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

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

Solution

a. Note that

$$S_1 = \left\{x \in \mathbb{R}^4 : \|Ax\|_2^2 \leq \pi\right\}$$

where

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

Convexity of S_1 follows as it is the π-sublevel set of the convex function

$$x \mapsto \|Ax\|_2^2$$

from $\mathbb{R}^4$ to $\mathbb{R}$. Indeed, the norm $\|\cdot\|_2$ is a convex function, the mapping

$$y \mapsto h_2(\|y\|_2) = \|y\|_2^2$$

from $\mathbb{R}^2$ to $\mathbb{R}$ is convex since it is a composition of the convex and nondecreasing function h_2 defined in (1) and the convex function $\|\cdot\|_2$, and the mapping

$$x \mapsto \left(\|\cdot\|_2^2 \circ A\right)(x) = \|Ax\|_2^2$$

from $\mathbb{R}^4$ to $\mathbb{R}$ is convex since it is the composition of the convex function $y \mapsto \|y\|_2^2$ from $\mathbb{R}^2$ to $\mathbb{R}$ and the affine mapping $x \mapsto Ax$ from $\mathbb{R}^4$ to $\mathbb{R}^2$.

b. *Alternative 1*: Note that S_2 is the 0-sublevel set of the convex function $(x, y) \mapsto y^2 - x$ from $\mathbb{R}^2$ to $\mathbb{R}$. This can be verified using, e.g., the second-order condition for convexity. Therefore, S_2 is convex.

Alternative 2: Let $f : \mathbb{R} \to \mathbb{R}$ be the convex function $x \mapsto x^2$. Define the set

$$C = \{\, (x, y) \in \mathbb{R}^2 : x^2 \leq y \,\}$$

and note that

$$\begin{aligned} C &= \{\, (x, y) \in \mathbb{R}^2 : f(x) \leq y \,\} \\ &= \text{epi} f. \end{aligned}$$

Thus, C is convex since it is equal to the epigraph of a convex function Moreover, note that

$$S_2 = T(C)$$

where T is the affine map $T : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$T(x, y) = (y, x)$$

for each $(x, y) \in \mathbb{R}^2$. Convexity of S_2 follows from that it is the image set of an affine map applied to a convex set.

c. Note that $1, 2 \in S_3$, but

$$0.5 \cdot 1 + (1 - 0.5) \cdot 2 = 1.5 \notin S_3,$$

which shows that S_3 is not convex.

d. Consider the points $a_0 = (0, 0)$ and $a_1 = (1, 0)$ that both lie in S_4. Create the convex combination

$$\begin{aligned} a &= \frac{1}{2} a_0 + \frac{1}{2} a_1 \\ &= \left(\frac{1}{2}, 0 \right). \end{aligned}$$

We claim that a is not in S_4.

This follows since for $x \in (0, 1)$ we have that $x^3 < x$ and so $x^3 - x < 0 \leq y^2$ for each $y \in \mathbb{R}$. Therefore $a \notin S_4$. This show that S_4 is not convex.

e. Let $\{e_1, e_2\}$ denote the standard basis in $\mathbb{R}^2$, where e_i is the vector whose ith coordinate is one while all the others are zeros. Then $e_1 \in S_5$ and $-e_1 \in S_5$. However,

$$0.5 e_1 + (1 - 0.5)(-e_1) = 0 \notin S_5,$$

which shows that S_5 is not convex.

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)

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)

Solution

a. The Hessian of f_1 is given by

$$\nabla^2 f_1(x) = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

for each $x \in \mathbb{R}^2$, which has positive eigenvalues 1 and 3. Therefore, the Hessian is positive definite for each $x \in \mathbb{R}^2$. We conclude that f is convex by the second-order condition for convexity.

b. Let $g : \mathbb{R} \to \mathbb{R}$ such that $t \mapsto e^{e^t}$. The function g is a composition of two continuous functions and is therefore also continuous. The fundamental theorem of calculus gives that

$$\begin{aligned} \nabla f_2(x) &= \nabla \int_0^x g(t)\mathrm{d}t \\ &= g(x) \end{aligned}$$

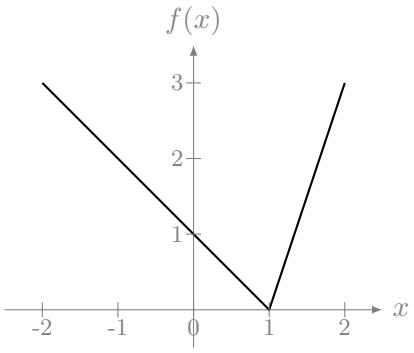

Figure 1 The function f in Problem 3

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

$$\begin{aligned}\nabla^2 f_2(x) &= e^{e^x} e^x \\ &= e^{e^x + x} \\ &> 0\end{aligned}$$

for each $x \in \mathbb{R}$. We conclude that f is convex by the second-order condition for convexity.

c. The function f_3 is equal to the conjugate function of $\cos(\cdot)$ and we know that all conjugate functions are convex. Therefore f_3 is convex.

d. Let

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then X and Y both have 0 as the smallest eigenvalue and the convex combination $0.5X + (1 - 0.5)Y = 0.5I$ has 0.5 as the smallest eigenvalue. Therefore,

$$0.5 = f_4(0.5X + (1 - 0.5)Y) > 0.5 f_4(X) + (1 - 0.5) f_4(Y) = 0.$$

This shows that f_4 is not convex.

e. Note that $f_5(x) = \min(x^2 + 10, -x^2) = -x^2$ and that this is a concave function that is not convex. Hence f_5 is not convex.

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 $\operatorname{prox}_{f^*}$. (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)

Solution

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

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

Thus,

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

Moreover, recall that ∂f is maximally monotone. Thus, $\partial f(1) = [-1, 3]$. Therefore, we conclude that

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 1, \\ [-1, 3] & \text{if } x = 1, \\ \{3\} & \text{if } x > 1. \end{cases}$$

b. Let $z \in \mathbb{R}$ and

$$\begin{aligned} x &= \operatorname{prox}_f(z) \\ &= \operatorname*{argmin}_{\tilde{x} \in \mathbb{R}} \left(f(\tilde{x}) + \frac{1}{2} \|\tilde{x} - z\|_2^2 \right). \end{aligned}$$

Fermat's rule gives that this is equivalent to that

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

Plugging in the expression for ∂f gives

$$0 \in \begin{cases} \{-1 + x - z\} & \text{if } x < 1, \\ [-1, 3] + 1 - z & \text{if } x = 1, \\ \{3 + x - z\} & \text{if } x > 1. \end{cases}$$

Solving for x we get that

$$\begin{aligned} x &= \operatorname{prox}_f(z) \\ &= \begin{cases} z + 1 & \text{if } z < 0, \\ 1 & \text{if } z \in [0, 4], \\ z - 3 & \text{if } z > 4. \end{cases} \end{aligned}$$

c. Let $s \in \mathbb{R}$. Recall that

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

Moreover, Fenchel-Young's equality states that $s \in \partial f(x)$ if and only if

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

where $x \in \mathbb{R}$.

We split into five cases:

- Suppose that $s \in (-1, 3)$. Then $s \in \partial f(x)$ implies that $x = 1$. Fenchel-Young's equality gives that

$$\begin{aligned} f^*(s) &= s \cdot 1 - f(1) \\ &= s - 0 \\ &= s. \end{aligned}$$

- Suppose that $s = -1$. Then $s \in \partial f(x)$ implies that $x \leq 0$. Fenchel-Young's equality gives that

$$\begin{aligned} f^*(s) &= (-1) \cdot x - \underbrace{f(x)}_{=-x+1} \\ &= -1. \end{aligned}$$

- Suppose that $s = 3$. Then $s \in \partial f(x)$ implies that $x \geq 0$. Fenchel-Young's equality gives that

$$\begin{aligned} f^*(s) &= 3 \cdot x - \underbrace{f(x)}_{=3x-3} \\ &= 3. \end{aligned}$$

- Suppose that $s < -1$. Let $t < 1$. Then

$$\begin{aligned} f^*(s) &= \sup_{x \in \mathbb{R}} \left(sx - f(x) \right) \\ &\geq st - f(t) \\ &= st - (-t + 1) \\ &= \underbrace{(s+1)}_{<0} t - 1 \to \infty \quad \text{as} \quad t \to -\infty. \end{aligned}$$

  Thus, $f^*(s) = \infty$ in this case.

- Suppose that $s > 3$. Let $t > 1$. Then

$$\begin{aligned} f^*(s) &= \sup_{x \in \mathbb{R}} \left(sx - f(x) \right) \\ &\geq st - f(t) \\ &= st - (3t - 3) \\ &= \underbrace{(s-3)}_{>0} t + 3 \to \infty \quad \text{as} \quad t \to \infty. \end{aligned}$$

  Thus, $f^*(s) = \infty$ in this case.

This covers all cases and we conclude that

$$f^*(s) = s + \iota_{[-1,3]}.$$

d. Let $z \in \mathbb{R}$. The Moreau decomposition gives that

$$\begin{aligned}
\operatorname{prox}_{f^*}(z) &= z - \operatorname{prox}_f(z) \\
&= z - \begin{cases} z+1 & \text{if } z < 0, \\ 1 & \text{if } z \in [0,4], \\ z-3 & \text{if } z > 4. \end{cases} \\
&= \begin{cases} -1 & \text{if } z < 0, \\ z-1 & \text{if } z \in [0,4], \\ 3 & \text{if } z > 4. \end{cases}
\end{aligned}$$

e. Note that

$$f^{**} = f,$$

since f is proper, closed and convex. Moreover, consider the function $g : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ such that

$$g = \mathrm{Id} + \iota_{\{-1,3\}}.$$

It is clear that

$$g \neq f^*,$$

since $f^* = \mathrm{Id} + \iota_{[-1,3]}$. However, a via graphical arguments (just draw the graph of both functions), one can show that g and f^* must have the same convex envelope, i.e.,

$$\operatorname{env} g = \operatorname{env} f^*.$$

But then we see that

$$\begin{aligned}
g^* &= (\operatorname{env} g)^* \\
&= (\operatorname{env} f^*)^* \\
&= (f^*)^* \\
&= f^{**} \\
&= f.
\end{aligned}$$

This is what we wanted to show.

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)

Solution

a. Note that

$$\begin{aligned} \infty > f^*(0) &= \sup_{x \in \mathbb{R}^n} \left(0^T x - f(x) \right) \\ &= - \inf_{x \in \mathbb{R}^n} f(x), \end{aligned}$$

or

$$-\infty < \inf_{x \in \mathbb{R}^n} f(x).$$

Since f is proper, we have that

$$\inf_{x \in \mathbb{R}^n} f(x) < \infty.$$

We conclude that

$$-\infty < \inf_{x \in \mathbb{R}^n} f(x) < \infty,$$

as desired.

b. Consider the proper, closed and convex function $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ such that

$$f(x) = \begin{cases} \dfrac{1}{x} & \text{if } x > 0, \\ \infty & \text{if } x \leq 0. \end{cases}$$

Note that

$$\begin{aligned} f^*(0) &= \sup_{x \in \mathbb{R}} \left(0 \cdot x - f(x) \right) \\ &= - \inf_{x \in \mathbb{R}_{++}} f(x) \\ &= 0 \end{aligned}$$

is finite. Moreover, note that

$$\partial f(x) = \begin{cases} \left\{ -\dfrac{1}{x^2} \right\} & \text{if } x > 0, \\ \emptyset & \text{if } x \leq 0, \end{cases}$$

which does not contain 0 for any $x \in \mathbb{R}$.
Moral: The function has a finite infimum but no minimizing argument.

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)

Solution

a. Note that

$$\begin{aligned} g(x) &= x^2 \\ &\leq \underbrace{\frac{1 - \cos(2\pi x)}{2}}_{\geq 0} \underbrace{x^2}_{\geq 0} + x^2 \\ &= f(x) \end{aligned}$$

or

$$g(x) \leq f(x)$$

for each $x \in \mathbb{R}$, since

$$\cos(2\pi x) \in [-1, 1]$$

for each $x \in \mathbb{R}$. Moreover, note that g is convex. This is easily seen from, e.g., the second-order condition for convexity. But then

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

by definition of the convex envelope $\text{env } f$ of f.

b. We know that

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

Since

$$g(10) = 100 \quad \text{and} \quad f(10) = 100,$$

we conclude that

$$(\text{env } f)(10) = 100.$$

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

$$m(x;\theta) = W_2\sigma_1(W_1x)$$

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

$$\phi(x_i) = \sigma_1(W_1x_i)$$

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., $\sigma_1 = \text{Id}$, 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

$$\underset{(W_1,W_2)\in\mathbb{R}^2}{\text{minimize}} \frac{1}{2}(W_2W_1)^2. \tag{4}$$

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

$$f(\theta) = \frac{1}{2}(W_2W_1)^2 \tag{5}$$

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

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

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

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

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

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

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)

Solution

a. Note that

$$W_2 \mapsto W_2\phi(x_i)$$

from $\mathbb{R}$ to $\mathbb{R}$ is linear for each $i = 1, \ldots, N$, and that

$$W_2 \mapsto L(W_2\phi(x_i), y_i)$$

from $\mathbb{R}$ to $\mathbb{R}$ is convex, since it is a convex function composed with a linear function, for each $i = 1, \ldots, N$. Therefore,

$$W_2 \mapsto \sum_{i=1}^{N} L(W_2\phi(x_i), y_i)$$

from $\mathbb{R}$ to $\mathbb{R}$ is convex, since it is a sum of convex function. This show that (3) is a convex optimzation problem.

b. *Alterative 1*: Let $\theta_1 = (1, 0)$ and $\theta_2 = (0, 1)$. Then

$$\begin{aligned} f\left(\frac{1}{2}\theta_1 + \frac{1}{2}\theta_2\right) &= f\left(\frac{1}{2}, \frac{1}{2}\right) \\ &= \frac{1}{2}\left(\frac{1}{2} \cdot \frac{1}{2}\right)^2 \\ &= \frac{1}{32} \end{aligned}$$

and

$$\begin{aligned} \frac{1}{2}f(\theta_1) + \frac{1}{2}f(\theta_2) &= \frac{1}{2} \cdot \frac{1}{2}(1 \cdot 0)^2 + \frac{1}{2} \cdot \frac{1}{2}(0 \cdot 1)^2 \\ &= 0. \end{aligned}$$

In particular,

$$f\left(\frac{1}{2}\theta_1 + \frac{1}{2}\theta_2\right) > \frac{1}{2}f\left(\theta_1\right) + \frac{1}{2}f\left(\theta_2\right),$$

and we conclude that f is not convex.

Alterative 2: The gradient is

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

for each $\theta = (W_1, W_2) \in \mathbb{R}^2$ and the Hessian is

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

for each $\theta = (W_1, W_2) \in \mathbb{R}^2$. If we pick $\theta = (1, 1)$, we get that

$$\nabla^2 f(1,1) = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}.$$

The eigenvalues are -1 and 3. Using the the second-order condition for convexity, we condculde that f is nonconvex since the Hessian $\nabla^2 f(\theta)$ is not positive semidefinite at the point $\theta = (1, 1)$.

c. Note that

$$\begin{aligned} f(\theta) &= \frac{1}{2}(W_2 W_1)^2 \\ &\geq 0 \end{aligned}$$

for each $\theta = (W_1, W_2) \in \mathbb{R}^2$ and that

$$f(0) = 0.$$

Therefore,

$$\min_{\theta \in \mathbb{R}^2} f(\theta) = 0.$$

Let $f^\star = 0$ in the following.

d. Let $\theta = (W_1, W_2) \in \mathbb{R}$. First, note that

$$\begin{gathered} \nabla f(\theta) = 0 \\ \Leftrightarrow \\ \begin{bmatrix} W_1 W_2^2 \\ W_1^2 W_2 \end{bmatrix} = 0 \\ \Leftrightarrow \\ W_1 = 0 \quad \text{or} \quad W_2 = 0. \end{gathered}$$

Second, using $f^* = 0$ from **c.**, we have that

$$\begin{gathered} f(\theta) = f^* \\ \Leftrightarrow \\ \frac{1}{2}(W_2 W_1)^2 = 0 \\ \Leftrightarrow \\ W_1 = 0 \quad \text{or} \quad W_2 = 0. \end{gathered}$$

This prove the claim.

e. Written coordinate-wise, the gradient update is

$$\begin{aligned} W_1^{(+)} &= W_1 - \gamma W_1 W_2^2, \\ W_2^{(+)} &= W_2 - \gamma W_1^2 W_2. \end{aligned}$$

We consider each of the four inequalities in

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

separately.

- Note that

$$\begin{gathered} W_1^{(+)} > 0 \\ \Leftrightarrow \\ W_1 - \gamma W_1 W_2^2 > 0 \\ \Leftrightarrow \\ \gamma < \frac{1}{W_2^2}, \end{gathered}$$

  since $W_1 > 0$ and $W_2 > 0$.
- Note that

$$\begin{gathered} W_2^{(+)} > 0 \\ \Leftrightarrow \\ W_2 - \gamma W_1^2 W_2 > 0 \\ \Leftrightarrow \\ \gamma < \frac{1}{W_1^2}, \end{gathered}$$

  since $W_1 > 0$ and $W_2 > 0$.
- Note that

$$\begin{gathered} W_1^{(+)} < W_1 \\ \Leftrightarrow \\ W_1 - \gamma W_1 W_2^2 < W_1 \\ \Leftrightarrow \\ 0 < \gamma, \end{gathered}$$

  since $W_1 > 0$ and $W_2 > 0$.

- Note that

$$\begin{gathered} W_2^{(+)} < W_2 \\ \Leftrightarrow \\ W_2 - \gamma W_1^2 W_2 < W_2 \\ \Leftrightarrow \\ 0 < \gamma, \end{gathered}$$

since $W_1 > 0$ and $W_2 > 0$.

We conclude that

$$\gamma \in \left(0, \min\left(\frac{1}{W_1^2}, \frac{1}{W_2^2}\right)\right),$$

which is nonempty, satisfy the requirements.

f. The hint gives that

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

holds for each integer $k \geq 0$. In particular, we are free to use the quadratic upper bound in (7) with $\theta_1 = \theta^{(k+1)}$ and $\theta_2 = \theta^{(k)}$:

$$\begin{aligned} f\left(\theta^{(k+1)}\right) &\leq f\left(\theta^{(k)}\right) + \nabla f\left(\theta^{(k)}\right)^T \left(\theta^{(k+1)} - \theta^{(k)}\right) + \frac{3}{2}\left\|\theta^{(k+1)} - \theta^{(k)}\right\|_2^2 \\ &\leq f\left(\theta^{(k)}\right) - \gamma\left\|\nabla f\left(\theta^{(k)}\right)\right\|_2^2 + \frac{3\gamma^2}{2}\left\|\nabla f\left(\theta^{(k)}\right)\right\|_2^2 \\ &= f\left(\theta^{(k)}\right) - \gamma\left(1 - \frac{3\gamma}{2}\right)\left\|\nabla f\left(\theta^{(k)}\right)\right\|_2^2 \end{aligned}$$

for each integer $k \geq 0$. This can be written as

$$\gamma\left(1 - \frac{3\gamma}{2}\right)\left\|\nabla f\left(\theta^{(k)}\right)\right\|_2^2 \leq f\left(\theta^{(k)}\right) - f\left(\theta^{(k+1)}\right)$$

for each integer $k \geq 0$. Summing from $k = 0$ until $k = K$ for some nonnegative interger K gives that

$$\begin{aligned} \sum_{k=0}^{K} \gamma\left(1 - \frac{3\gamma}{2}\right)\left\|\nabla f\left(\theta^{(k)}\right)\right\|_2^2 &\leq \sum_{k=0}^{K} f\left(\theta^{(k)}\right) - f\left(\theta^{(k+1)}\right) \\ &= f\left(\theta^{(0)}\right) - f\left(\theta^{(K+1)}\right) \end{aligned}$$

or

$$\sum_{k=0}^{K}\left\|\nabla f\left(\theta^{(k)}\right)\right\|_2^2 \leq \frac{f\left(\theta^{(0)}\right)}{\gamma\left(1 - \frac{3\gamma}{2}\right)},$$

since $\gamma \in (0, 2/3)$ implies that $\gamma(1 - 3\gamma/2) > 0$ and $f \geq 0$. Letting $K \to \infty$ gives that

$$\sum_{k=0}^{\infty}\left\|\nabla f\left(\theta^{(k)}\right)\right\|_2^2 \leq \frac{f\left(\theta^{(0)}\right)}{\gamma\left(1 - \frac{3\gamma}{2}\right)}.$$

Therefore,

$$\left\|\nabla f\left(\theta^{(k)}\right)\right\|_2^2 \to 0 \quad \text{as} \quad k \to \infty,$$

and we conclude that

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

as desired.