Machine Learning Exercise Sheet 06

Optimization

Homework

1 Convexity of functions

Problem 1: Given n convex functions $g_i : \mathbb{R}^{d_i} \to \mathbb{R}$ for $i \in \{1, \dots, n\}$, prove or disprove that the function

a) $h(\boldsymbol{x}) = g_2(g_1(\boldsymbol{x}))$ is convex (here $d_1 \in \mathbb{N}$, $d_2 = 1$),

b) $h(\boldsymbol{x}) = g_2(g_1(\boldsymbol{x}))$ is convex if g_2 is non-decreasing (here $d_1 \in \mathbb{N}$, $d_2 = 1$),

c) $h(\boldsymbol{x}) = \max(g_1(\boldsymbol{x}), \dots, g_n(\boldsymbol{x}))$ is convex (here all $d_i \in \mathbb{N}$).

Notation: for $x_0, x_1 \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$ we introduce $x_\lambda = \lambda x_1 + (1-\lambda)x_0$.

a) Statement is false: consider $g_1(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2$ and $g_2(x) = -x$. Both functions are convex, but for $h(\boldsymbol{x}) = -\|\boldsymbol{x}\|_2^2$ it holds with $\boldsymbol{x}_0 = \boldsymbol{0}$, $\boldsymbol{x}_1 \neq \boldsymbol{0}$ and $\lambda \in (0,1)$ that

$$h(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_0) = -\lambda^2 \|\boldsymbol{x}_1\|_2^2 \text{ and } \lambda h(\boldsymbol{x}_1) + (1-\lambda)h(\boldsymbol{x}_0) = -\lambda \|\boldsymbol{x}_1\|_2^2.$$

Since $\lambda \in (0,1)$ we get $\lambda^2 < \lambda \Rightarrow -\lambda^2\|\boldsymbol{x}_1\|_2^2 > -\lambda\|\boldsymbol{x}_1\|_2^2 \Rightarrow h(\boldsymbol{x}_\lambda) > \lambda h(\boldsymbol{x}_1) + (1-\lambda)h(\boldsymbol{x}_0)$ contradicting the definition of convexity.

b) Statement is true: consider $\boldsymbol{x}_0, \boldsymbol{x}_1 \in \mathbb{R}^d$ and $\lambda \in (0,1)$. For h we prove convexity using definitions of the given properties of g_1 and g_2:

$$\begin{aligned}
&g_1 \text{ convex} && \Rightarrow && g_1(\boldsymbol{x}_\lambda) \leq \lambda g_1(\boldsymbol{x}_1) + (1-\lambda) g_1(\boldsymbol{x}_0) && \\
&g_2 \text{ non-decresing} && \Rightarrow && g_2(g_1(\boldsymbol{x}_\lambda)) \leq g_2(\lambda g_1(\boldsymbol{x}_1) + (1-\lambda) g_1(\boldsymbol{x}_0)) && (1)\\
&g_2 \text{ convex} && \Rightarrow && g_2(\lambda g_1(\boldsymbol{x}_1) + (1-\lambda) g_1(\boldsymbol{x}_0)) \leq \lambda g_2(g_1(\boldsymbol{x}_1)) + (1-\lambda) g_2(g_1(\boldsymbol{x}_0)) && (2)\\
&(1) \text{ and } (2) && \Rightarrow && h(\boldsymbol{x}_\lambda) \leq \lambda h(\boldsymbol{x}_1) + (1-\lambda) h(\boldsymbol{x}_0) &&
\end{aligned}$$

c) Statement is true: first, note that for the function $\max : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$ and $\lambda \geq 0$ we have

$$\max(\lambda \boldsymbol{y}_1, \dots, \lambda \boldsymbol{y}_n) = \lambda \max(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n) \quad (3)$$

$$\max(\boldsymbol{y}_1 + \boldsymbol{z}_1, \dots, \boldsymbol{y}_n + \boldsymbol{z}_n) = \boldsymbol{y}_{\bar{i}} + \boldsymbol{z}_{\bar{i}} \leq \max(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n) + \max(\boldsymbol{z}_1, \dots, \boldsymbol{z}_n) \quad (4)$$

$$\text{if } \boldsymbol{y}_i \geq \boldsymbol{z}_i \text{ for all } i = 1, \dots, N \text{ then } \max(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n) \geq \max(\boldsymbol{z}_1, \dots, \boldsymbol{z}_n) \quad (5)$$

Upload a single PDF file with your homework solution to Moodle by 24.11.2019, 23:59 CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

Now it follows ($\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_\lambda$ and λ as in b) arbitrary) that

$$
\begin{aligned}
&g_i \text{ convex} && \Rightarrow && g_i(\boldsymbol{x}_\lambda) \leq \lambda g_i(\boldsymbol{x}_1) + (1-\lambda) g_i(\boldsymbol{x}_0) \\
&(5) && \Rightarrow && \overbrace{\max(g_1(\boldsymbol{x}_\lambda), \ldots, g_n(\boldsymbol{x}_\lambda))}^{h(\boldsymbol{x}_\lambda)} \leq \qquad (6) \\
& && && \max\left(\lambda g_1(\boldsymbol{x}_1) + (1-\lambda) g_1(\boldsymbol{x}_0), \ldots, \lambda g_n(\boldsymbol{x}_1) + (1-\lambda) g_n(\boldsymbol{x}_0)\right) \\
&(3)+(4) && \Rightarrow && \max\left(\lambda g_1(\boldsymbol{x}_1) + (1-\lambda) g_1(\boldsymbol{x}_0), \ldots, \lambda g_n(\boldsymbol{x}_1) + (1-\lambda) g_n(\boldsymbol{x}_0)\right) \leq \qquad (7) \\
& && && \lambda \underbrace{\max(g_1(\boldsymbol{x}_1), \ldots, g_n(\boldsymbol{x}_1))}_{h(\boldsymbol{x}_1)} + (1-\lambda) \underbrace{\max(g_1(\boldsymbol{x}_0), \ldots, g_n(\boldsymbol{x}_0))}_{h(\boldsymbol{x}_0)} \\
&(6)+(7) && \Rightarrow && h(\boldsymbol{x}_\lambda) \leq \lambda h(\boldsymbol{x}_1) + (1-\lambda) h(\boldsymbol{x}_0)
\end{aligned}
$$

2 Optimization / Gradient descent

Problem 2: 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.

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 $\boldsymbol{x}^{(0)} = (0, 0)$ with a constant learning rate $\tau = 1$.

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

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

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?

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 $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ forever. We can fix this by decreasing the learning rate (adaptive stepsize, etc.).

Problem 3: Load the notebook `06_homework_optimization.ipynb` from Piazza. Fill in the missing code and run the notebook. Convert the evaluated notebook to pdf and add it to the printout of your homework.

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 and how to convert them to other formats, consult the Jupyter documentation and nbconvert documentation.

The solution notebook is uploaded on Piazza.

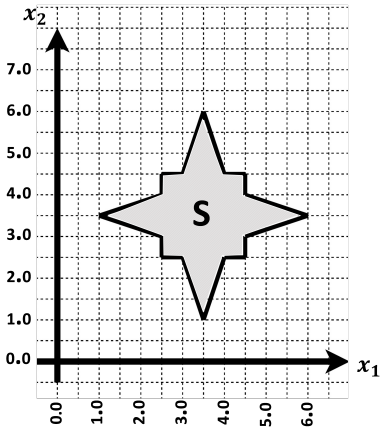
Problem 4: Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the following convex function:

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

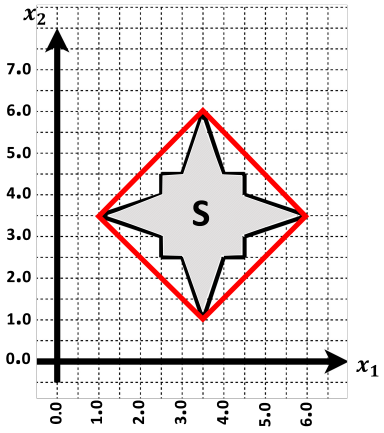
a) Consider the following shaded region S in $\mathbb{R}^2$. Is this region convex? Why?

b) Find the maximizer $\boldsymbol{x}^*$ of f over the shaded region S. Justify your answer.

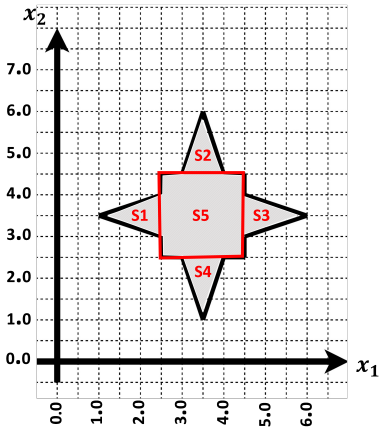
c) Assume that we are given an algorithm `ConvOpt`(f, D) that takes as input a convex function f and convex region D, and returns the minimum of f over D. Using the `ConvOpt` algorithm, how would you find the global minimum of f over the shaded region S?



(a) initial non-convex set



(b) convex hull



(c) union of convex sets

Upload a single PDF file with your homework solution to Moodle by 24.11.2019, 23:59 CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

a) It is not because we can choose two points in $\boldsymbol{S}$ such that the line connecting the points does not completely resides in $\boldsymbol{S}$, for example $(1.0, 3.5)^T$ and $(3.5, 6.0)^T$ (see Figure 1b).

b) According to the lecture, maximizer of the convex function over a non-convex set can be computed by solving the task over its convex hull (see Figure 1b). Since convex hull is convex we need to compute the value of f only on the four corners $(1.0, 3.5)^T, (3.5, 1.0)^T, (6.0, 3.5)^T$ and $(3.5, 6.0)^T$. Afterwards, we pick the corner at which the value of f is the largest.

$$f(1.0, 3.5) = e^{4.5} - 5 \cdot \log(3.5) = 83.7533, \qquad f(3.5, 1.0) = e^{4.5} - 5 \cdot \log(1.0) = 90.01$$
$$f(6.0, 3.5) = e^{9.5} - 5 \cdot \log(3.5) = 13353.46, \qquad f(3.5, 6.0) = e^{9.5} - 5 \cdot \log(6.0) = 13350.76\,.$$

Therefore, the maximizer of f over $\boldsymbol{S}$ is $(6.0, 3.5)$.

c) We can partition the shaded region $\boldsymbol{S}$ to the following five convex regions $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_5$ (see Figure 1c). Afterwards, we run the `ConvOpt` algorithm separately for the 5 regions and obtain

$$m_i = \min_{\boldsymbol{x} \in \boldsymbol{S}_i} f(\boldsymbol{x}) = \texttt{ConvOpt}(f, \boldsymbol{S}_i)\,.$$

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

In-class Exercises

Problem 5: 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 $f(x, y) = y \cdot x^3 - y \cdot x^2 + y^2 + y + 4$.

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 $\mathbb{R}^+$:

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

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

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

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 $\lambda = 0.5$:

$$0.5 f(-4, 1) + 0.5 f(0, 1) = 0.5 \cdot (-74) + 0.5 \cdot 6 = -34$$
$$f\big(0.5 \cdot (-4) + 0.5 \cdot 0, 0.5 \cdot 1 + 0.5 \cdot 1\big) = f(-2, 1) = -6 \; > -34$$

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

Problem 6: Prove that the following functions $f : \mathbb{R}^d \to \mathbb{R}$ are convex:

a) $D \subset \mathbb{R}^d$ bounded and closed set and f is defined by $f(\boldsymbol{x}) = \max_{\boldsymbol{w} \in D} \boldsymbol{x}^T \boldsymbol{w}$.

b) f is the objective function of logistic regression, that is

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

Upload a single PDF file with your homework solution to Moodle by 24.11.2019, 23:59 CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

a) We will need some facts similar to (3) and (4) that we used in Problem 1c. Let g_1 and g_2 be two functions that have finite maxima on D (e.g. later $g_1(\boldsymbol{w}) = \lambda \boldsymbol{x}_1^T \boldsymbol{w}$ and $g_2(\boldsymbol{w}) = (1-\lambda)\boldsymbol{x}_0^T \boldsymbol{w}$), then

$$\max_{\boldsymbol{w}\in D} g_1(\boldsymbol{w}) + g_2(\boldsymbol{w}) = g_1(\bar{\boldsymbol{w}}) + g_2(\bar{\boldsymbol{w}}) \leq \max_{\boldsymbol{w}\in D} g_1(\boldsymbol{w}) + \max_{\boldsymbol{w}\in D} g_2(\boldsymbol{w})\,.$$

With $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_\lambda$ and λ as before we get

$$\begin{aligned} f(\boldsymbol{x}_\lambda) = \max_{\boldsymbol{w}\in D} \boldsymbol{x}_\lambda^T \boldsymbol{w} = \max_{\boldsymbol{w}\in D} \overbrace{\lambda \boldsymbol{x}_1^T \boldsymbol{w}}^{=g_1(\boldsymbol{w})} + \overbrace{(1-\lambda)\boldsymbol{x}_0^T \boldsymbol{w}}^{=g_2(\boldsymbol{w})} &\leq \max_{\boldsymbol{w}\in D} \lambda \boldsymbol{x}_1^T \boldsymbol{w} + \max_{\boldsymbol{w}\in D} (1-\lambda)\boldsymbol{x}_0^T \boldsymbol{w} \\ &= \lambda \max_{\boldsymbol{w}\in D} \boldsymbol{x}_1^T \boldsymbol{w} + (1-\lambda) \max_{\boldsymbol{w}\in D} \boldsymbol{x}_0^T \boldsymbol{w} \\ &= \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_0)\,. \end{aligned}$$

b) First, notice that if we can prove that the following two functions

$$g_1(\boldsymbol{w}) = -\ln \sigma(\boldsymbol{w}^T \boldsymbol{x}_i) \qquad \text{and} \qquad g_2(\boldsymbol{w}) = -\ln(1 - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i))$$

are convex for all $\boldsymbol{x}_i \in \mathbb{R}^d$, the whole function f must also be convex since any linear combination (with non-negative weights, which is true since y_i and $1 - y_i$ are non-negative) of two or more convex functions is also convex.

To prove that the first function is convex we will use the second-order condition of convexity. A function $f(x)$ which is twice-differentiable is convex if and only if its Hessian matrix (matrix of second-order partial derivatives) is positive semi-definite (see Problem 8). To compute the Hessian matrix we first calculate the derivative of the sigmoid function:

$$\begin{aligned} \frac{d}{dx}\sigma(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x)\frac{e^{-x}}{1+e^{-x}} = \sigma(x)\left(1 + \frac{e^{-x} - 1 - e^{-x}}{1+e^{-x}}\right) &= \sigma(x)\left(1 - \frac{1}{1+e^{-x}}\right) \\ = \sigma(x)(1-\sigma(x))\,. \end{aligned}$$

Using this, we can derive the Hessian:

$$\begin{aligned} \boldsymbol{\nabla}_{\boldsymbol{w}}^2 \left[-\ln \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)\right] &= \boldsymbol{\nabla}_{\boldsymbol{w}} \left[\boldsymbol{\nabla}_{\boldsymbol{w}} (-\ln \sigma(\boldsymbol{w}^T \boldsymbol{x}_i))\right] \\ &= \boldsymbol{\nabla}_{\boldsymbol{w}} \left[-\boldsymbol{x}_i \frac{\sigma(\boldsymbol{w}^T \boldsymbol{x}_i)(1 - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i))}{\sigma(\boldsymbol{w}^T \boldsymbol{x}_i)}\right] \\ &= \boldsymbol{\nabla}_{\boldsymbol{w}} \left[\boldsymbol{x}_i (\sigma(\boldsymbol{w}^T \boldsymbol{x}_i) - 1)\right] \\ &= \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)(1 - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)) \boldsymbol{x}_i \boldsymbol{x}_i^T\,. \end{aligned}$$

Next, we show that this Hessian matrix is positive semi-definite:

$$\begin{aligned} \forall \boldsymbol{z}: \quad & \boldsymbol{z}^T \boldsymbol{\nabla}_{\boldsymbol{w}}^2 \left[-\ln \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)\right] \boldsymbol{z} \\ &= \boldsymbol{z}^T \left[\sigma(\boldsymbol{w}^T \boldsymbol{x}_i)(1 - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)) \boldsymbol{x}_i \boldsymbol{x}_i^T\right] \boldsymbol{z} \\ &= \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)(1 - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i))(\boldsymbol{x}_i^T \boldsymbol{z})^2 \geq 0 \end{aligned}$$

To prove that the second function is convex, we first notice:

$$-\ln(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)) = -\ln\left(1-\frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}_i}}\right) = -\ln\left(\frac{e^{-\boldsymbol{w}^T\boldsymbol{x}_i}}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}_i}}\right)$$
$$= \boldsymbol{w}^T\boldsymbol{x}_i - \ln\left(\frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}_i}}\right) = \boldsymbol{w}^T\boldsymbol{x}_i - \ln\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)$$

This is a sum of two convex functions, since the affine function $\boldsymbol{w}^T\boldsymbol{x}_i$ is convex and we just showed that $-\ln\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)$ is convex. Hence, $-\ln(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))$ is convex as well.

Problem 7: 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 if there is a local minimum at $\boldsymbol{x}^*$ then $\boldsymbol{\nabla} f(\boldsymbol{x}^*) = 0$ and if $\boldsymbol{\nabla} f(\boldsymbol{x}^*) = 0$ then $\boldsymbol{x}^*$ is a global minimum.

At a (local) minimum point $\boldsymbol{\theta}^*$ the gradient must be zero. Otherwise we could follow the gradient to get an even lower value.

$$f(\boldsymbol{\theta}^* - \epsilon\boldsymbol{\nabla} f(\boldsymbol{\theta}^*)) = f(\boldsymbol{\theta}^*) - \epsilon||\boldsymbol{\nabla} f(\boldsymbol{\theta}^*))||_2^2 + O(\epsilon^2||\boldsymbol{\nabla} f(\boldsymbol{\theta}^*)||_2^2) < f(\boldsymbol{\theta}^*)$$

for sufficiently small ϵ.

Now, using the first order criterion we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + (\boldsymbol{y}-\boldsymbol{x})\boldsymbol{\nabla} f(\boldsymbol{x})$$

If we replace $\boldsymbol{x}$ with $\boldsymbol{\theta}^*$ and plug in $\boldsymbol{\nabla} f(\boldsymbol{\theta}^*) = 0$ we get: $f(\boldsymbol{y}) \geq f(\boldsymbol{\theta}^*)$ for all $\boldsymbol{y}$, meaning $\boldsymbol{\theta}^*$ is a global minimum.

Problem 8: Show that a twice differentiable function $f : D \to \mathbb{R}$ with a convex domain $D \subset \mathbb{R}^d$ is convex if and only if its Hessian is positive semi-definite on the whole domain D.

For all $\boldsymbol{x}_0, \boldsymbol{x}_1 \in D$ and some $\boldsymbol{x}_\lambda$ as before for a $\lambda \in [0,1]$ (depending on $\boldsymbol{x}_0, \boldsymbol{x}_1$) it holds that

$$f(\boldsymbol{x}_1) = f(\boldsymbol{x}_0) + \boldsymbol{\nabla} f(\boldsymbol{x}_0)^T(\boldsymbol{x}_1-\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}_1-\boldsymbol{x}_0)^T\boldsymbol{\nabla}^2 f(\boldsymbol{x}_\lambda)(\boldsymbol{x}_1-\boldsymbol{x}_0). \tag{8}$$

If $\boldsymbol{\nabla}^2 f(\boldsymbol{x})$ is positive semi-definite for all $\boldsymbol{x} \in D$ we get that the quadratic term in (8) is non-negative and therefore $f(\boldsymbol{x}_1) \geq f(\boldsymbol{x}_0) + \boldsymbol{\nabla} f(\boldsymbol{x}_0)^T(\boldsymbol{x}_1-\boldsymbol{x}_0)$ which is the first-order convexity condition.

Now assume $\boldsymbol{\nabla}^2 f(\boldsymbol{x})$ is not positive semi-definite for some $\boldsymbol{x} \in D$, so there exists d such that $d^T\boldsymbol{\nabla}^2 f(\boldsymbol{x})d < 0$. Consider (8) for $\boldsymbol{x}_0 = \boldsymbol{x}$ and $\boldsymbol{x}_1 = \boldsymbol{x} + \epsilon\boldsymbol{d}$ for $\epsilon > 0$. For sufficiently small ϵ we get $\boldsymbol{d}^T\boldsymbol{\nabla}^2 f(\boldsymbol{x}_\lambda)\boldsymbol{d} < 0$ since $\boldsymbol{\nabla}^2 f$ is continuous. Therefore, the quadratic term in (8) for our choice of $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ is now negative and we get $f(\boldsymbol{x}_1) < f(\boldsymbol{x}_0) + \boldsymbol{\nabla} f(\boldsymbol{x}_0)^T(\boldsymbol{x}_1-\boldsymbol{x}_0)$ contradicting the first-order convexity condition.