

## Machine Learning Exercise Sheet 06

### Optimization

Exercise sheets consist of two parts: homework and in-class exercises. You solve the homework exercises on your own or with your registered group and upload it to Moodle for a possible grade bonus. The in-class exercises will be solved and explained during the tutorial. You do not have to upload any solutions of the in-class exercises.

### In-class Exercises

**Problem 1:** Prove or disprove whether the following functions  $f : D \rightarrow \mathbb{R}$  are convex

- a)  $D = (1, \infty)$  and  $f(x) = \log(x) - x^3$ ,
- b)  $D = \mathbb{R}^+$  and  $f(x) = -\min\{\log(3x+1), -x^4 - 3x^2 + 8x - 42\}$
- c)  $D = (-10, 10) \times (-10, 10)$  and  $f(x, y) = y \cdot x^3 - y \cdot x^2 + y^2 + y + 4$ .

a) The second derivative of  $f$  is  $\frac{d^2 f(x)}{dx^2} = \frac{d}{dx} \left( \frac{1}{x} - 3x^2 \right) = -\frac{1}{x^2} - 6x$ , which is negative on the given set  $D$  and therefore  $f$  is not convex.

b) Transform min to max  

$$-\min\{\log(3x+1), -x^4 - 3x^2 + 8x - 42\} = \max\{-\log(3x+1), x^4 + 3x^2 - 8x + 42\}.$$

$\max(g_1(x), g_2(x))$  is convex if both  $g_1$  and  $g_2$  are convex on  $D = \mathbb{R}^+$  (see Exercise Sheet 6, Problem 1c).  $g_1(x) = -\log(3x+1)$  is convex since the second derivative is positive on  $\mathbb{R}^+$ :

$$\frac{d^2}{dx^2} (-\log(3x+1)) = \frac{d}{dx} \left( -\frac{3}{3x+1} \right) = \frac{9}{(3x+1)^2} > 0$$

$g_2(x) = x^4 + 3x^2 - 8x + 42$  is also convex:

$$\frac{d^2}{dx^2} (x^4 + 3x^2 - 8x + 42) = \frac{d}{dx} (4x^3 + 6x - 8) = 12x^2 + 6 > 0$$

Thus  $f$  is convex.

- c) For the function  $f(x, y)$  to be convex (on  $D$ ) it has to hold for all  $x_1, x_2, y \in D$  and  $\lambda \in (0, 1)$  that

$$\lambda f(x_1, y) + (1 - \lambda)f(x_2, y) \geq f(\lambda x_1 + (1 - \lambda)x_2, y).$$

It does not hold in our case, consider  $y = 1, x_1 = -4, x_2 = 0$  and  $\lambda = 0.5$ :

$$0.5f(-4, 1) + 0.5f(0, 1) = 0.5 \cdot (-74) + 0.5 \cdot 6 = -34$$

$$f(0.5 \cdot (-4) + 0.5 \cdot 0, 0.5 \cdot 1 + 0.5 \cdot 1) = f(-2, 1) = -6 > -34$$

Thus  $f(x, y)$  is not convex.

**Problem 2:** Prove that the following function (the loss function of logistic regression)  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex:

$$f(\mathbf{w}) = -\ln p(\mathbf{y} \mid \mathbf{w}, \mathbf{X}) = -\sum_{i=1}^N (y_i \ln \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \ln(1 - \sigma(\mathbf{w}^T \mathbf{x}_i))) .$$

First, let's simplify the above expression. For this we will need the following two facts

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z} \quad \text{and} \quad 1 - \sigma(z) = \sigma(-z) = \frac{1}{1 + e^z},$$

which implies that

$$\ln \sigma(z) = \ln \left( \frac{e^z}{1 + e^z} \right) = z - \ln(1 + e^z) \quad \text{and} \quad \ln(1 - \sigma(z)) = -\ln(1 + e^z).$$

Plugging this into the definition of the loss function we obtain

$$f(\mathbf{w}) = -\sum_{i=1}^N (y_i \ln \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \ln(1 - \sigma(\mathbf{w}^T \mathbf{x}_i)))$$

$$= -\sum_{i=1}^N (y_i (\mathbf{w}^T \mathbf{x}_i - \ln(1 + e^{\mathbf{w}^T \mathbf{x}_i})) - (1 - y_i) \ln(1 + e^{\mathbf{w}^T \mathbf{x}_i})))$$

$$= \sum_{i=1}^N (-y_i (\mathbf{w}^T \mathbf{x}_i) + \ln(1 + e^{\mathbf{w}^T \mathbf{x}_i}))$$

We know that  $\mathbf{w}^T \mathbf{x}_i$  is a convex (and concave) function of  $\mathbf{w}$ . Therefore, the first term  $-y_i(\mathbf{w}^T \mathbf{x}_i)$  is also convex.

Now, if we show that  $\ln(1 + e^z)$  is a nondecreasing and convex function of  $z$  on  $\mathbb{R}$ , we will be able to use the convexity preserving operations to prove that  $f(\mathbf{w})$  is convex.

The first derivative of  $\ln(1 + e^z)$  is

$$\frac{d}{dz} \ln(1 + e^z) = \frac{e^z}{1 + e^z} = \sigma(z),$$

which is positive for all  $z \in \mathbb{R}$ , which means that  $\ln(1 + e^z)$  is an nondecreasing function.

The second derivative is

$$\frac{d^2}{dz^2} \ln(1 + e^z) = \frac{d}{dz} \sigma(z) = \sigma(z)\sigma(-z),$$

which is also positive for all  $z \in \mathbb{R}$ , which means that  $\ln(1 + e^z)$  is a convex function.

Using the following two facts

1. Sum of convex functions is convex
2. Composition of a convex function with a convex nondecreasing function is convex

we can verify that  $f(\mathbf{w})$  is indeed convex in  $\mathbf{w}$  on  $\mathbb{R}^d$ .

**Problem 3:** Prove that for differentiable convex functions each local minimum is a global minimum. More specifically, given a differentiable convex function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , prove that

- a) if  $\mathbf{x}^*$  is a local minimum, then  $\nabla f(\mathbf{x}^*) = \mathbf{0}$ .
- b) if  $\nabla f(\mathbf{x}^*) = \mathbf{0}$ , then  $\mathbf{x}^*$  is a global minimum.

We will show that if the gradient at some point  $\mathbf{x}^*$  is not equal to zero, then this point cannot be a local optimum — we could simply follow the direction of the negative gradient and end up in a point with a lower value of the function.

More formally, suppose  $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$  for some  $\mathbf{x}^*$ . Then by Taylor's theorem for a sufficiently small  $\varepsilon > 0$  we get

$$\begin{aligned} f(\mathbf{x}^* - \varepsilon \nabla f(\mathbf{x}^*)) &= f(\mathbf{x}^*) - (\varepsilon \nabla f(\mathbf{x}^*))^T \nabla f(\mathbf{x}^*) + O(\varepsilon^2 \|\nabla^2 f(\mathbf{x}^*)\|_2^2) \\ &= f(\mathbf{x}^*) - \varepsilon \|\nabla f(\mathbf{x}^*)\|_2^2 + O(\varepsilon^2 \|\nabla f(\mathbf{x}^*)\|_2^2) \\ &< f(\mathbf{x}^*) \end{aligned}$$

Which means that  $\mathbf{x}^*$  is not a local optimum. Therefore, the gradient must be equal to zero for any local optimum  $\mathbf{x}^*$ .

We will prove (b) using the first order criterion for convexity.

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^T \nabla f(\mathbf{x}).$$

If we plug in  $\mathbf{x}^*$  and use the fact that  $\nabla f(\mathbf{x}^*) = \mathbf{0}$  we get:  $f(\mathbf{y}) \geq f(\mathbf{x}^*)$  for all  $\mathbf{y}$ , meaning  $\mathbf{x}^*$  is a global minimum.

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

## Homework

### 1 Convexity of functions

**Problem 4:** Assume that  $f : \mathbb{R} \rightarrow \mathbb{R}$  and  $g : \mathbb{R} \rightarrow \mathbb{R}$  are convex functions. Prove or disprove the following statements:

- a) The function  $h(x) = g(f(x))$  is convex.
- b) The function  $h(x) = g(f(x))$  is convex if  $g$  is non-decreasing.

*Note: For this exercise you are not allowed to use the convexity preserving operations from the lecture.*

- a) Statement is false. Proof by counterexample: Suppose  $f(x) = x^2$  and  $g(z) = -x$ . Since the function  $h(x) = g(f(x)) = -x^2$  is twice differentiable, we can inspect its second derivative:

$$\frac{d^2}{dx^2} h(x) = -2.$$

Since the second derivative is negative for all  $x$ , we conclude that the function  $h$  is not convex.

(Note: It would actually be sufficient to show that the second derivative is negative for a single value of  $x$ )

- b) Statement is true. Suppose  $x_0, x_1 \in \mathbb{R}$  and  $\lambda \in (0, 1)$ . We will use a shorthand notation  $x_\lambda = \lambda x_1 + (1 - \lambda)x_0$ .

We will prove the convexity of  $h$  using the definition of convexity and the properties of  $f$  and  $g$ :

$$\begin{aligned} f \text{ convex} &\Rightarrow f(x_\lambda) \leq \lambda f(x_1) + (1 - \lambda)f(x_0) \\ g \text{ non-decreasing} &\Rightarrow g(f(x_\lambda)) \leq g(\lambda f(x_1) + (1 - \lambda)f(x_0)) & (1) \\ g \text{ convex} &\Rightarrow g(\lambda f(x_1) + (1 - \lambda)f(x_0)) \leq \lambda g(f(x_1)) + (1 - \lambda)g(f(x_0)) & (2) \\ (1) \text{ and } (2) &\Rightarrow g(f(x_\lambda)) \leq \lambda g(f(x_1)) + (1 - \lambda)g(f(x_0)) \\ &\Leftrightarrow h(x_\lambda) \leq \lambda h(x_1) + (1 - \lambda)h(x_0). \end{aligned}$$

Therefore  $h$  is convex.

### 2 Optimization / Gradient descent

**Problem 5:** You are given the following objective function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$

$$f(x_1, x_2) = 0.5x_1^2 + x_2^2 + 2x_1 + x_2 + \cos(\sin(\sqrt{\pi})).$$

- a) Compute the minimizer  $x^*$  of  $f$  analytically.

As  $f$  is a sum of convex functions, it is convex. To find the global minimizer, we compute the gradient and set it to zero

$$\nabla f(x_1, x_2) = \begin{pmatrix} x_1 + 2 \\ 2x_2 + 1 \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} x_1^* \\ x_2^* \end{pmatrix} = \begin{pmatrix} -2 \\ -\frac{1}{2} \end{pmatrix}.$$

- b) Perform 2 steps of gradient descent on  $f$  starting from the point  $\mathbf{x}^{(0)} = (0, 0)$  with a constant learning rate  $\tau = 1$ .

We already know how to compute the gradient from a).

$$\begin{aligned} \text{first step} \quad \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} &= \begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \end{pmatrix} - \tau \begin{pmatrix} x_1^{(0)} + 2 \\ 2x_2^{(0)} + 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 1 \begin{pmatrix} 0 + 2 \\ 0 + 1 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix} \\ \text{second step} \quad \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \end{pmatrix} &= \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} - \tau \begin{pmatrix} x_1^{(1)} + 2 \\ 2x_2^{(1)} + 1 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix} - 1 \begin{pmatrix} -2 + 2 \\ -2 + 1 \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \end{pmatrix} \end{aligned}$$

## Assignment Project Exam Help

- c) Will the gradient descent procedure from Problem b) ever converge to the true minimizer  $\mathbf{x}^*$ ? Why or why not? If the answer is no, how can we fix it?

By performing one more iteration of gradient descent we observe that

$$\begin{pmatrix} x_1^{(3)} \\ x_2^{(3)} \end{pmatrix} = \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \end{pmatrix} - \tau \begin{pmatrix} x_1^{(2)} + 2 \\ 2x_2^{(2)} + 1 \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \end{pmatrix} - 1 \begin{pmatrix} -2 + 2 \\ 0 + 1 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix}.$$

That is, we are stuck iterating between  $\mathbf{x}^{(1)}$  and  $\mathbf{x}^{(2)}$  forever. We can fix this by decreasing the learning rate (adaptive stepsize, etc.).

**Problem 6:** Load the notebook `exercise_06_notebook.ipynb` from Moodle. Fill in the missing code and run the notebook. Export (download) the evaluated notebook as PDF and add it to your submission.

*Note: We suggest that you use Anaconda for installing Python and Jupyter, as well as for managing packages. We recommend that you use Python 3.*

*For more information on Jupyter notebooks, consult the Jupyter documentation. Instructions for converting the Jupyter notebooks to PDF are provided on Piazza.*

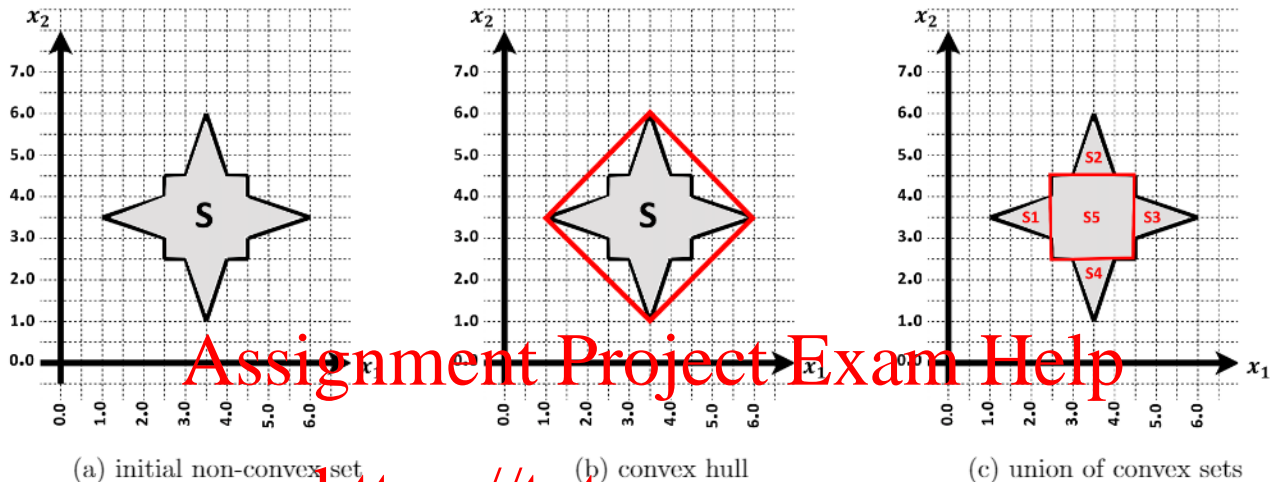
The solution notebook is uploaded to Moodle.



**Problem 7:** Let  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  be the following convex function:

$$f(x_1, x_2) = e^{x_1+x_2} - 5 \cdot \log(x_2)$$

- a) Consider the following shaded region  $S \subset \mathbb{R}^2$ . Is this region convex? Why?
- b) Assume that we are given an algorithm  $\text{ConvOpt}(f, D)$  that takes as input a convex function  $f$  and convex region  $D$ , and returns the minimum of  $f$  over  $D$ . Using the  $\text{ConvOpt}$  algorithm, how would you find the global minimum of  $f$  over the shaded region  $S$ ?



- a) It is not because we can choose two points in  $S$  such that the line connecting the points does not completely reside in  $S$ , for example  $(1.0, 3.5)^T$  and  $(3.5, 6.0)^T$  (see Figure 1b).
- b) We can partition the shaded region  $S$  to the following five convex regions  $S_1, \dots, S_5$  (see Figure 1c). Afterwards, we run the  $\text{ConvOpt}$  algorithm separately for the 5 regions and obtain

$$m_i = \min_{x \in S_i} f(x) = \text{ConvOpt}(f, S_i).$$

Finally, the minimum over the whole  $S$  can be computed as the smallest of these values, that is  $\min_{x \in S} f(x) = \min(m_1, \dots, m_5)$ .

