Machine Learning Exercise Sheet 06

Optimization

Exercise sheets consist of two parts: In-class exercises and homework. The in-class exercises will be solved and discussed during the tutorial. The homework is for you to solve at home and further engage with the lecture content. There is no grade bonus and you do not have to upload any solutions. Note that the order of some exercises might have changed compared to last year's recordings.

In-class Exercises

Problem 1: Prove or disprove whether the following functions $f : D \to \mathbb{R}$ 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 $f(x, y) = y \cdot x^3 - y \cdot x^2 + y^2 + y + 4$.

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

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

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

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

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

P1 a) $f'(x) = \frac{1}{x} - 3x^2$

$f''(x) = -\frac{1}{x^2} - 6x$

$\because x \in (1, +\infty)$

$\therefore f''(x) < 0$. non-convex

b) $f(x) = \max(\underbrace{-\log(3x+1)}_{f_1(x)}, \underbrace{x^4 + 3x^2 - 8x + 42}_{f_2(x)})$

$f_1'(x) = -\frac{3}{3x+1}$ $\quad f_2'(x) = 4x^3 + 6x - 8$

$f_1''(x) = \frac{9}{(3x+1)^2} > 0$ $\quad f_2''(x) = 12x^2 + 6 > 0$

$\therefore f(x)$ is convex

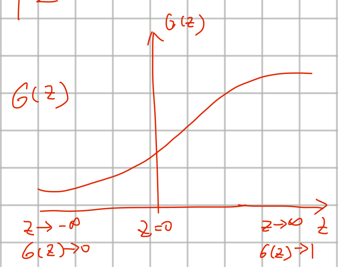
c) $H = \begin{pmatrix} y \cdot 6x - y^2 & 3x^2 - 2x \\ 3x^2 - 2x & 2 \end{pmatrix}$

$\det H = 2y6x - y^2 - (3x^2 - 2x)^2$

$y = 1 \quad x = -4$.

$\det H = -48 - 1 - \cdots < 0$ $\quad \therefore$ not convex

P2



$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$1-\sigma(z) = 1 - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}} = \frac{1}{e^{z}+1} = \sigma(-z)$$

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

$$\ln(1-\sigma(z)) = \ln(\sigma(-z)) = -\ln(1+e^{z})$$

$$f(w) = -\sum_{i=1}^{N}\left[y_i \ln \sigma(w^T x_i) + (1-y_i)\ln(1-\sigma(w^T x_i))\right]$$

$$= -\sum_{i=1}^{N}\left[y_i\left(w^T x_i - \cancel{\ln(1+e^{w^T x_i})}\right) - \cancel{(1-y_i)\ln(1+e^{w^T x_i})}\right]$$

$$= -\sum_{i=1}^{N}\left[y_i w^T x_i - \cancel{y_i \ln(1+e^{w^T x_i})} - \ln(1+e^{w^T x_i}) + \cancel{y_i \ln(1+e^{w^T x_i})}\right]$$

$$= -\sum_{i=1}^{N}\left[-y_i w^T x_i + \ln(1+e^{w^T x_i})\right]$$

$$\frac{\partial}{\partial z} = \frac{1}{1+e^{z}} \cdot e^{z} = \frac{1}{1+e^{-z}} = \sigma(z) > 0$$

$$\frac{\partial}{\partial z}(\sigma(z)) = (1+e^{-z})^{-2} \cdot e^{-z} = \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}}$$

$$= \sigma(z) \cdot \sigma(-z) > 0$$

P3 Given f is continously differentiable, x^* is locdl mimum

$\Rightarrow \nabla f(x^*) = 0$

Then If f is convex, differentiable, domain open set

$\rightarrow \nabla f$ is contine

Taylor $\rightarrow$ $f(x+p) = f(x) + \nabla f(x)^T p + \underbrace{o(\|p\|_2^2)}_{\text{Cauchy form}}$

Given $f: \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, $p \in \mathbb{R}^d$

$\Rightarrow f(x+p) = f(x) + \nabla f(x+tp)^T p \quad t \in (0,1)$

Assume $\nabla f(x^*) \neq 0 \quad p = -\nabla f(x^*)$

$$p^T \nabla f(x^*) = -\nabla f(x^*)^T \nabla f(x^*) = -\|\nabla f(x^*)\|_2^2 < 0$$

$$g(\xi) := p^T \nabla f(x^* + \xi p) = -\nabla f(x^*)^T \nabla f(x^* + \xi p) = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i} f(x^*) \frac{\partial}{\partial x_i} f(x^* + \xi p)$$

$$g(0) = p^T \nabla f(x^*) < 0$$

$$f(x + \xi' p) = f(x) + \nabla f(x + \underbrace{t\xi'}_{:=\xi} p)^T p \xi'$$

$$= f(x) + \nabla f(x + \xi p)^T p \xi'$$

(b) if $\nabla f(x^*) = 0 \rightarrow x^*$ is a global minimum

$x, y \in \mathbb{R}^d$: $f(y) \geq f(x) + (y-x)^T \nabla f(x)$

x^* : $f(y) \geq f(x^*) + \underbrace{(y-x)^T \nabla f(x^*)}_{0}$

$\Rightarrow f(y) \geq f(x^*)$

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

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

$$\sigma = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z} \qquad 1-\sigma = \frac{1}{1+e^z}$$

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

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

$$f(w) = \sum_{i=1}^{N} \left(-y_i\,(w^Tx_i - \ln(1+e^{w^Tx_i})) + (1-y_i)\ln(1+e^{w^Tx_i})\right)$$

$$\left(-y_i w^Tx_i + \cancel{y_i \ln(1+e^{w^Tx_i})} + \ln(1+e^{w^Tx_i}) - \cancel{y_i \ln(1+e^{w^Tx_i})}\right)$$

$$= \sum_{i=1}^{N} \left(-y_i w^Tx_i + \ln(1+e^{w^Tx_i})\right)$$

w^Tx_i is linear $\rightarrow$ convex and concave

$-y_i(w^Tx_i) \longrightarrow$ convex

$$f_2(x) = \ln(1+e^z) \qquad f_2'(x) = \frac{e^z}{1+e^z} = \sigma(z)$$

$$f_2''(x) = \sigma(z)\cdot\sigma(-z) > 0 \rightarrow \text{convex.}$$

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

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

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

If gradient at x^* is not zero, then it's not local optimum

$\nabla f(x^*) \neq 0 \rightarrow \varepsilon > 0$

$$f(x^* - \varepsilon\nabla f(x^*)) = f(x^*) - (\varepsilon\nabla f(x^*))^T \nabla f(x^*) + O(\varepsilon^2 \|\nabla^2 f(x^*)\|_2^2)$$

$$= f(x^*) - \varepsilon\|\nabla f(x^*)\|_2^2 + O(\varepsilon^2\|\nabla f(x^*)\|_2^2) < f(x^*)$$

Means x^* is not local optimum

$$f(y) \geq f(x) + (y-x)^T \nabla f(x)$$

if $\nabla f(x^*) = 0$.

$f(y) \geq f(x^*)$ for all $y \longrightarrow x^*$ is global minimum

Homework

1 Convexity of functions

Problem 4: Assume that $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ 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.

2 Optimization / Gradient descent

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

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

a) Compute the minimizer $\boldsymbol{x}^*$ of f analytically.

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

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

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.

suppose $f(x) = x^2 \quad g(x) = -x$

$h(x) = -x^2 \quad h''(x) = -2 < 0$

non-convex.

$x_\lambda = \lambda x_1 + (1-\lambda) x_0$

$f(x_\lambda) \le \lambda f(x_1) + (1-\lambda) f(x_0) \Leftrightarrow f$ convex.

$g(f(x_\lambda)) \le g(\lambda f(x_1)) + g((1-\lambda) f(x_0)) \Leftrightarrow g$ non-decreasing

$g(\lambda f(x_1) + (1-\lambda) f(x_0)) \le \lambda g(f(x_1)) + (1-\lambda) g(f(x_0)) \Leftrightarrow g$ convex

$\Rightarrow$ ① ② $\quad g(f(x_\lambda)) \le \lambda g(f(x_1)) + (1-\lambda) g(f(x_0))$

$\Leftrightarrow \quad h(x_\lambda) \le \lambda h(x_1) + (1-\lambda) h(x_2)$

$\nabla f(x) = \begin{pmatrix} x_1 + 2 \\ 2x_2 + 1 \end{pmatrix} = 0$

$x_1 = -2 \quad x_2 = -\frac{1}{2}$

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

$= \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 + 2 \\ 0 + 1 \end{pmatrix}$

$= \begin{pmatrix} -2 \\ -1 \end{pmatrix}$

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

$= \begin{pmatrix} -2 \\ -1 \end{pmatrix} - \tau \begin{pmatrix} 0 \\ -1 \end{pmatrix}$

$= \begin{pmatrix} -2 \\ 0 \end{pmatrix}$

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

no coverge to truth x^*

P4

$$\frac{\partial}{\partial x} h(x) = \frac{\partial g}{\partial f} \frac{\partial f}{\partial x} = g'(f(x)) \cdot f'(x)$$

$$\frac{\partial}{\partial x} \cdots = \frac{\partial g'(f(x))}{\partial x} \cdot f'(x) + g'(f(x)) \cdot \underbrace{\frac{\partial f'(x)}{\partial x}}_{f''(x)}$$

$$= \frac{\partial g'}{\partial f} \cdot \frac{\partial f}{\partial x} \cdot f'(x) + \cdots$$

$$= \underbrace{g''(f(x))}_{\geq 0} \cdot \underbrace{f'(x)^2}_{\geq 0} + \underbrace{g'(f(x))}_{?} \cdot \underbrace{f''(x)}_{\geq 0}$$

P7 $x, y \in S \rightarrow \theta x + (1-\theta) y \in S$, $\theta \in [0,1]$

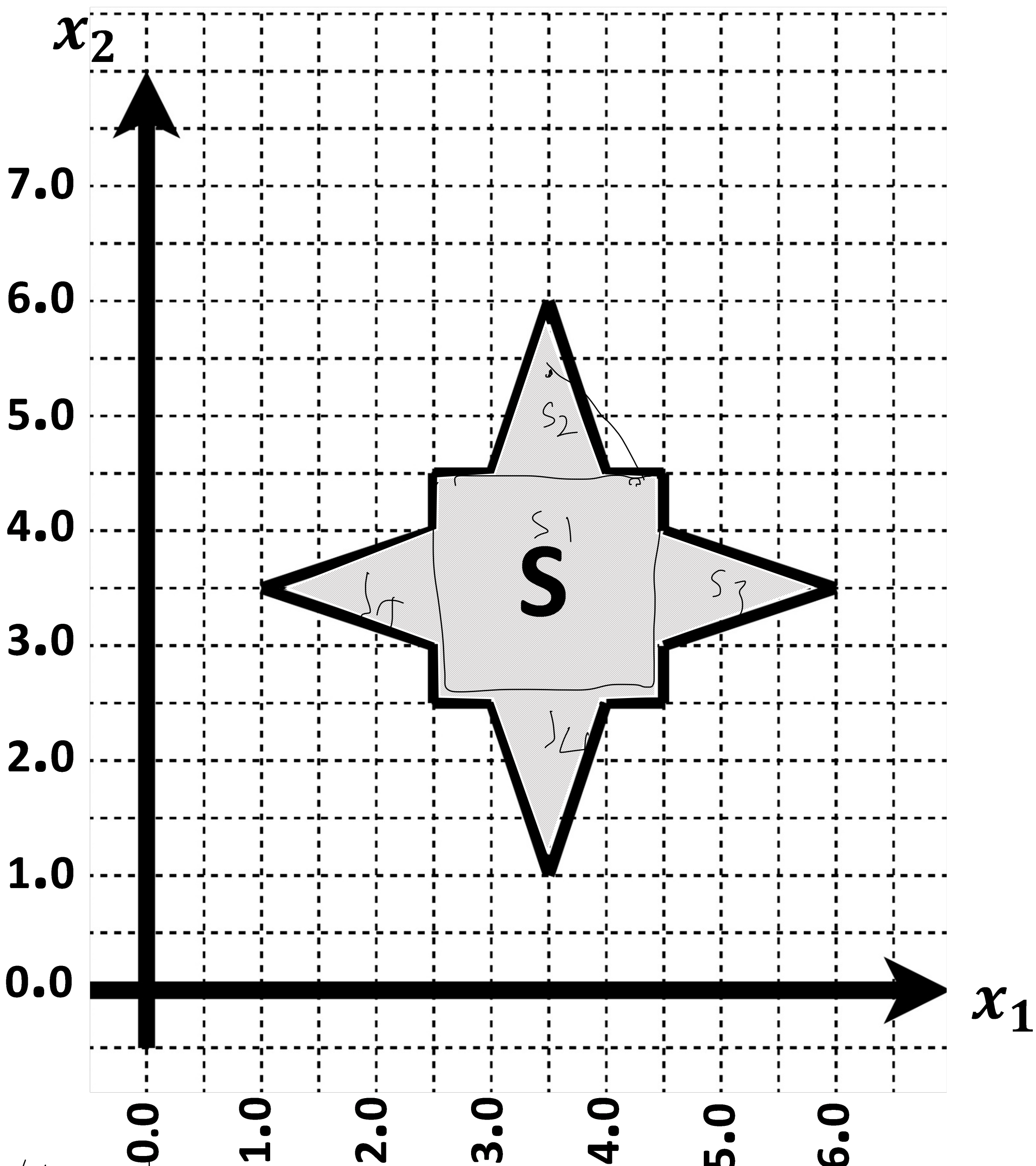
Problem 7: Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be the following convex function:

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

a) Consider the following shaded region $S \subset \mathbb{R}^2$. Is this region convex? Why?

no the point on the line, which connect two points is not in the shape

b) Assume that we are given an algorithm $\texttt{ConvOpt}(f, D)$ that takes as input a convex function f and convex region D, and returns the minimum of f over D. Using the ConvOpt algorithm, how would you find the global minimum of f over the shaded region S?



partition the region

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

$\min_{x \in S_i} f(x) = \min(m_1, \ldots m_5)$