6.390: Midterm 1, Fall 2025

Exam – Batch 0 | Serial 0

- This exam is closed-book, and you may **not** use any electronic devices (including computers, calculators, phones, etc.). The total exam time is 1.5 hours.
- One reference sheet (8.5 in. by 11 in.) with notes on both sides is permitted. Blank scratch paper will also be provided if needed. You do **not** need to submit your reference sheet or the scratch paper.
- Each exam has a unique batch number and serial number. Your exam's batch and serial numbers appear on every page. You **only** need to write your name and Kerberos on this front page.
- The problems are not necessarily presented in any order of difficulty.
- Please write all answers in the provided boxes. If you need more space, clearly indicate near the answer box where to find your work.
- Unless otherwise specified, for all multiple-choice questions please **select all that apply.** If you want to change your selections, please **write your final answers clearly** instead of marking over your selected options.
- If you have a question, please **come to us directly**. You may also raise your hand, but if we do not see you, please approach us.
- You may **not** discuss the details of the exam with anyone other than the course staff until exam grades have been assigned and released.

Name: ______________________ Kerberos: ______________________

Question:	1	2	3	4	Total
Points:	19	23	30	28	100
Score:					

Linear Regression

1. Given a training dataset with 3 data points having 2-dimensional features:

$$D_{\text{train}} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)})\} = \{([1, 0], 2), ([0, 1], 1), ([1, 1], 3)\}$$

Alice wants to find an optimal linear hypothesis $h(x) = \theta^T x$ (no offset term) to minimize the mean-squared error (MSE).

(a) (3 points) Give X and Y such that $J(\theta) = \frac{1}{n}(X\theta - Y)^T(X\theta - Y)$ represents the MSE of the linear hypothesis for this dataset.

$X =$ $Y =$

(b) (4 points) Bob thought adding another feature could help make a more informed decision. So, they went out and collected another piece of information. Now the dataset has 3 features:

$$D_{\text{train}} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)})\} = \{([1, 0, 1], 2), ([0, 1, 1], 1), ([1, 1, 3], 3)\}$$

Does there exist at least one optimal θ^* that minimizes the MSE? ◯ Yes ◯ No

If yes, such a θ^* exists, answer the next two yes/no questions. If no, leave them blank:

- Can we apply the closed-form solution formula to find such a θ^*? ◯ Yes ◯ No
- Can we apply gradient descent to find such a θ^*? ◯ Yes ◯ No

Brief explanation:

(c) (4 points) Charlie thought perhaps it'd help to add one more feature with random integer values:

$$D_{\text{train}} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)})\} = \{([1, 0, 1, 3], 2), ([0, 1, 1, 7], 1), ([1, 1, 3, 2], 3)\}$$

Does there exist at least one optimal θ^* that minimizes the MSE? ◯ Yes ◯ No

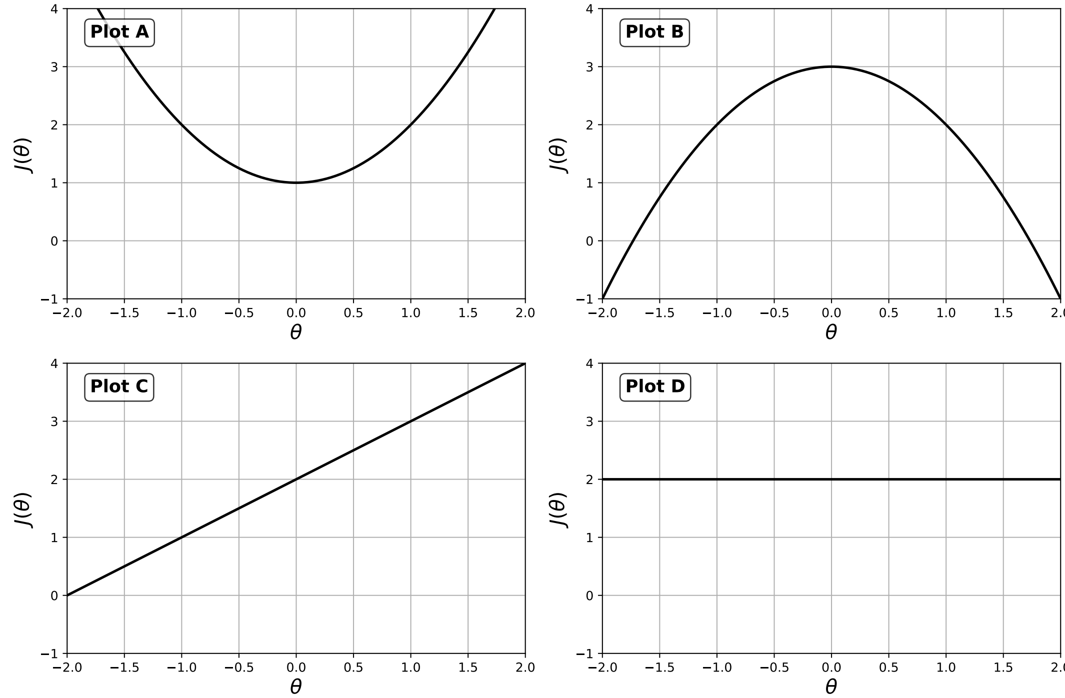
If yes, such a θ^* exists, answer the next two yes/no questions. If no, leave them blank:

- Can we apply the closed-form solution formula to find such a θ^* ? ◯ Yes ◯ No
- Can we apply gradient descent to find such a θ^*? ◯ Yes ◯ No

Brief explanation:

The parts below assume a general setting, instead of specific to any given data set.

(d) (4 points) Consider the following four plots of objective functions $J(\theta)$ vs. $\theta \in \mathbb{R}$:



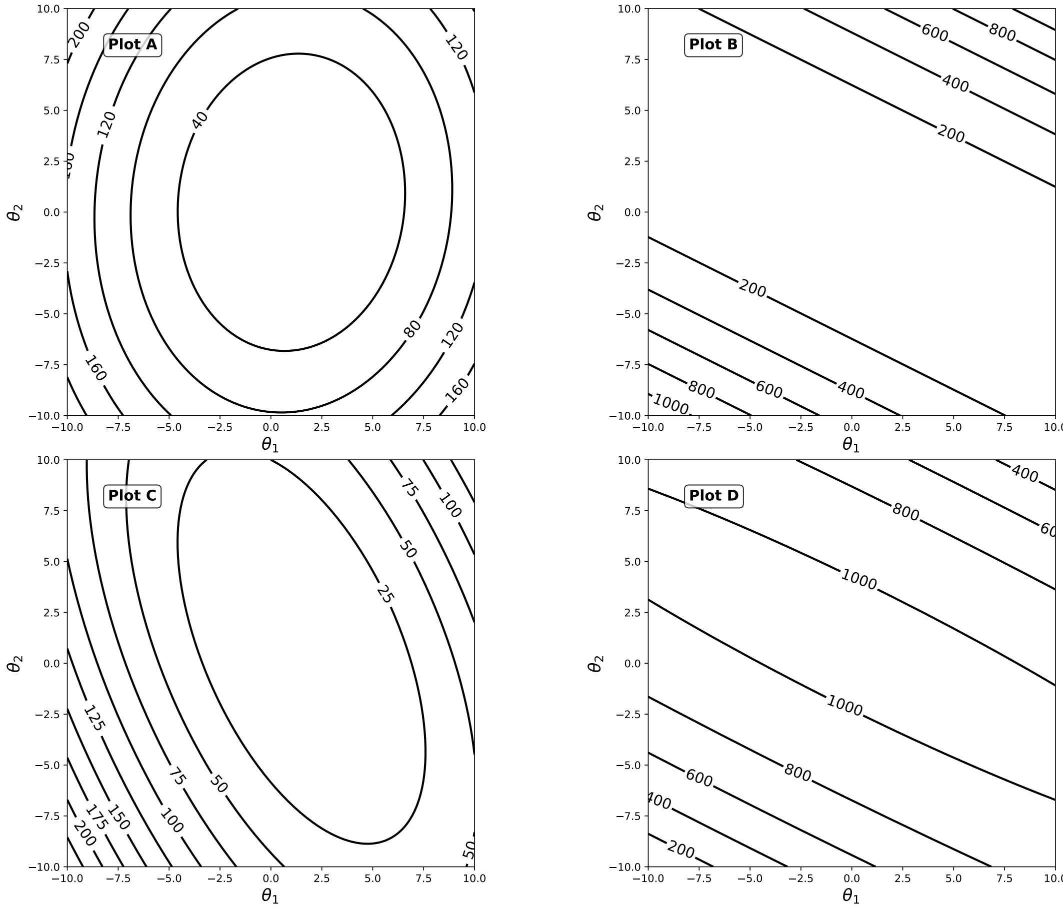
Which of these plots could possibly represent an MSE of a linear hypothesis $h(x) = \theta^T x$ (no offset term) on some data set?

Reminder (copied from the exam cover page instructions):
Unless otherwise specified, for all multiple-choice questions please **select all that apply.**
If you want to change any of your initial selections, please **write your final answers clearly** instead of marking directly on the option choices.

○ Plot A ○ Plot B ○ Plot C ○ Plot D

Brief explanation:

(e) (4 points) Consider the following four 2D contour plots of objective functions $J(\theta)$ vs. $\theta \in \mathbb{R}^2$:



Which of these plots could possibly represent an MSE of a linear hypothesis $h(x) = \theta^T x$ (no offset term) on some dataset?

◯ Plot A ◯ Plot B ◯ Plot C ◯ Plot D

Brief explanation:

Regularization and Cross-validation

2. In this problem, we investigate how the hyperparameter λ in ridge regression influences the learned parameters.

   (a) (6 points) We're minimizing the ridge regression objective function. For a range of λ values, we used the closed-form solution for getting the optimal parameters.

   We then used these parameters to get the MSE on the training data set.

   As we increase λ, which of the following best describes the MSE on the training data?

   ◯ A monotonically increasing curve (MSE increases as λ increases)
   ◯ A monotonically decreasing curve (MSE decreases as λ increases)
   ◯ A U-shaped curve (MSE decreases then increases)
   ◯ A constant horizontal line (MSE stays the same)

   Brief explanation:

   (b) As we saw in class, one common approach to evaluate and choose λ is *cross-validation*:

   1. Divide data $\mathcal{D}$ into $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{validation}}$.
   2. Further divide $\mathcal{D}_{\text{train}}$ into k chunks $\mathcal{D}_1, \ldots, \mathcal{D}_k$.
   3. For each candidate value of λ:
      (a) For $i = 1$ to k:
         i. Train a ridge regressor h_i using $\mathcal{D}_{\text{train}} \setminus \mathcal{D}_i$ (i.e., all training chunks except $\mathcal{D}_i$)
         ii. Compute the chunk-i validation error $E_i(h_i)$ on $\mathcal{D}_i$.
      (b) Compute the average validation error $E_\lambda := \frac{1}{k}\sum_{i=1}^{k} E_i(h_i)$.
   4. Choose λ^* with __________.
   5. Retrain a final model h^* using __________ to ship.

   i. (3 points) Fill in the blank (using either words or mathematical expressions):

   Choose λ^* with __________.

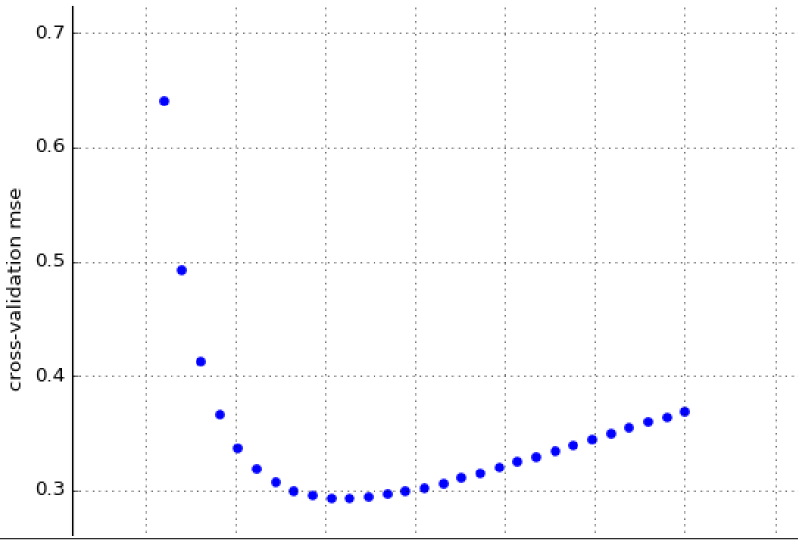
   ii. (3 points) For this blank, "Retrain a final model h^* using __________ to ship", what's the appropriate data set to use:

   ◯ the full validation set $\mathcal{D}_{\text{validation}}$
   ◯ the full training set $\mathcal{D}_{\text{train}}$
   ◯ the union of the training and validation sets $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{validation}}$
   ◯ $\mathcal{D}_{\text{train}} \setminus \mathcal{D}_i$, i.e., all training chunks except $\mathcal{D}_i$ where i is a randomly chosen chunk

iii. (5 points) True/False: In line 5 (retraining the final model), we use the same objective function as in the cross-validation steps, which includes the regularization term $\lambda\|\theta\|^2$.

◯ True ◯ False
Brief explanation:

iv. (6 points) Plot below shows the validation error E_λ for a range of λ values.
Notice that the horizontal axis denotes λ. However, we lost all the tick values. In particular, this axis is not necessarily increasing in λ value.
Fortunately, we do have a record that $\lambda = 1.11$ is the approximate value of λ that gives the minimum validation error.
What is the approximate value of λ at the leftmost tick mark?



◯ $\lambda = 0.1$ ◯ $\lambda = 10$
Brief explanation:

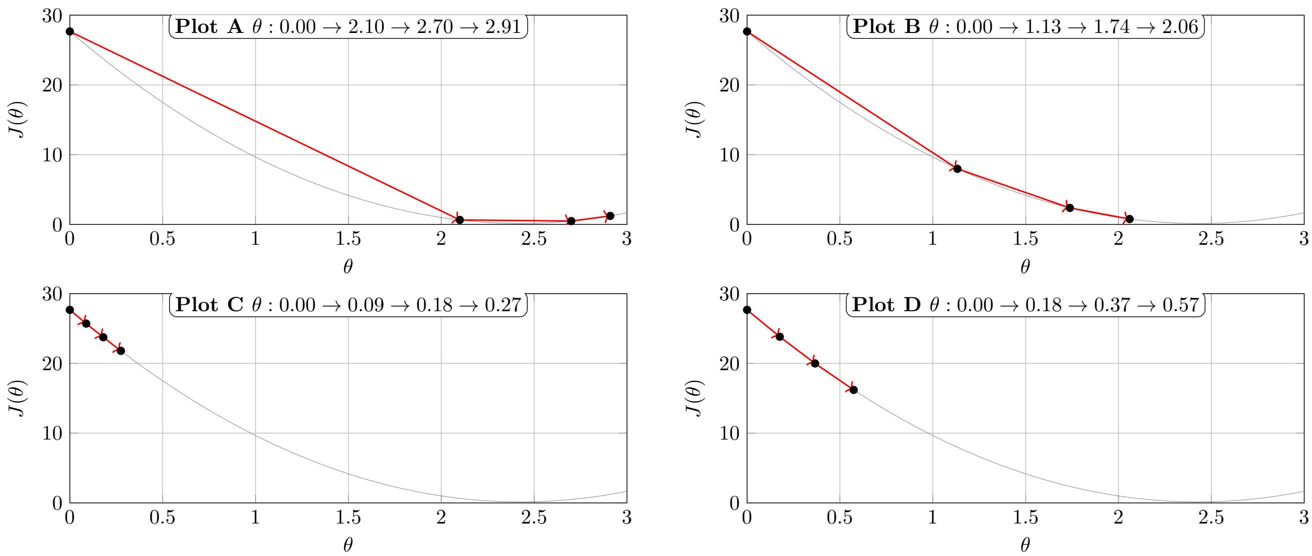
Gradient Descent

3. (a) Given a training dataset with 3 data points having 1-dimensional features:

$$\mathcal{D}_{\text{train}} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)})\} = \{(1, 3), (2, 5), (3, 7)\}$$

We aim to learn a linear regressor $h(x; \theta) = \theta x$ (no offset term) to minimize the MSE $J(\theta)$. We were able to show that the optimal $\theta^* = \frac{17}{7}$ using the closed-form solution. In this part, we focus on understanding the behavior of gradient descent instead.

i. (5 points) We first try gradient descent (GD), with initial parameter $\theta = 0$, a constant learning rate $\eta > 0$ and we run it for 3 iterations. Which of the following could be a possible plot for this GD run?
Hint: Focus on conceptual reasoning.



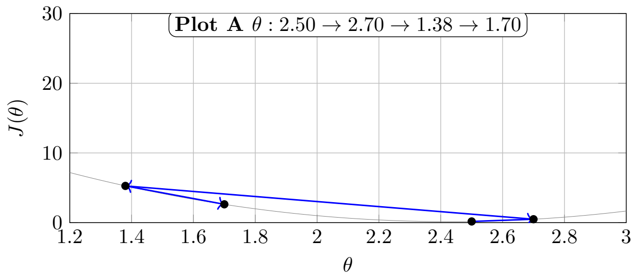
○ Plot A ○ Plot B ○ Plot C ○ Plot D
Brief explanation:

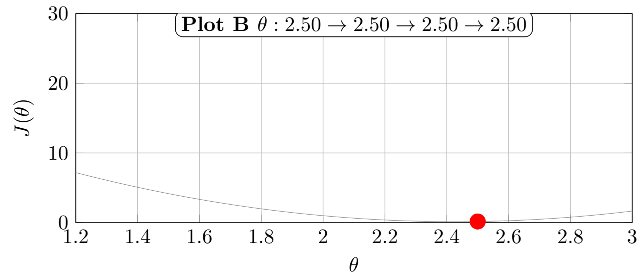
ii. (6 points) Suppose we run GD with initial parameter $\theta = 2$ and a constant learning rate $\eta > 0$ for one iteration. Let's call the initial parameter value θ_{old} (so $\theta_{\text{old}} = 2$) and the updated parameter value θ_{new}. What is the range of η such that $J(\theta_{\text{new}}) \leq J(\theta_{\text{old}})$? You can use the fact that the optimal θ^* is $\frac{17}{7}$ if needed.

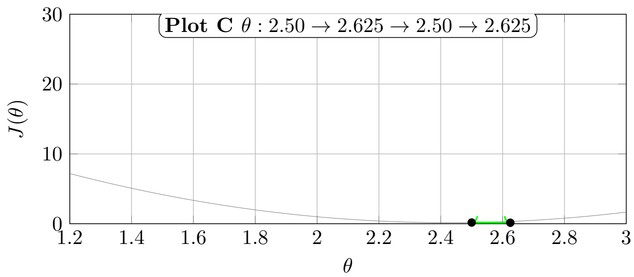
iii. (3 points) Now we run stochastic gradient descent (SGD) on the same dataset with initial parameter $\theta = 0$ and a constant learning rate $\eta > 0$.
After the first iteration of SGD, how many possible values are there for the resulting updated parameter value?

◯ 1 ◯ 2 ◯ 3 ◯ Not enough info to determine.

iv. (7 points) We run stochastic gradient descent (SGD) to minimize $J(\theta)$ with initial parameter $\theta = 2.5$ and a constant learning rate $\eta = 0.125$. We run the algorithm for 3 iterations. Which of the following could be a possible plot for this SGD run?







◯ Plot A ◯ Plot B ◯ Plot C
Brief explanation:

(b) (5 points) Given a training data set with 3 data points having 2-dimensional features:

$$D_{\text{train}} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)})\} = \{([3,0],3), ([0,3],6), ([3,3],9)\}$$

We want to minimize the ridge objective function:

$$J(\theta) = \frac{1}{3}\|X\theta - Y\|^2 + \lambda\|\theta\|^2$$

where $\lambda = 1$.
Compute the gradient $\nabla J(\theta)$ and evaluate the gradient at the initial parameter values $\theta = [1,1]^T$.

(c) (4 points) Consider the following objective function:

$$J(\theta) = \theta^T\theta + \|X\theta - Y\|^2 + \|\theta^T X^T Y\|^2$$

Which of the following expressions could correctly represent $\nabla_\theta J$?

- ○ $\nabla_\theta J = 2\theta^T + X(X\theta - Y) + XY$
- ○ $\nabla_\theta J = \theta^T + 2X^T(X\theta^T - Y) + \theta^T X^T Y$
- ○ $\nabla_\theta J = 2\theta + 2X^T X\theta - 2X^T Y + 2X^T Y$
- ○ $\nabla_\theta J = 2\theta + X^T X\theta - X^T Y + X^T Y X^T Y$
- ○ $\nabla_\theta J = 2\theta + 2X^T(X\theta - Y) + 2(\theta^T X^T Y)X^T Y$
- ○ $\nabla_\theta J = 2\theta + 2X^T X\theta - 2X^T Y + X^T Y$

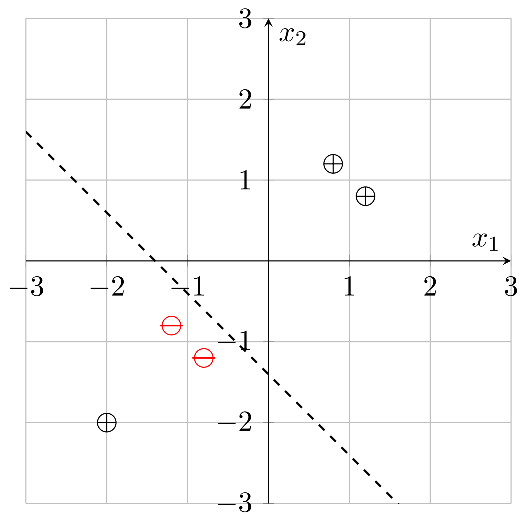
Linear Classification

4. (a) Given a training dataset with 5 data points having 2-dimensional features for binary classification, where $\oplus$ represents positive class ($y = 1$) and $\ominus$ represents negative class ($y = 0$). The plot below shows the dataset and logistic regression results.

The dashed line is the decision boundary, given by the optimal parameters from using a logistic regression hypothesis: $h(x) = \sigma(\theta^T x + \theta_0)$ to minimize the negative log-likelihood loss:

$$\mathcal{L}_{\text{NLL}}(g, y) = -(y \log g + (1 - y) \log(1 - g)).$$

The decision boundary intersects the axes at $(-1.4, 0)$ and $(0, -1.4)$.



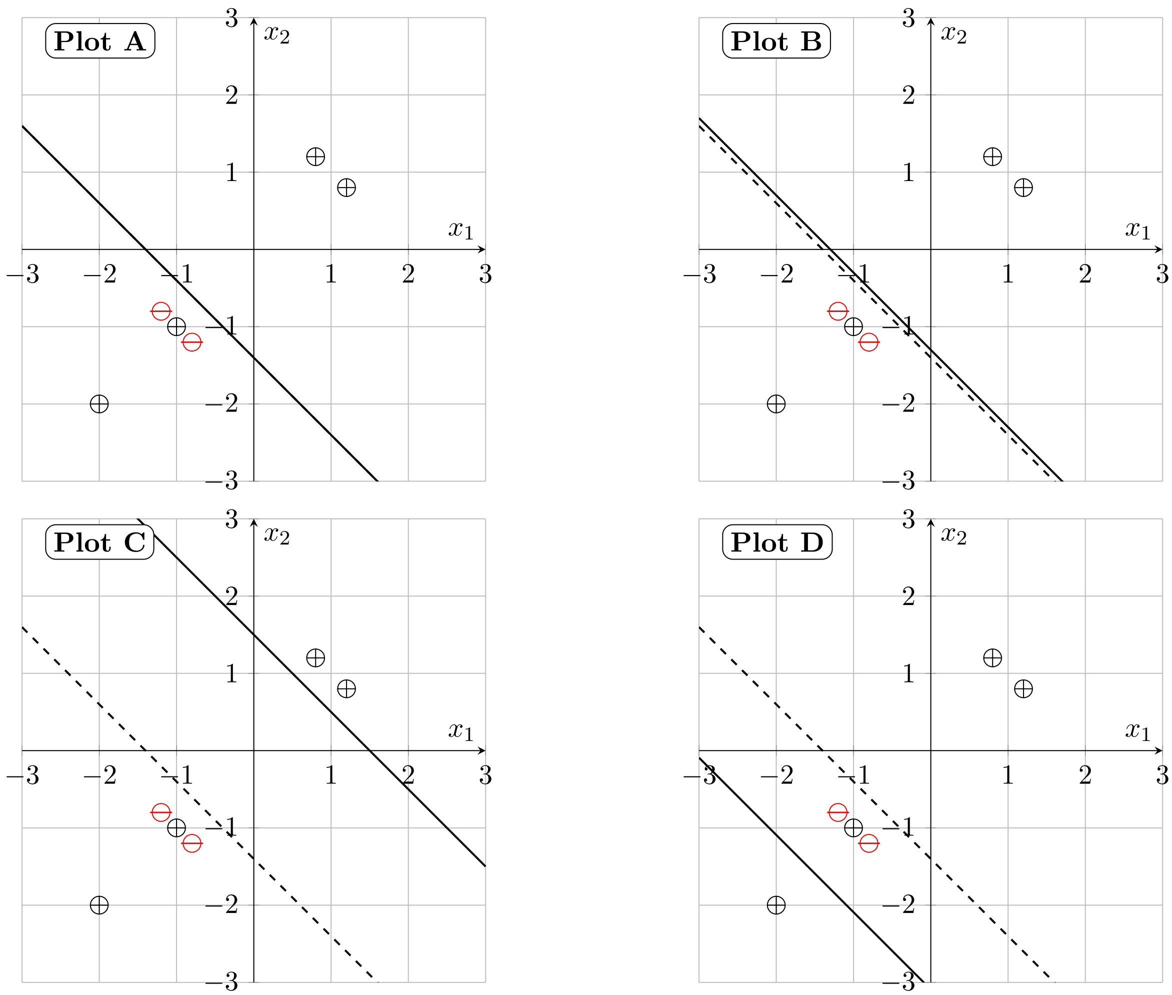
i. (4 points) Can we determine the normal vector direction? If yes, give the normal vector direction, as numerical values. If no, explain why not.

◯ Yes ◯ No
Normal vector direction, or why not possible to determine:

ii. (4 points) Can we determine $h(x)$ at point $(-2, -2)$? If yes, give it as a numerical value. If no, explain why it is not possible to determine.

◯ Yes ◯ No
$h(x)$ at point $(-2, -2)$, or why not possible to determine:

iii. (4 points) We add a new positive data point at $(-1, -1)$ and rerun logistic regression. We obtain the new optimal parameters to draw the decision boundary.
In each plot, the dashed line shows the boundary learned from the original 5-point dataset; the solid line shows a decision boundary for the expanded 6-point dataset.
Which **one** of the plots correctly shows the new decision boundary (solid line) given by the optimal parameters for the 6-point dataset?
Hint: Focus on conceptual reasoning.



◯ Plot A ◯ Plot B ◯ Plot C ◯ Plot D
Brief explanation:

(b) Now consider a classification problem with one feature. We consider two kinds of linear classifiers:
Binary Logistic Classification: where $h(x) = \sigma(\theta^T x + \theta_0)$ with $\theta = 1$ and $\theta_0 = -1$.
3-Class Softmax Classification: where $h(x) = \text{softmax}(\theta^T x + \theta_0)$ with $\theta = \begin{bmatrix} 1 & 2 & 0 \end{bmatrix}$ and $\theta_0^T = \begin{bmatrix} 0 & -1 & 1 \end{bmatrix}$.

i. (4 points) For $x = -2$, what are the predicted classes?

Predicted class for binary classifier: ◯ Positive class ◯ Negative class
Predicted class for 3-class classifier: ◯ Class 1 ◯ Class 2 ◯ Class 3
Brief explanation:

ii. (4 points) We increase the binary logistic classifier offset by 10, from $\theta_0 = -1$ to $\theta_0 = 9$. What is the new predicted class for $x = -2$?

Predicted class for the modified binary classifier: ◯ Positive class ◯ Negative class
Brief explanation:

iii. (4 points) We increase the softmax offsets by 10, from $\theta_0^T = \begin{bmatrix} 0 & -1 & 1 \end{bmatrix}$ to $\theta_0^T = \begin{bmatrix} 10 & 9 & 11 \end{bmatrix}$. What is the new predicted class for $x = -2$?

Predicted class for the modified 3-class classifier: ◯ Class 1 ◯ Class 2 ◯ Class 3
Brief explanation:

iv. (4 points) What is the range of x such that increasing the offset parameters by 10 leaves the predicted classes unchanged for both classifiers? Show your work.

This is the end of the exam.