Midterm Exam, Fall 2024
ESE 577

Instructor: Jorge Mendez-Mendez

Date: Thursday October 10th, 2024

Do not tear exam booklet apart!

- This is a closed book exam. One sheet (8 1/2 in. by 11 in.) of notes, front and back, are permitted. Calculators are not permitted.
- The total exam time is 2.5 hours.
- The problems are not necessarily in any order of difficulty.
- Record all your answers in the places provided. If you run out of room for an answer, continue on a blank page and mark it clearly.
- If a question seems vague or under-specified to you, make an assumption, write it down, and solve the problem given your assumption.
- If you absolutely *have* to ask a question, come to the front.
- **Write your name on every piece of paper.**

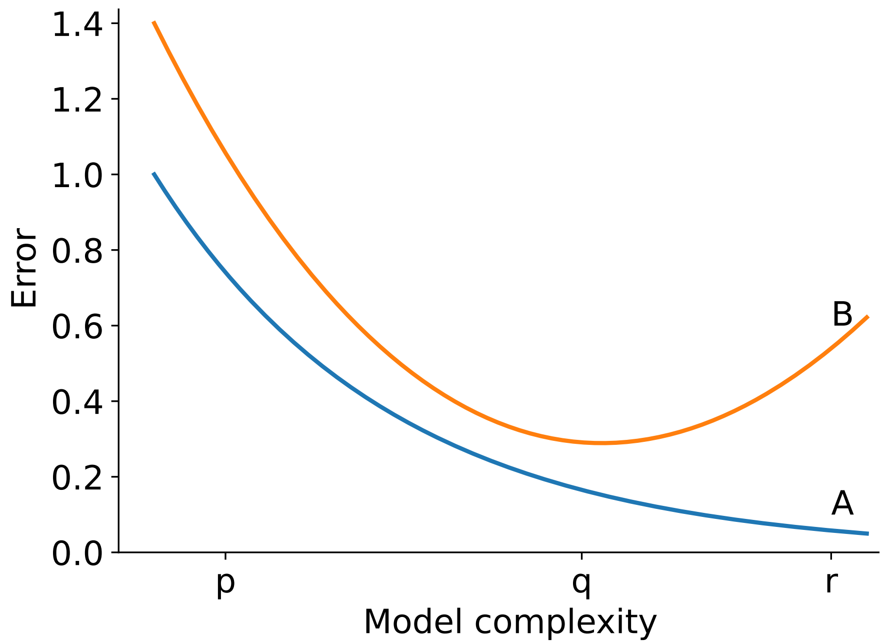
Name: ______________________ SBU email: ______________________

Question	Points
1	12
2	19
3	22
4	20
5	9
6	18
Total:	100

Name: ____________________________

Model Evaluation

1. (12 points) The following plot shows the training and validation errors as a function of the model complexity. For each of the following questions about the plots, provide a one-sentence justification for your answer.



(a) Which of the curves likely corresponds to the training and validation error?

◯ Curve A is the training error and curve B is the validation error

◯ Curve B is the validation error and curve A is the training error

◯ Both curves are the training error

◯ Both curves are the validation error

Justification:

(b) What type of fit occurs at the point labeled "p" on the horizontal axis?

◯ Underfitting.

◯ Proper fit.

◯ Overfitting

Justification:

(c) What type of fit occurs at the point labeled "r"?

◯ Underfitting.

◯ Proper fit.

◯ Overfitting

Justification:

(d) What type of fit occurs at the point labeled "q"?

◯ Underfitting.

◯ Proper fit.

◯ Overfitting

Justification:

(e) Suppose that the curves above correspond to a linear hypothesis trained to minimize the ridge regression objective, given by:

$$J(X, y; \theta) = \frac{1}{n}\left(\sum_{i=1}^{n}\left(\theta^\top x^{(i)} - y^{(i)}\right)\right) + \lambda\|\theta\|^2 \ .$$

Which of the following could correspond to the value being plotted in the horizontal axis?

◯ θ

◯ λ

◯ $\frac{1}{\lambda}$

◯ $y^{(i)}$

Justification:

Name: ______________________

Regression or Repetition

2. (19 points) Suppose that we are given a small dataset and we would like to learn the parameters of a linear regressor hypothesis taking the form $h(x) = \theta^\top x + \theta_0$ for fitting the data.

   (a) Consider the following dataset $\mathcal{D}_1$ containing three data points (in feature-label pairs):

x	y
-4	-3
2	3
-1	6

   Suppose that we would like to minimize:

$$J_1(\theta, \theta_0; \mathcal{D}_1) = \frac{1}{n}\sum_{i=1}^{n}(\theta^\top x^{(i)} + \theta_0 - y^{(i)})^2 \ .$$

   i. For J_1, we know that there exist θ^* and θ_0^* that minimize it. Can we find θ^* via the analytical solution formula? (No need for justification)

      ◯ Yes.

      ◯ No.

   ii. Suppose we know that for J_1, one set of minimizing parameters has $\theta_0^* = 3$. What is the corresponding unknown θ^*?

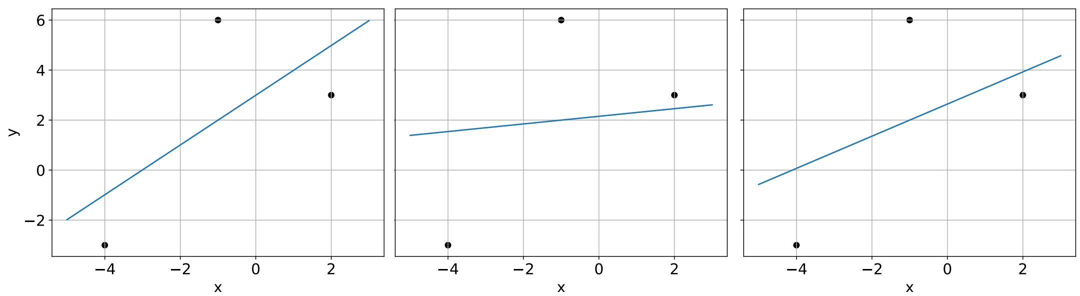
   iii. What is J_1^*, the minimum value achievable of J_1?

(b) Suppose instead of J_1, we try to minimize:

$$J_2(\theta, \theta_0; \mathcal{D}_1, \lambda) = \frac{1}{n}\sum_{i=1}^{n}(\theta^\top x^{(i)} + \theta_0 - y^{(i)})^2 + \lambda\|\theta\|^2 \ ,$$

with $\lambda = 0.1$, 10, and 100. Identify the λ used to generate plot (III).

Which value of λ corresponds to which plot?



(I) (II) (III)

$\lambda = 0.1$ corresponds to plot: ________

$\lambda = 10$ corresponds to plot: ________

$\lambda = 100$ corresponds to plot: ________

(c) Suppose we add a second feature for each of the three datapoints in $\mathcal{D}_1$, to obtain the new dataset $\mathcal{D}_2$:

x_1	x_2	y
-4	8	-3
2	-4	3
-1	2	6

Suppose that we would like to minimize:

$$J_3(\theta, \theta_0; \mathcal{D}_2) = \frac{1}{n}\sum_{i=1}^{n}(\theta^\top x^{(i)} + \theta_0 - y^{(i)})^2 \ .$$

i. For J_3, we also know that there exist θ^* and θ_0^* that minimize it. Can we find θ^* via the analytical solution formula? If yes, provide such θ^*; if not, briefly justify why not.

◯ Yes.

◯ No.

Justification or $\theta^* =$

Name: ______________________________

ii. Compare J_3^*, the minimum value achievable of J_3, with J_1^*. Which option below is true? Briefly justify your choice.

- ◯ $J_3^* > J_1^*$
- ◯ $J_3^* = J_1^*$
- ◯ $J_3^* < J_1^*$
- ◯ It depends

Justification:

Name: ______________________

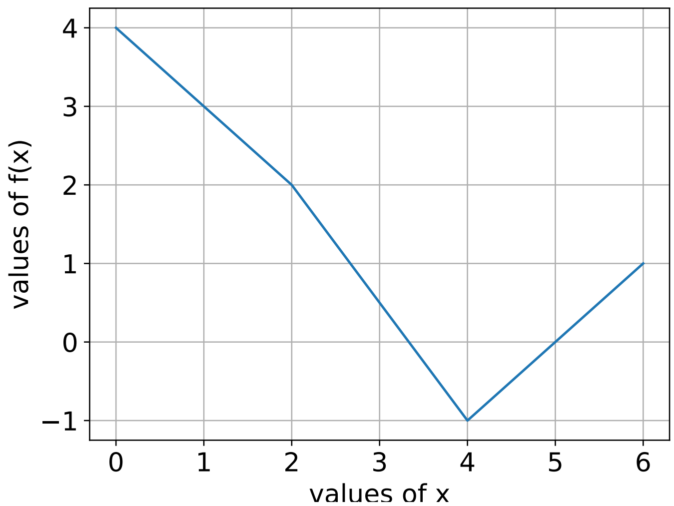
Gradient Descent in Pictures

3. (22 points) John is using standard gradient descent iterations:

$$x^{(k+1)} = x^{(k)} - \eta \nabla f(x^{(k)}) \qquad (k = 0, 1, 2, \ldots)$$

on a variety of functions f.

(a) First, John applies gradient descent to a piecewise-linear function $f : \mathbb{R} \mapsto \mathbb{R}$ with the (partial) graph shown in the figure below:



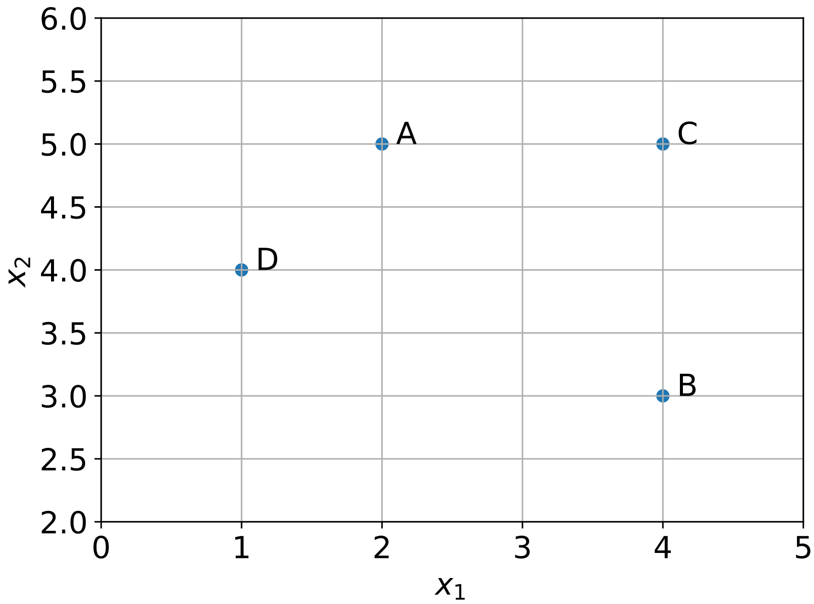
At points $x = 2$ and $x = 4$, John uses $\nabla f(2) = -1.5$ and $\nabla f(4) = 0$.

i. Starting from the initial guess $x^{(0)} = 0.5$, and using a step size $\eta = 2$, what will be the values of $x^{(1)}$, $x^{(2)}$, and $x^{(3)}$?

ii. John discovers that, starting with $x^{(0)} = 5$, there are many values of $\eta > 0$ for which the gradient descent iterations produce oscillations of period 2 within the range $(0, 6)$ (i.e., $x^{(k+2)} = x^{(k)} \in (0, 6)$ for all $k = 0, 1, 2, \ldots$). Find all such values of η.

Name: ______________________________

(b) After mastering one-dimensional optimization, John applies gradient descent to a smooth convex function $f : \mathbb{R}^2 \mapsto \mathbb{R}$, with $\eta = 0.1$, resulting in the sequence of points $x^{(0)} = A$, $x^{(1)} = B$, $x^{(2)} = C$, $x^{(3)} = x^{(4)} = D$ shown on the plot below:



i. Find $\nabla f(x^{(0)})$, $\nabla f(x^{(1)})$, $\nabla f(x^{(2)})$, and $\nabla f(x^{(3)})$.

ii. Is this statement true or false: "given the information provided, the point D **must** be a global minimum of function f"? Briefly justify your choice.

○ True

○ False

Justification:

Name: ______________________________

CTO Logistic Regression

4. (20 points) You have a training dataset $\mathcal{D}$ with each input $x^{(i)}$ consisiting of d binary features $x_1^{(i)}, \ldots, x_d^{(i)}$, where all $x_j^{(i)} \in \{0, 1\}$ and a binary label $y^{(i)} \in \{0, 1\}$. After hours struggling with *underfitting* (high training-set error) when using logistic regression to make predictions, you decide to call your friend, the CTO of the famous start-up *ClosedAI*, who gives you a few ideas to get a lower training error. For each of them, specify whether 1) it generally reduces underfitting; 2) it will not make a difference; or 3) it generally worsens underfitting. **Note that we are only talking about training loss, not test loss. Justify each answer.**

(a) "Combine and conquer": For each data point, augment the dimensions by adding the *and* of each feature with its neighbor; that is $[x_1^{(i)}, \ldots, x_d^{(i)}] \mapsto [x_1^{(i)}, \ldots x_d^{(i)}, x_1^{(i)} \cdot x_2^{(i)}, \ldots, x_{d-1}^{(i)} \cdot x_d^{(i)}]$.

◯ Reduces underfitting

◯ No change

◯ Worsens underfitting

Justification:

(b) "Two is bigger than one": Use two sigmoids instead of one; that is, instead of parameterizing the solution as $g^{(i)} = \sigma(\theta^\top x^{(i)} + \theta_0)$, parameterize it as $g^{(i)} = \sigma(\sigma(\theta^\top x^{(i)} + \theta_0))$.

◯ Reduces underfitting

◯ No change

◯ Worsens underfitting

Justification:

(c) "Ensembling is always good": Fit a regular logistic regression using the standard sigmoid function with base e and a second logistic regression using a base-2 σ_2 sigmoid function, and return the most confident result (farthest away from 0.5).

◯ Reduces underfitting

◯ No change

◯ Worsens underfitting

Justification:

(d) "Label distillation": Before fitting the dataset, for every data-point i, add its label as a feature; that is, $[x_1^{(i)}, \ldots, x_d^{(i)} \mapsto [x_1^{(i)}, \ldots, x_d^{(i)}, y^{(i)}]$.

- ○ Reduces underfitting
- ○ No change
- ○ Worsens underfitting

Justification:

(e) "Not quite a 2-layer neural network": Fit a regular logistic regression and obtain a prediction $g^{(i)}$ for each element $x^{(i)}$. Add $g^{(i)}$ as a feature: $[x_1^{(i)}, \ldots x_d^{(i)}, g^{(i)}]$ and fit a new logistic regression to the new dataset.

- ○ Reduces underfitting
- ○ No change
- ○ Worsens underfitting

Justification:

Name: ______________________________

Wolfie Madness

5. (9 points) Over the last few years, Wolfie has been seen running through some of the popular lectures on campus. There is no warning as to when or where Wolfie would show up (for example, you might be mid-quiz in 577 and then BAM!, a huge seawolf is sprinting behind your professor in his traditional red uniform).

   New student Tyler is a big fan, and is interested in seeing if he can figure out how Wolfie chooses when and where to appear in class.

   (a) First, Tyler interviews former students about what they can remember about the Wolfie encounters, in order to collect some data.

   For each of the following, suggest an appropriate feature encoding, and provide the dimension of the feature encoding.

   i. Each interviewee remembers the class they were in when they had an encounter. It was always one of {ESE 557, ESE 534, ESE 506, ESE 358, or ESE 301}.

   Encoding:

   Encoding dimension:

   ii. There has long been a rumor on campus that the professors are in on Wolfie's appearances, and they wear distinctive items on the days when Wolfie would show up. By looking at the lecture videos, Tyler can find out if the professor was wearing a bowtie, a hat, or a plaid coat. A professor could wear zero, one, or more of these items.

   Encoding:

   Encoding dimension:

   iii. Tyler can also find out the number of students who were present in lecture for each encounter. This might range from zero students showing up (during busy weeks), to a maximum of 522 students attending (full capacity in the largest lecture hall, Javits-100).

   Encoding:

   Encoding dimension:

(b) Using all of this data, Tyler now wants to make predictions about how likely it might've been to see Wolfie.

He wants to design a simple neural network for the task. What should the dimension of Tyler's *output* prediction be (that is, the shape of his output)? what would be a good choice of activation function on the output layer? Briefly justify your answer.

Dimension:

Activation function:

Justification:

Deep neural networks

6. (18 points) Nori thinks about ReLU units and wonders whether there's a better alternative for an activation function, and decides to explore the LURe function, defined as:

$$f_{\text{LURe}} = \min(z, 0) \ .$$

(a) What is the derivative of this function $\frac{df_{\text{LURe}}(z)}{dz}$?

(b) Nori's friend Ori thinks this is cool and suggests making a neural network with two activation functions per layer, so that:

$$a^l = f_{\text{LURe}}(f_{\text{ReLU}}(z^l)) \ .$$

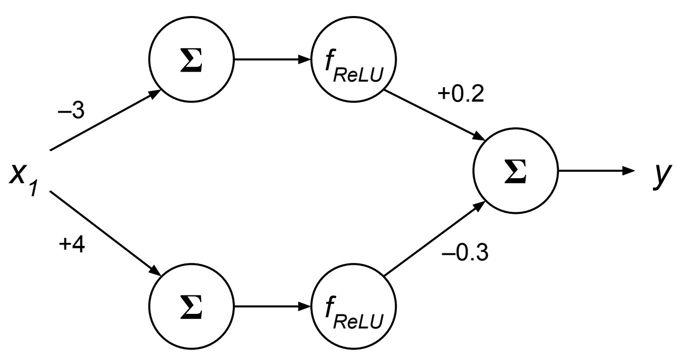
Explain what effect this will have on the network.

(c) Nori's other friend Dori thinks we should try this trick with two ReLUs, so that:

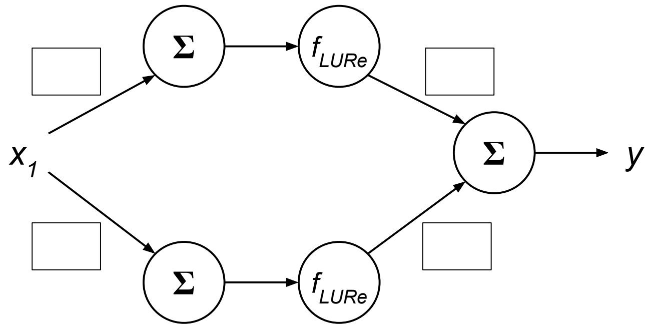
$$a^l = f_{\text{ReLU}}(f_{\text{ReLU}}(z^l)) \ .$$

Explain what effect this will have on the network.

(d) Nori finds a neural network trained by Smaug in his treasure pile that takes a single-dimensional input (so $d = 1$) and looks like this:



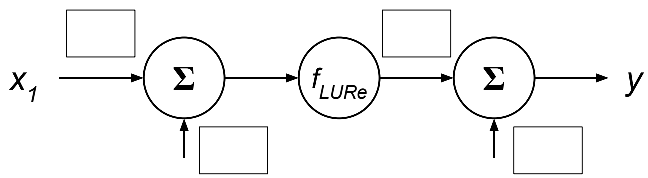
He sees that it computes $\hat{y} = -0.1 \cdot f_{\text{ReLU}}(-5x) + 0.4 \cdot f_{\text{ReLU}}(5x)$ and is very curious to see if he can replace those ReLU activation units with his own LURes. Please help him find another neural network that computes exactly the same function as the one above (that is, maps any input x to the same output as the original one). Provide a set of weights that achieves this in the boxes on the diagram below.



(e) Ori, Dori, and Nori have this dataset with their heights and whether they like beer:

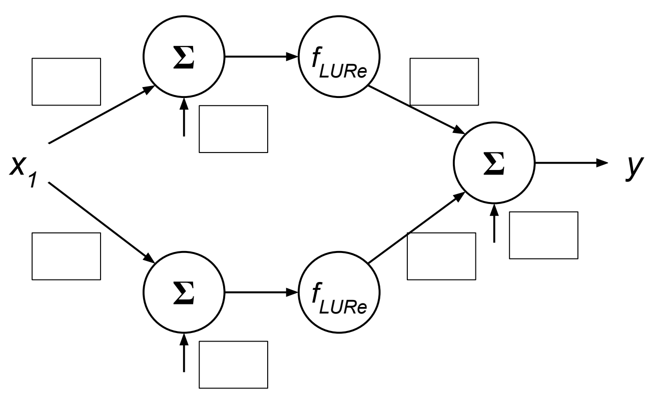
$$x^{(1)} = [1] \qquad y^{(1)} = 1$$
$$x^{(2)} = [2] \qquad y^{(2)} = 0$$
$$x^{(3)} = [3] \qquad y^{(3)} = 1$$

They make a two-layer neural network, as shown below. Are there weights and biases for this network that will predict their data correctly? If so, specify them in the boxes on the network diagram. If not, argue why not.



Or justification:

(f) Thorin suggests they add one more unit, so that they have the architecture below. Are there weights and biases for this network that will predict their data correctly? If so, specify them in the boxes on the network diagram. If not, argue why not.



Or justification: