\right\|_2 \to 0 \quad \text{as} \quad k \to \infty.$$

Hint: If $\gamma \in (0, 2/3)$, **e.** implies that

$$\theta^{(k)} \in \mathbb{R}^2_{++} \quad \text{and} \quad \left\|\theta^{(k)}\right\|_\infty \leq 1$$

for each integer $k \geq 0$. This can be shown using induction. You are free to use this result here. (1 p)