Midterm Exam, Fall 2025
ESE 577
SOLUTIONS

Instructor: Jorge Mendez-Mendez

Date: Thursday October 9th, 2025

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 80 minutes.
- 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	21
2	16
3	18
4	21
5	24
Total:	100

Name: _____

Linear Regression

1. (21 points) 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
2	0
-3	-8
4	-2

   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. Call ${\theta^{(1)}}^*$ and ${\theta_0^{(1)}}^*$ the parameters that minimize J_1. Can we find ${\theta^{(1)}}^*$ and ${\theta_0^{(1)}}^*$ via the analytical solution formula? (No need for justification.)

   ◯ Yes.

   ◯ No.

   ANSWER: Yes. With a single feature, there is no possibility of collinear features or $n < d$ (since $d = 1$), so the matrix inverse will be well-defined.

   ii. Find a set of values ${\theta^{(1)}}^*$ and ${\theta_0^{(1)}}^*$ that minimize J_1.

   ANSWER: $\theta = 1, \theta_0 = -\frac{13}{3}$. We have that

$$3J_1 = (2\theta + \theta_0)^2 + (-3\theta + \theta_0 + 8)^2 + (4\theta + \theta_0 + 2)^2$$

   The gradient w.r.t. θ give us:

$$\begin{aligned}\nabla_\theta J_1 &= 4(2\theta + \theta_0) - 6(-3\theta + \theta_0 + 8) + 8(4\theta + \theta_0 + 2) \\ &= 58\theta + 6\theta_0 - 32 = 0 \end{aligned} \quad (1)$$

   The gradient w.r.t. θ_0 gives us:

$$\begin{aligned}\nabla_{\theta_0} J_1 &= 2(2\theta + \theta_0) + 2(-3\theta + \theta_0 + 8) + 2(4\theta + \theta_0 + 2) \\ &= 6\theta + 6\theta_0 + 20 = 0 \end{aligned} \quad (2)$$

   Subtracting (1)−(2), we get:

$$52\theta - 52 = 0 \Rightarrow \theta = 1$$

   Substituting θ into (2) gives:

$$6 + 6\theta_0 + 20 = 0 \Rightarrow \theta_0 = \frac{-26}{6} = -\frac{13}{3} \approx -4.33$$

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

Name: _____

x_1	x_2	y
2	2.0001	0
-3	-3.0001	-8
4	4.0001	-2

Suppose that we would like to minimize:

$$J_2(\theta, \theta_0; \mathcal{D}_2) = \frac{1}{n}\sum_{i=1}^{n}(\theta^\top x^{(i)} + \theta_0 - y^{(i)})^2 \ .$$

i. Can we find ${\theta^{(2)}}^*$ and ${\theta_0^{(2)}}^*$ that minimize J_2 via the analytical solution formula? (No need for justification.)

◯ Yes.

◯ No.

ANSWER: Yes. Even though the features are nearly collinear, they are not collinear numerically and so the matrix inverse is well defined.

ii. Compare ${\theta^{(1)}}^*$ and ${\theta^{(2)}}^*$. Which option below is true? Briefly justify your choice.

◯ $\|{\theta^{(1)}}^*\| \ll \|{\theta^{(2)}}^*\|$

◯ $\|{\theta^{(1)}}^*\| \gg \|{\theta^{(2)}}^*\|$

◯ $\|{\theta^{(1)}}^*\| \approx \|{\theta^{(2)}}^*\|$

◯ It depends

Justification:

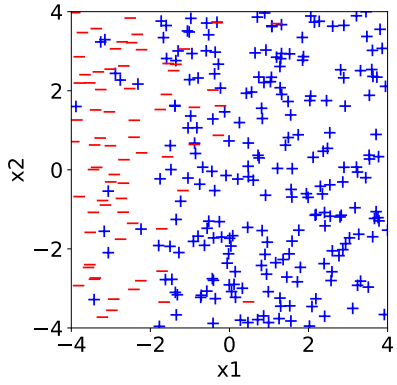
ANSWER: $\|{\theta^{(1)}}^*\| \ll \|{\theta^{(2)}}^*\|$. Note that in $\mathcal{D}_2$, the 2nd feature is linearly dependent on the 1st feature. Intuitively, this means that the 2nd feature gives us nothing additional to learn from towards alinear hypothesis, i.e. the 2nd feature is redundant.

Name: _____

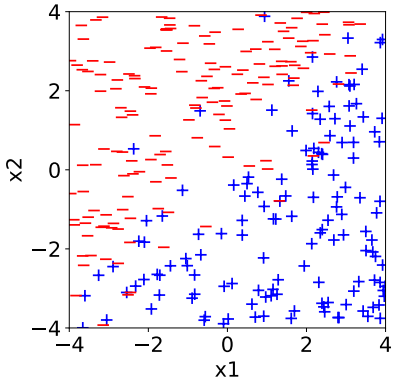
Logistic Mix-Up

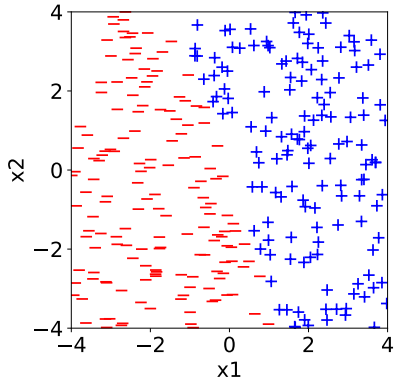
2. (16 points) Below we show six different datasets in 2 dimensions. Points displayed as '–' have a negative label and points displayed as '+' have a positive label. We learned an unregularized linear logistic classifier for each of them. But they all got mixed up! Please help us match each set of parameters to the dataset they came from. (Each set of parameters is used exactly once.)

(a)

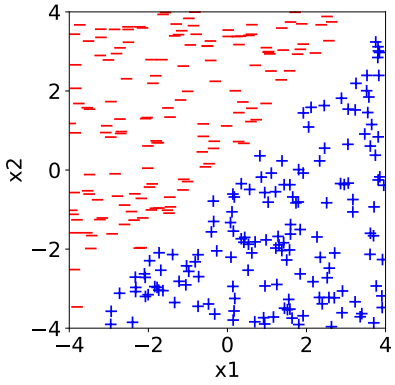


(b)





(c)



(d)

Name: _____

i. $\theta = \begin{bmatrix} 10 & -10 \end{bmatrix}^\top$, $\theta_0 = 0$: ◯ (a) ◯ (b) ◯ (c) ◯ (d)

ii. $\theta = \begin{bmatrix} 1 & -1 \end{bmatrix}^\top$, $\theta_0 = 0$: ◯ (a) ◯ (b) ◯ (c) ◯ (d)

iii. $\theta = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$, $\theta_0 = 2$: ◯ (a) ◯ (b) ◯ (c) ◯ (d)

iv. $\theta = \begin{bmatrix} 9 & 3 \end{bmatrix}^\top$, $\theta_0 = 0$: ◯ (a) ◯ (b) ◯ (c) ◯ (d)

ANSWER: (d), (b), (a), (c). There is a fast and a slow way to answer this question. The slow way is to find the line that each separator represents and determine which plot(s) can be well separated by that line. The fast way is to use the normal vectors' directions to determine where the positives should lie. For example, $\theta = \begin{bmatrix} 10 & -10 \end{bmatrix}^\top$ and $\theta = \begin{bmatrix} 1 & -1 \end{bmatrix}^\top$ both point to the lower right, which matches plots (b) and (d). Because (d) is linearly separable and (b) is not, the scale of parameters for (d) will be greater. $\theta = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$ points to the right, which only matches (a). $\theta = \begin{bmatrix} 9 & 3 \end{bmatrix}^\top$ points to the upper right, and the scale for (c) parameters must be large because the data set is separable. (Note that, using the slow way, you needed to use the same reasoning to distinguish (d)/(b).)

Name: _____

Classification Features

3. (18 points) Seawolves Game Day Predictions (25 points)

   You're using machine learning to predict the attendance at Stony Brook Seawolves basketball games. You have collected data from past games and need to encode the features properly before training a regression model.

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

   (a) **Opponent:** The Seawolves regularly play against CAA teams, and there are 13 of them.

   Encoding:

   Encoding dimension:

   ANSWER: This is a categorical variable, and there is no relationship between each class. We can use a one-hot vector of size 13, in which only a single class is marked at a time.

   (b) **Game importance:** People may be more likely to attend a game if it is more important. We identify three levels: "regular season", "playoffs", and "championship".

   Encoding:

   Encoding dimension:

   ANSWER: We expect attendance to be higher for championship games than for playoff games, and for playoff games than for regular season games, but we do not know how much more attendance to expect. A thermometer encoding is a good choice: [1,0,0] for regular season, [1,1,0] for playoffs, and [1,1,1] for championship.

   (c) **Tip-off Time:** Games start at the hour sharp (e.g., 11 am, 12 pm, 7 pm) on certain hours of the day. Peak times are 11 am – 2 pm and 7 pm – 11 pm, with lower attendances in-between.

Name: _____

Encoding:

Encoding dimension:

ANSWER: A one-hot encoding of dimension 24 is a good choice here (if we had better knowledge of which hours games could be scheduled on—e.g., no games at 2 am—we could lower this dimension). A normalized scalar encoding is not a good choice, since we know that the relationship between time and attendance is non-linear, since earlier and later times are the peaks of attendance.

Name: _____

Gradient Descent

4. (21 points) We will do standard gradient descent iterations:

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

on the following piecewise function:

$$f(x) = \begin{cases} (x+1)^2 & \text{if } x \leq 3 \\ (x-5)^2 + 12 & \text{if } x > 3 \end{cases} \ .$$

For the purposes of this question, assume that the gradient at $x = 3$ is $\nabla f(x)\Big|_{x=3} = 0$.

(a) Starting from the initial guess $x^{(0)} = -6$ and using a step size $\eta = 1$, what will be the values of x for the first three iterations of gradient descent, $x^{(1)}$, $x^{(2)}$, and $x^{(3)}$?

Name: _____

$x^{(1)} =$

$x^{(2)} =$

$x^{(3)} =$

ANSWER: $x^{(1)} = 4, x^{(2)} = 6, x^{(3)} = 4$.

$$\nabla f(x) = \begin{cases} 2(x+1) & \text{if } x < 3 \\ 0 & \text{if } x = 3 \\ 2(x-5) & \text{if } x > 3 \end{cases}$$

$$\begin{aligned} \nabla f(x)\Big|_{x=x^{(0)}=-6} &= 2(-6+1) = -10 \\ x^{(1)} &= x^{(0)} - \eta(-10) = 4 \end{aligned}$$

$$\begin{aligned} \nabla f(x)\Big|_{x=x^{(1)}=4} &= -2 \\ x^{(2)} &= x^{(1)} - \eta(-2) = 6 \end{aligned}$$

$$\begin{aligned} \nabla f(x)\Big|_{x=x^{(2)}=6} &= 2(6-5) = 2 \\ x^{(3)} &= x^{(2)} - \eta(2) = 4 \end{aligned}$$

(b) For each of the following statements, state whether the statement is true or false, and briefly (in one sentence) justify your choice.

i. For any function f, finding the point at which $\nabla f = 0$ is equivalent to finding the global minimum of that function

Name: _____

TRUE FALSE (Circle one)

Justification:

ANSWER: False: for non-convex functions, other points such as local minima, global/local maxima, and saddle points may also have $\nabla f = 0$.

ii. For the piecewise function f in part (a), starting from $x^{(0)} = -6$ and using a step size $\eta = 1$—the same values from part (a)—we can converge to a minimum (local or global).

TRUE FALSE (Circle one)

Justification:

ANSWER: False: as we found in part (a), the first three iterates are: $4, 6, 4$. Because we use a constant step size $\eta = 1$, the next step will bring us back to $x = 6$ and the algorithm will oscillate perpetually between 4 and 6.

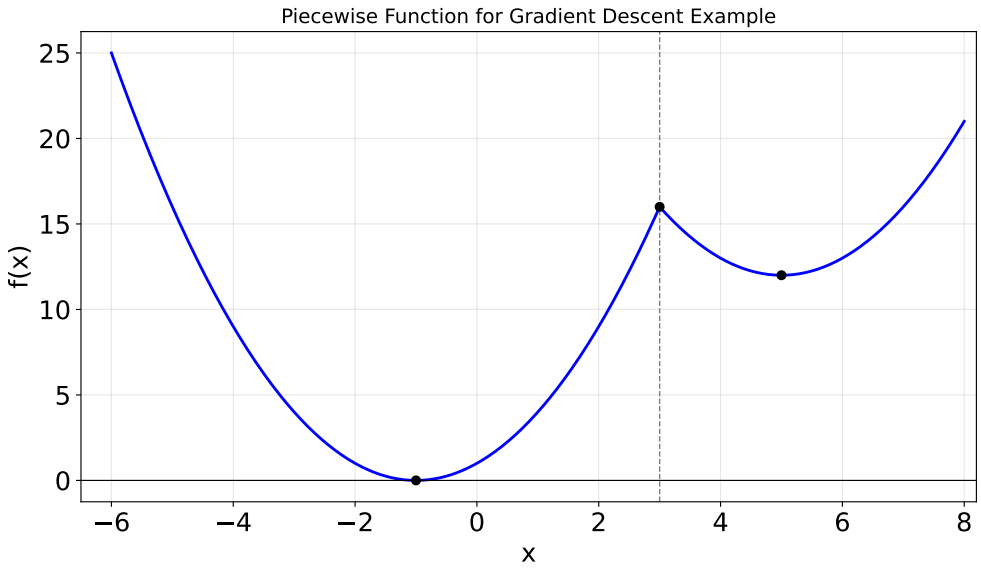
iii. For the piecewise function f in part (a), starting from $x^{(0)} = 8$ and using a step size $\eta = 1e-3$ fails to find find the global minimum.

Name: _____

TRUE FALSE (Circle one)

Justification:

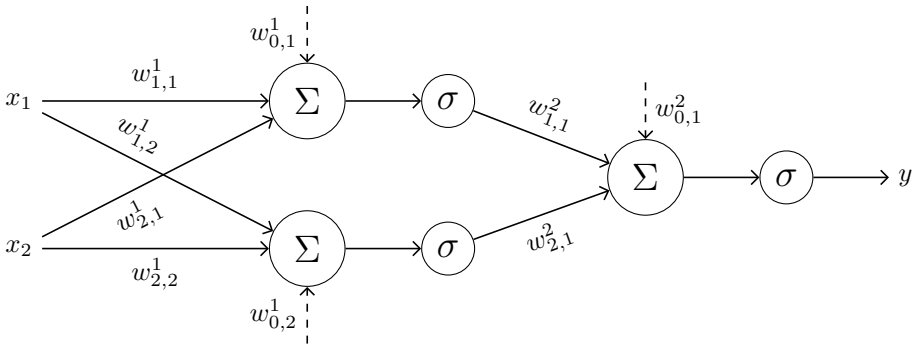
ANSWER: True: at $x = 8$ using a small step size will converge to the local minimum at $x = 5$. Plotting the functions as shown below was an easy way to determine this:



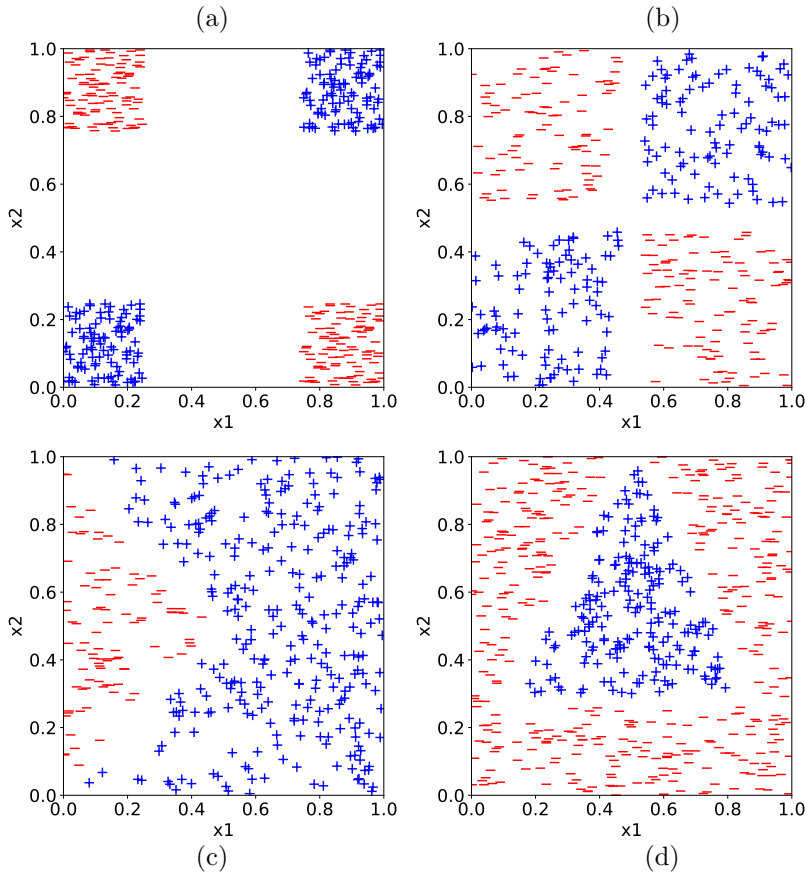
Name: _____

Neural Networks

5. (24 points) We will explore what types of datasets can be perfectly separated by a simple neural network with two hidden units with sigmoid activation, as the one shown below. The two input features are x_1 and x_2, arrows into the summation nodes are are annotated with the weights for the linear combination, biases are represented with dashed arrows, and σ represents the sigmoid activation.



Below we show six different datasets in 2 dimensions. Points displayed as '–' have a negative label and points displayed as '+' have a positive label. For each dataset, state whether it can be perfectly separated by the neural net. If it can, explain a possible transformation learned by each neuron that achieves perfect accuracy (e.g., "neuron one learns a linear separator that transforms points to the right of 0.8 to 1 and others to 0"). If not, briefly (in 1 sentence) explain why not.



Name: _____

(a)

Can be separated (circle one): Yes No
Justification:

ANSWER: Yes. One hidden neuron learns a linear separator that maps points in the bottom right corner to 1 and others to 0, the other hidden neuron learns a similar separator that maps points in the upper right corner to 1 and others to 0. The result is that all '–' are mapped to $0, 0$ and all '+' are mapped to either $1, 0$ or $0, 1$. The output neuron learns a linear separator with slope -1 that maps the '+' to 1 and the '–' to 0.

(b)

Can be separated (circle one): Yes No
Justification:

ANSWER: No. No separator like those described in (a) can map the bottom-right or top-left corners correctly to $1, 0$ or $0, 1$. If instead we use a vertical separator to send points to the right/left of 0.5 to 0/1, and a horizontal separator to send points above/below 0.5 to 0/1, we end up with a non-linearly separable dataset (the X-OR!), which cannot be separated by a single output node.

(c)

Can be separated (circle one): Yes No
Justification:

ANSWER: Yes. One hidden neuron learns a linear separator with slope -1 that maps points above $x_2 = -x_1 + 1$ to 1 and others two zero, and the other hidden neuron learns a similar separator with slope 1 that maps points below $x_2 = x_1$ to 1 and others to 0. This becomes an 'OR' problem: all positive points are $0, 1$, $1, 0$, or $1, 1$, and all negative points are $0, 0$. The output neuron learns a linear separator with slope -1 that maps the $0, 0$ to 0 and others to 1.

(d)

Can be separated (circle one): Yes No
Justification:

ANSWER: No. We need three linear separators to define the edges of the triangle as our transformations, but we only have two hidden units.