Problem 1. (Logistic loss properties)
In class, we developed the logistic loss for binary classification as

$$L(w) = \frac{1}{N}\sum_{i=1}^{N} y^i \log(1+e^{-z^i}) + (1-y^i)\log(1+e^{z^i}), \tag{0.1}$$

where e is the exponential function, log refers to the natural logarithm, $z^i = w_0 + w_1 x_1^i + \cdots + w_d x_d^i$, and our data pairs are $\{(x^i, y^i)\}_{i=1}^N$ with $x^i \in \mathbb{R}^d, y^i \in \{0,1\}$.

1. Compute the derivative $\nabla L(w) \in \mathbb{R}^{d+1}$ of the logistic loss function. Do you see a similarity and a difference between this derivative and the derivative of the least-squares linear regression?
2. Explain an approach to verify the convexity of $L(w)$.
3. Now, consider a constant classifier by assuming the only parameter of the logistic regression is w_0. In other words, for all data points, we consider only the constant feature 1. Show that the optimal w_0 is a function of the fraction of positive examples, namely, $w_0^\star$ is a function of $\frac{1}{N}|\{i \,|\, y^i = 1\}|$ (where for a set S, $|S|$ denotes its cardinality).

Problem 2. (Empirical distribution and expectation)
Suppose we flip a (possibly biased) coin 5 times and observe the outcomes

$$(\mathrm{H}, \mathrm{T}, \mathrm{H}, \mathrm{H}, \mathrm{T}).$$

Here "H" stands for heads and "T" for tails. Recall that the *empirical distribution* $\hat{p}$ of a random variable based on a set of samples $\{s_i\}_{i=1}^N$ with $s_i \in \{H, T\}$ for all $i = 1, \ldots, N$, is defined as

$$\hat{p}(s) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{s_i = s\}}, \quad \text{for each outcome } s \in \{H, T\}.$$

1. Write the empirical distribution $\hat{p}$ of this coin based on the samples.
2. Let $f(\mathrm{H}) = 1$ and $f(\mathrm{T}) = 0$. Compute the *empirical expectation* $\mathbb{E}_{s\sim\hat{p}}[f(s)]$ under $\hat{p}$.
3. Is it possible for the empirical distribution $\hat{p}$ to assign probability 0 to an outcome that has positive true probability?

Problem 3. (Neural network)
Consider a neural network

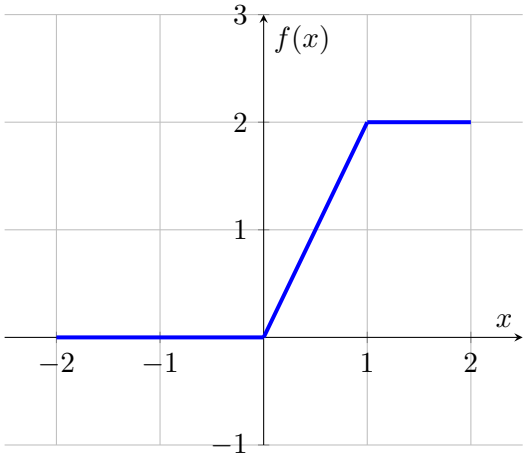
$$f : [-2, 2] \to \mathbb{R}, \quad f(x) = W^{[1]T} g\left(W^{[0]T} x + b^{[0]}\right) + b^{[1]},$$

with a single hidden layer and the ReLU activation function[1] $g(x) := \max(0, x)$. Supposing that there are two nodes in the hidden layer, determine the weight matrices $W^{[0]} \in \mathbb{R}^{1\times 2}, W^{[1]} \in \mathbb{R}^{2\times 1}$ and the biases $b^{[0]} \in \mathbb{R}^2, b^{[1]} \in \mathbb{R}$ such that:

1. $f(x) = x$.

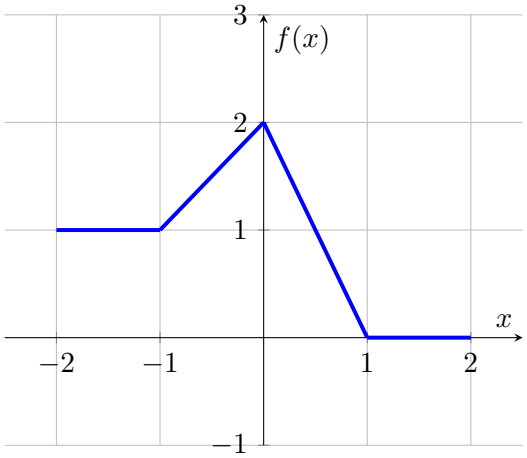
[1] Activation functions are applied elementwise to each node of a hidden layer.

2. f has the following graph:



Supposing that there are three nodes in the hidden layer, determine the weights $W^{[0]} \in \mathbb{R}^{1\times 3}$, $W^{[1]} \in \mathbb{R}^{3\times 1}$ and the biases $b^{[0]} \in \mathbb{R}^3, b^{[1]} \in \mathbb{R}$ such that:

3. f has the following graph:



Problem 4. (Neural network for regression)
Consider a data set $\{x^i, y^i\}_{i=1}^N$, with $x^i \in \mathbb{R}^d$ and $y^i \in \mathbb{R}^m$. Let our predictor be a neural network $f : \mathbb{R}^d \to \mathbb{R}^m$.

1. Consider a neural network with one hidden layer having 3 nodes and an output layer having m nodes, with activation function $g : \mathbb{R} \to \mathbb{R}$ for each node. Draw this network. How many parameters need to be determined in this network?

2. Recall the definition of an affine function from your "Background and notations.pdf" file (posted on Moodle, 8 September - 14 September). Show that if the activation function for each node in the hidden layer and each node in the output layer is the identity, $g(z) = z$, $\forall z \in \mathbb{R}$, then the neural network predictor is the same as a linear predictor with $m(d + 1)$ parameters to be determined. Hence, the problem is the same as linear regression.

3. Now, consider the activation function $g(z) = \tanh(z)$ for each node. Write the predictor f.

(a) Let $d = 2$, $m = 1$. Suppose all biases are 0 and we pick weights $\hat{\boldsymbol{W}}_{ij}^{[1]} = 1$ for all i, j, and $\hat{\boldsymbol{W}}_{ij}^{[2]} = -1$ for all i, j. For $x^{\text{test}} = [4, -2]^\top$, evaluate $f(x^{\text{test}})$.

(b) Write the mean-squared loss function $L(\boldsymbol{W}, \boldsymbol{b})$. What is the domain of this function?

(c) Let $\frac{\partial L}{\partial \boldsymbol{W}}$ and $\frac{\partial L}{\partial \boldsymbol{b}}$ denote the partial gradients of the loss with respect to weights and biases, respectively. Write the gradient descent rule for minimizing the loss function L.[2] Would you expect this procedure to find the optimal $\boldsymbol{W}$ and $\boldsymbol{b}$?

[2] You can use $\frac{\partial L}{\partial \boldsymbol{W}}$ and $\frac{\partial L}{\partial \boldsymbol{b}}$ in your expressions without determining them explicitly.