Exercise 1:

One popular classification model is **logistic regression**. The medical research group from last week is eager to find out if it can be used to predict whether a patient admitted to the hospital will require intensive care. This is a binary classification task with target space $\mathcal{Y} = \{0, 1\}$, with $y = 1$ if the patient requires intensive care and $y = 0$ if not. The feature space is the same as before: $\mathcal{X} = (\mathbb{R}_0^+)^3$, with $\mathbf{x}^{(i)} = (x_{age},\ x_{blood\,pressure},\ x_{weight})^{(i)} \in \mathcal{X}$ for $i = 1, 2, \ldots, n$ observations.

Before the group trains a logistic regression model, researcher Holger remarks they could just as well fit a linear model (LM), as in the case of a binary classification task, both models would make identical predictions. Therefore, he comes up with the following hypothesis space:

$$\mathcal{H} = \left\{\pi : \mathcal{X} \to [0, 1] \mid \pi(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x}\right\} \tag{1}$$

1) Are predictions and hypothesis space of a logistic regression model and an LM identical for a binary classification task? If not, explain why they could differ and write down the correct hypothesis space.

Researcher Lisa knows that logistic regression follows a discriminant approach, meaning the discriminant functions are optimized directly via empirical risk minimization (ERM). She remembers the general form of ERM:

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f) = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right) \tag{2}$$

Additionally, she recalls the Bernoulli loss function of the logistic regression model in statistics:

$$L\left(y, \pi(\mathbf{x})\right) = -\, y \ln(\pi(\mathbf{x})) - (1 - y) \ln(1 - \pi(\mathbf{x})) \tag{3}$$

Lastly, she recollects how logistic regression models the posterior probabilities $\pi(\mathbf{x} \mid \boldsymbol{\theta})$ of the labels – the estimated linear scores are "squashed" through the logistic function s:

$$\pi(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta}^T \mathbf{x}\right)}{1 + \exp\left(\boldsymbol{\theta}^T \mathbf{x}\right)} = \frac{1}{1 + \exp\left(-\boldsymbol{\theta}^T \mathbf{x}\right)} = s\left(\boldsymbol{\theta}^T \mathbf{x}\right) \tag{4}$$

Given (2) − (4), she figures one could formulate the explicit ERM problem, but leaves the task to you.

2) Write down the explicit form of the ERM problem for estimating the parameter vector $\boldsymbol{\theta}$.

Later, the research group trains the logistic regression model and receives a corresponding parameter estimate $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0,\ \hat{\theta}_{age},\ \hat{\theta}_{blood\,pressure},\ \hat{\theta}_{weight})$. Researcher Son, who has worked all night on the research problem, finds a function scribbled on his personal notes. He remembers it was useful in the context of a logistic regression model, but does not recall how:

$$h\left(\mathbf{x}^{(i)} \mid \hat{\boldsymbol{\theta}},\ \alpha\right) = \mathbb{I}_{[\alpha, 1]}\left(\frac{1}{1 + \exp(-\hat{\boldsymbol{\theta}}^T \mathbf{x}^{(i)})}\right), \quad \alpha \in\ (0, 1) \tag{5}$$

3) What purpose does the function serve in the case of a trained logistic regression model with estimated parameters $\hat{\boldsymbol{\theta}}$? Explain the role of the parameter α.

Researcher Son is curious about why the loss function of the logistic regression model in (3) is called *Bernoulli loss*. He seems certain that he can connect it to the Bernoulli distribution, which has the following probability mass function:

$$\mathbb{P}(Y = y) = \pi^y (1 - \pi)^{1-y}, \quad y \in \{0, 1\} \tag{6}$$

4) Derive the log-likelihood function ℓ of a single Bernoulli distributed random variable Y. How is it related to the loss function used for ERM in (3)?