**Question 1 [50 marks].** Let $\sigma(w) = \frac{1}{1+e^{-w}}$ be the logistic function. Let $\mathbf{x} = \{x_i\}_{i=1}^s$ be some data samples, where $x_i \in \mathbb{R}$ for all $i = 1, \ldots, s$ and

$$E(\mathbf{w} = (w_1, w_2, \ldots, w_s)) = -\log \sigma\left(\mathbf{x}^\top \mathbf{w}\right) = \log\left(1 + e^{-\sum_{i=1}^{s} x_i w_i}\right),$$

be the logistic regression cost function.

(a) Prove $E(\mathbf{w})$ is a convex function.

*Hint:* You may wish use the fact that a twice differentiable function $\sigma(w) : \mathbb{R} \to \mathbb{R}$ is convex if and only if its second derivative satisfies

$$\sigma''(w) \geq 0, \quad \forall w \in \mathbb{R}.$$

(b) Show that $E(\mathbf{w})$ is an $L$-smooth function for $L = \frac{1}{4}\|\mathbf{x}\|^2$.

*Hint: You may wish to use the following inequality (without proving it)*

$$|\sigma(a) - \sigma(b)| \leq \frac{1}{4}|a - b|,$$

*valid for any* $a, b \in \mathbb{R}$.

We say that a random variable $\xi$ follows the Rayleigh distribution with parameter $\alpha$ if its probability density function is given by

$$p(x|\alpha) = \begin{cases} 2\alpha x e^{-\alpha x^2}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Let the parameter $\alpha$ be equal to the maximum between the fifth digit of your student ID and one.

(c) Calculate the variance $\mathrm{Var}[\xi]$ of a random variable $\xi$ drawn from the Rayleigh distribution with the parameter $\alpha$ as described above.

(d) Suppose you get $s$ i.i.d. samples $\{\xi_i\}_{i=1}^s$ drawn from the Rayleigh distribution with the parameter $\alpha$ described above. Derive the optimisation problem for the negative log-likelihood estimator for $\alpha$ of the form

$$\hat{\alpha} = \min_{\alpha > 0} E(\alpha),$$

where $E(\alpha)$ is an energy function.

(e) Solve the maximum likelihood estimator problem for the parameter $\alpha$ based on these samples.

**Question 2 [50 marks].** Suppose you are given $s = 5$ samples $\{(x_i, y_i)\}_{i=1}^5$ with

$$\begin{aligned} (x_1, y_1) &= (-2, -1), \\ (x_2, y_2) &= (-1, 1), \\ (x_3, y_3) &= (0, 3), \\ (x_4, y_4) &= (1, 5), \\ (x_5, y_5) &= (2, 7). \end{aligned}$$

In this question you are asked to build a one-feature linear regression model

$$y_i \approx f_\theta(x_i) := \theta x_i,$$

where $\theta$ is the only weight parameter.

(a) Start with the mean-squared error (MSE) as a measure of a deviation, where

$$\mathrm{MSE}(\theta) = \frac{1}{2s}\sum_{i=1}^{s}(y_i - f_\theta(x_i))^2. \tag{1}$$

Run the grid search algorithm to minimise $\mathrm{MSE}(\theta)$ over the grid $\theta \in G$, with $G = \{1, 2, 3, 4\}$. Fill in the missing entries of the following table by evaluating $y_i - \theta x_i$ for corresponding values of the parameters, and find the minimiser $\hat{\theta}$ given by

$$\hat{\theta} = \arg\min_{\theta \in G} \mathrm{MSE}(\theta).$$

| | $\theta = 1$ | $\theta = 2$ | $\theta = 3$ | $\theta = 4$ |
|---|---|---|---|---|
| $x_1 = -2,\ y_1 = -1$ | | | | |
| $x_2 = -1,\ y_2 = 1$ | | | | |
| $x_3 = 0,\ y_3 = 3$ | | | | |
| $x_4 = 1,\ y_4 = 5$ | | | | |
| $x_5 = 2,\ y_5 = 7$ | | | | |
| $\mathrm{MSE}(\theta)$ | | | | |

(b) Derive the gradient descent update formula that aims at minimising $\mathrm{MSE}(\theta)$.

Solution: The gradient descent update rule takes the form

$$\theta^{(k+1)} = \theta^{(k)} - \tau \nabla \mathrm{MSE}\left(\theta^{(k)}\right).$$

The gradient of MSE defined in (1) is equal to

$$\nabla \mathrm{MSE}(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \frac{1}{2s} \sum_{i=1}^{s} (y_i - \theta x_i)^2 = \frac{1}{s} \sum_{i=1}^{s} x_i (\theta x_i - y_i) = \theta \overline{\mathbf{x}^2} - \overline{\mathbf{xy}},$$

where $\overline{\mathbf{x}^2} = \frac{1}{s} \sum_{i=1}^{s} x_i^2$ and $\overline{\mathbf{xy}} = \frac{1}{s} \sum_{i=1}^{s} x_i y_i$. The gradient descent update formula then takes the form

$$\theta^{(k+1)} = \theta^{(k)} \left(1 - \tau \overline{\mathbf{x}^2}\right) + \tau \overline{\mathbf{xy}}.$$

(c) Let $\theta^{(0)}$ be the sixth digit of your student ID. Perform two steps of the gradient descent method with the step-size set to $\tau = \frac{1}{2}$ to evaluate $\theta^{(1)}$ and $\theta^{(2)}$. Comment on your findings.

(d) Let $a$ and $b$ be the maximum between the seventh, respectively eighth, digit of your student ID and one. The leaky rectifier function

$$R_{a,b}(x) = \begin{cases} ax, & x \geq 0, \\ -bx, & x < 0. \end{cases}$$

can also be written via the maximisation problem

$$R_{a,b}(x) = \max_{p \in [-b,a]} xp.$$

Show that the smoothed leaky rectifier function

$$R_{a,b,\tau}(x) := \max_{p \in [-b,a]} xp - \frac{\tau}{2}|p|^2$$

for a parameter $\tau > 0$ has the closed-form solution

$$R_{a,b,\tau}(x) = \begin{cases} ax - \frac{\tau}{2}a^2, & x > a\tau, \\ \frac{x^2}{2\tau}, & -b\tau < x \leq a\tau, \\ -bx - \frac{\tau}{2}b^2, & x \leq -b\tau. \end{cases}$$

(e) Consider a regularised optimisation problem of the form

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}} \left[\mathrm{MSE}(\theta) + R_{a,b,\tau}(\theta)\right], \tag{2}$$

where the MSE and the parametric rectifier $R_{a,b,\tau}$ functions are defined above. Derive the gradient descent update formula that aims at solving (2). Let $\theta^{(0)}$ be the seventh digit of your student ID. Perform two steps of the gradient descent starting from $\theta^{(0)}$ with step-size set to $\tau = \frac{1}{2}$ to evaluate $\theta^{(1)}$ and $\theta^{(2)}$.