6.390: Midterm 1, Fall 2025

Solutions

- 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.

Solution: $X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$, $Y = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$

Recall that our linear hypothesis takes the form $h(x) = \theta^T x$ for any individual input vector. We can let the i^{th} row of X be equal to $x^{(i)T}$ for $i = 1, 2, 3$, i.e., each row of X is a data point (transposed). Similarly, the i^{th} row of Y can be equal to $y^{(i)}$.

(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:

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

Solution:

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

As the columns of X are linearly independent, the matrix $X^T X$ is invertible. In this case, the mean-squared error objective function is strictly convex, so there exists a unique global minimum. Additionally, the MSE objective function is differentiable everywhere, meaning that we can perform gradient descent.

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

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

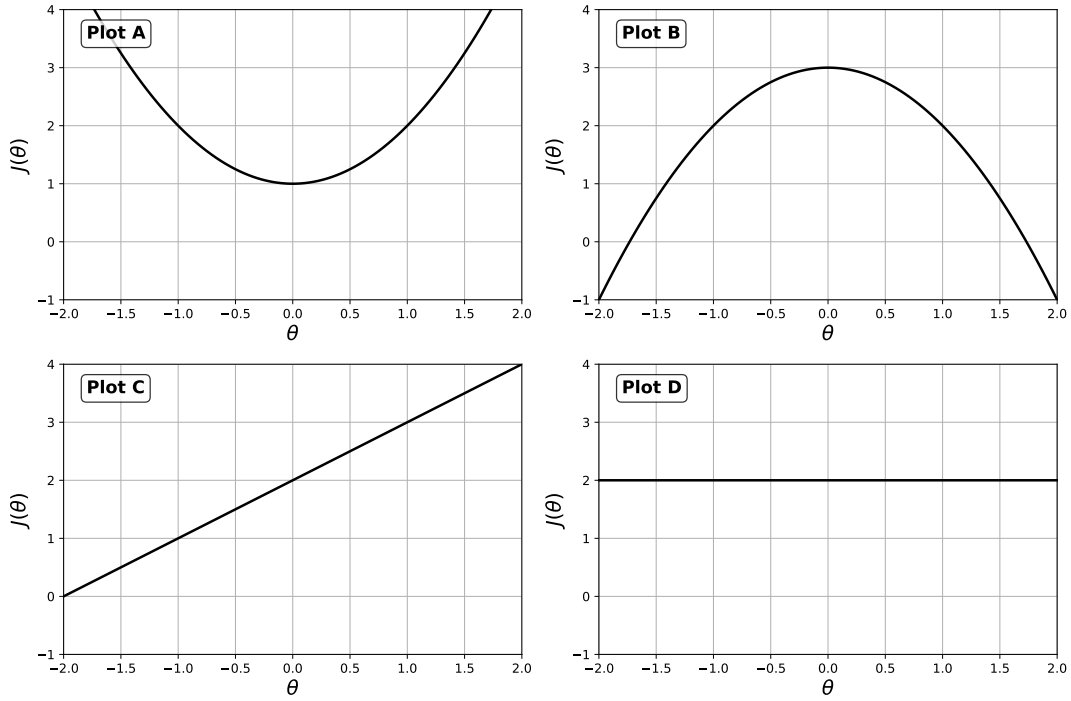
Solution:

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

With 3 data points and 4 features, multiple solutions can achieve zero MSE. The closed-form solution will not be valid because $X^T X$ is not invertible. Gradient descent will still work and converge to some solution that minimizes MSE.

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.

Solution: ☑ Plot A ☐ Plot B ☐ Plot C ☑ Plot D

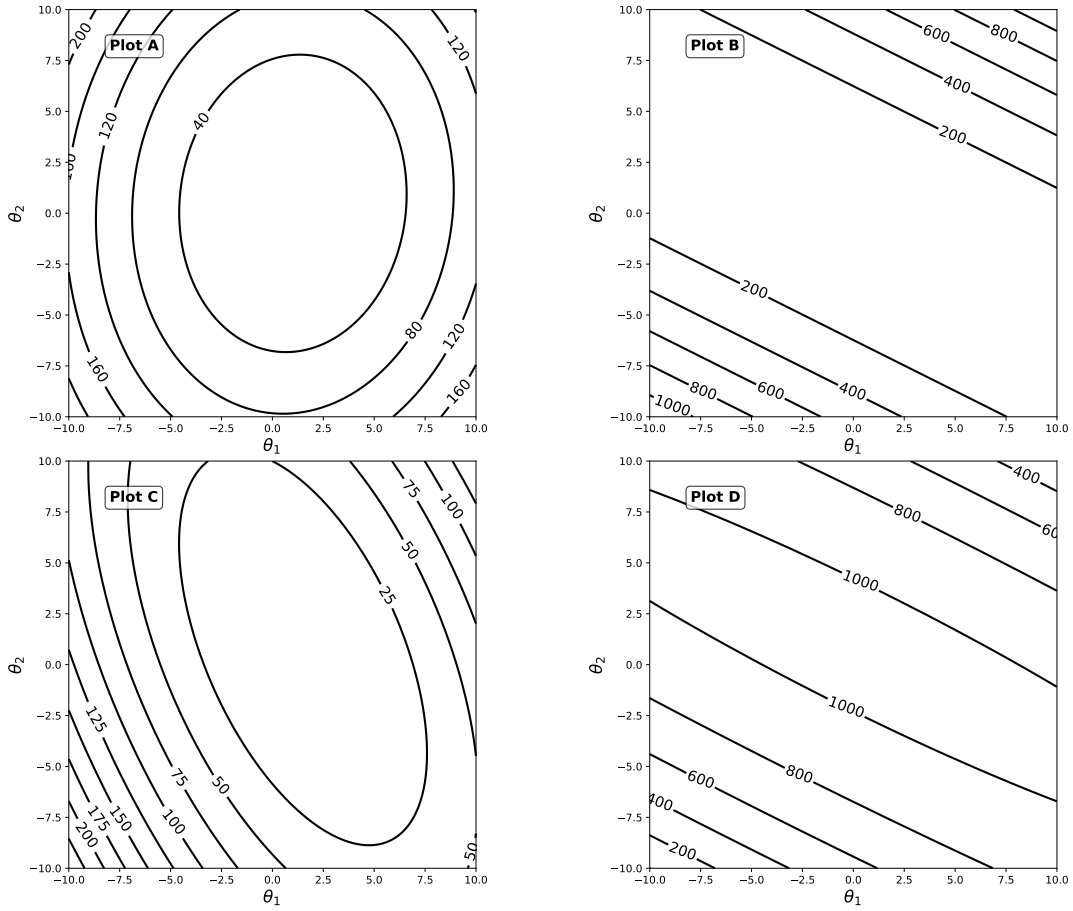
Key observation is that the MSE is an average of squared errors, so it is always non-negative.

This immediately rules out Plot B and Plot C, which are unbounded below.

Plot A is the standard convex MSE objective parabola.

Plot D, as we discussed in class, can occur if $h(0) = 0$, which would happen if all training data points have a zero feature.

(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?

Solution: ☑ Plot A ☑ Plot B ☑ Plot C ○ Plot D

Explanation:

Plots A and C correspond to a "bowl", which we expect when there is a unique solution to the MSE objective function.

Plot B corresponds to the "half pipe", which we expect when there are infinitely many solutions to the MSE objective function.

Plot D is an "upside-down bowl", unbounded from below, which cannot represent the MSE of a linear hypothesis.

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?

   Solution: (a) A monotonically increasing curve (MSE increases as λ increases).

   Explanation: As λ increases, the regularization term $\lambda\|\theta\|_2^2$ becomes more dominant in the ridge objective function, forcing the parameters to be smaller. This typically leads to higher training error (MSE) because the model becomes more constrained and less able to fit the training data perfectly. With stronger regularization, the model trades off some training accuracy for better generalization.

   (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):

   Solution: smallest validation error E_λ

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

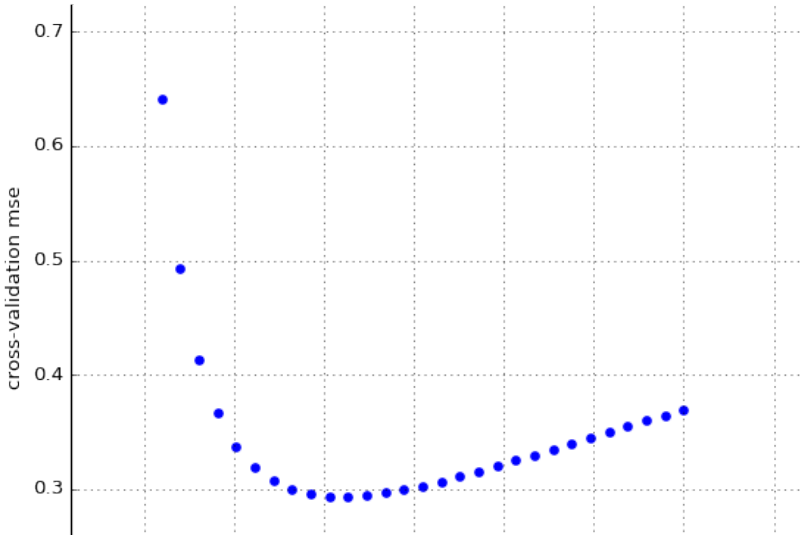
   Solution: As we saw in recitation 2, the correct answer is "the full training set $\mathcal{D}_{\text{train}}$". We expect to later test our final model h* on the unseen data $\mathcal{D}_{\text{validation}}$

   Note that we also give full credit for "the union of the training and validation sets $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{validation}}$" since this is the answer to a very similar homework question. But you should compare and contrast the two questions to understand the difference.

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 $\lambda\|\theta\|^2$.

Solution: True. When retraining the final model, we use the same ridge regression objective function with the selected λ^* value so the final shipped model achieves the "best" trade off between training error and regularization.

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?



Solution: $\lambda = 0.1$.
On both sides of the extreme, the validation error is high. This makes sense because the model is either overfitting (when λ is small) or underfitting (when λ is large). Purely based on the absolute validation error, it's slightly still unclear which side is which.
As λ increases, the regularization term $\lambda\|\theta\|^2$ becomes more dominant in the ridge objective function, forcing the parameters to be smaller. Asymptotically, the parameters will approach 0, leading to a constant validation error.
The left side is too steep to exhibit this trend. Hence the leftmost tick mark must be $\lambda = 0.1$.

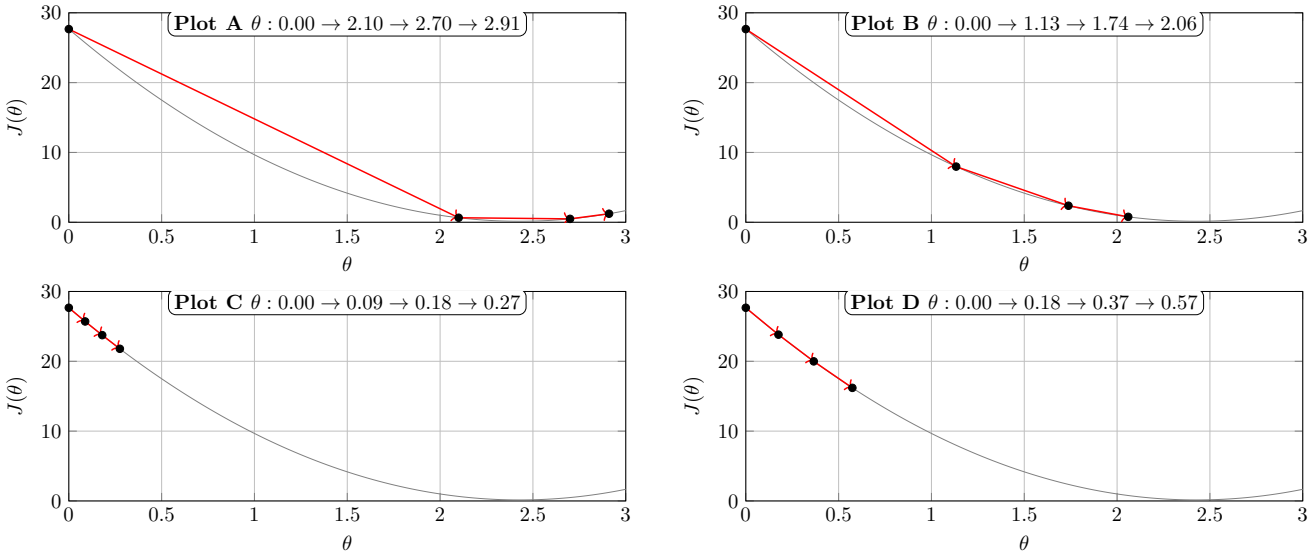
Gradient Descent

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.



Solution: ◯ Plot A ☑ Plot B ◯ Plot C ◯ Plot D

Brief explanation:
Plot A not possible. At 2.7 the gradient is positive, so it cannot move to a larger value of 2.91.
Plot B behaves as expected, with the trajectory progressing towards to global minimum and the steps decreasing in magnitude as the gradient itself is decreasing.
Plot C is not possible, as θ change is shown to be constant. However, since θ change is the product of the learning rate and the gradient magnitude, and since the gradient magnitude is linear in θ hence not a constant, θ change cannot be constant.
Plot D is also not possible. Between 0.00, 0.18, and 0.37, θ value change is increasing magnitude. By using the same argument as in Plot C, θ value change is the product of the learning rate and the gradient magnitude, and since the gradient magnitude at 0 is larger than the gradient magnitude at 0.18, θ value change needs should be smaller moving from 0.18 than from 0.00. Hence Plot D is not possible.

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

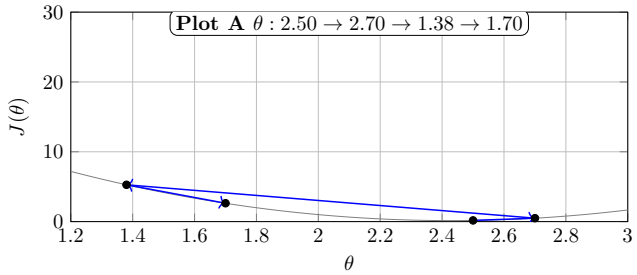
updated parameter value θ_{new}. What is the range of η such that $J(\theta_{\text{new}}) \leq J(\theta_{\text{old}})$? You can use the fact that the optimal θ^* is $\frac{17}{7}$ if needed.

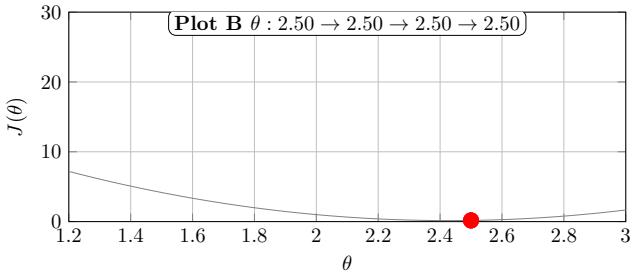
Solution:
Answer: $0 < \eta \leq \frac{3}{14}$
Solution: Given the MSE objective function: $J(\theta) = \frac{14}{3}\theta^2 - \frac{68}{3}\theta + \frac{83}{3}$
The gradient is: $\nabla J(\theta) = \frac{28}{3}\theta - \frac{68}{3}$
At $\theta_{\text{old}} = 2$: $\nabla J(2) = \frac{28}{3} \cdot 2 - \frac{68}{3} = \frac{-12}{3} = -4$
The update rule: $\theta_{\text{new}} = 2 - \eta \cdot (-4) = 2 + 4\eta$
For a quadratic function, the condition $J(\theta_{\text{new}}) \leq J(\theta_{\text{old}})$ means we can take a step that lands us anywhere between our current position and the point that's symmetric across the optimal point.
The optimal point is $\theta^* = \frac{17}{7}$. The symmetric point across $\theta = 2$ is at $2\theta^* - 2 = 2 \cdot \frac{17}{7} - 2 = \frac{34}{7} - \frac{14}{7} = \frac{20}{7}$.
Setting $\theta_{\text{new}} = \frac{20}{7}$: $\frac{20}{7} = 2 + 4\eta$
Solving: $\eta = \frac{\frac{20}{7} - 2}{4} = \frac{\frac{20}{7} - \frac{14}{7}}{4} = \frac{\frac{6}{7}}{4} = \frac{6}{28} = \frac{3}{14}$
Therefore: $0 < \eta \leq \frac{3}{14}$

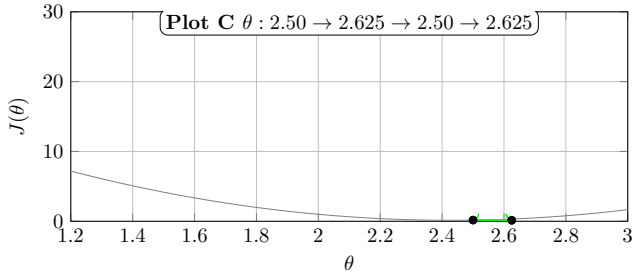
iii. (3 points) Now we run stochastic gradient descent (SGD) on the same dataset with initial parameter $\theta = 0$ and a constant learning rate $\eta > 0$.
After the first iteration of SGD, how many possible values are there for the resulting updated parameter value?

Solution: ◯ 1 ◯ 2 ☑ 3 ◯ Not enough info to determine.
As there are three possible data points, there are three possible results after one iteration of SGD.

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?







Solution:
Plot B and Plot C are possible.
For Plot B: We keep sampling the same second data point for 3 times.
For Plot C:

Setup:

We have training data

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

and a single-parameter model $h_\theta(x) = \theta x$. The squared loss on a point (x, y) is

$$\ell(\theta; x, y) = (\theta x - y)^2,$$

with gradient

$$\nabla_\theta \ell(\theta; x, y) = 2x(\theta x - y).$$

Thus the SGD update is

$$\theta \leftarrow \theta - \eta \cdot 2x(\theta x - y), \qquad \eta = 0.125.$$

Step 1: $\theta = 2.5,$ sample $(1,3):$ $\nabla = 2\cdot 1(2.5-3) = -1,$ $\theta' = 2.5 - 0.125(-1) = 2.625$

Step 2: $\theta = 2.625,$ sample $(2,5):$ $\nabla = 2\cdot 2(5.25-5) = 1,$ $\theta' = 2.625 - 0.125(1) = 2.5$

Step 3: $\theta = 2.5,$ sample $(1,3):$ $\nabla = 2\cdot 1(2.5-3) = -1,$ $\theta' = 2.5 - 0.125(-1) = 2.625$

Oscillation: $2.5 \to 2.625 \to 2.5 \to 2.625 \to \ldots$ by alternating between $(1,3)$ and $(2,5)$.

(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 $\theta = [1,1]^T$.

Solution:
Solution:
The ridge objective function is: $J(\theta) = \frac{1}{3}\|X\theta - Y\|^2 + \lambda\|\theta\|^2$
Expanding: $J(\theta) = \frac{1}{3}(X\theta - Y)^T(X\theta - Y) + \lambda\theta^T\theta$
Taking the gradient: $\nabla J(\theta) = \frac{1}{3} \cdot 2X^T(X\theta - Y) + 2\lambda\theta$
$\nabla J(\theta) = \frac{2}{3}X^T(X\theta - Y) + 2\lambda\theta$
At $\theta = [1,1]^T$:

$$X\theta = \begin{bmatrix} 3 & 0 \\ 0 & 3 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 6 \end{bmatrix}$$

$$X\theta - Y = \begin{bmatrix} 3 \\ 3 \\ 6 \end{bmatrix} - \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix} = \begin{bmatrix} 0 \\ -3 \\ -3 \end{bmatrix}$$

$$X^T(X\theta - Y) = \begin{bmatrix} 3 & 0 & 3 \\ 0 & 3 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ -3 \\ -3 \end{bmatrix} = \begin{bmatrix} -9 \\ -18 \end{bmatrix}$$

$$\nabla J(\theta) = \frac{2}{3}\begin{bmatrix} -9 \\ -18 \end{bmatrix} + 2 \cdot 1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -6 \\ -12 \end{bmatrix} + \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} -4 \\ -10 \end{bmatrix}$$

Answer: $\nabla J([1,1]^T) = \begin{bmatrix} -4 \\ -10 \end{bmatrix}$

(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 $\nabla_\theta J$?

Solution:
The correct answer is: $\nabla_\theta J = 2\theta + 2X^T(X\theta - Y) + 2(\theta^T X^T Y)X^T Y$
The answer is most striaghtforward by inspecting shapes and that J is quadratic in θ (so the gradient is linear in θ).
Below is the full derivation:
Derivation: $J(\theta) = \theta^T\theta + \|X\theta - Y\|^2 + \|\theta^T X^T Y\|^2$
Taking derivatives term by term: 1. $\frac{\partial}{\partial\theta}(\theta^T\theta) = 2\theta$ 2. $\frac{\partial}{\partial\theta}(\|X\theta - Y\|^2) = 2X^T(X\theta - Y)$ 3. $\frac{\partial}{\partial\theta}(\|\theta^T X^T Y\|^2) = \frac{\partial}{\partial\theta}((\theta^T X^T Y)^2) = 2(\theta^T X^T Y) \cdot X^T Y$
Therefore: $\nabla_\theta J = 2\theta + 2X^T(X\theta - Y) + 2(\theta^T X^T Y)X^T Y$

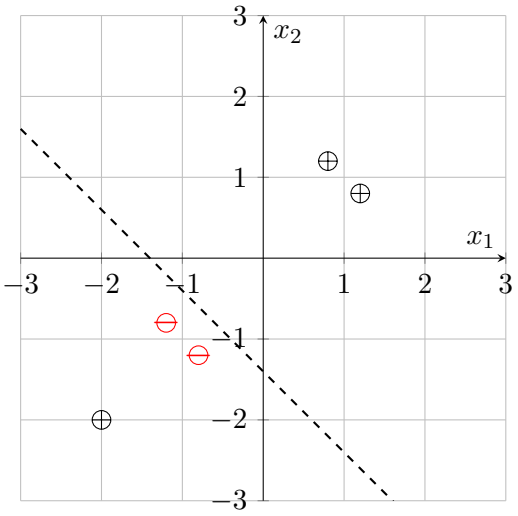
Linear Classification

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 $(0, -1.4)$.



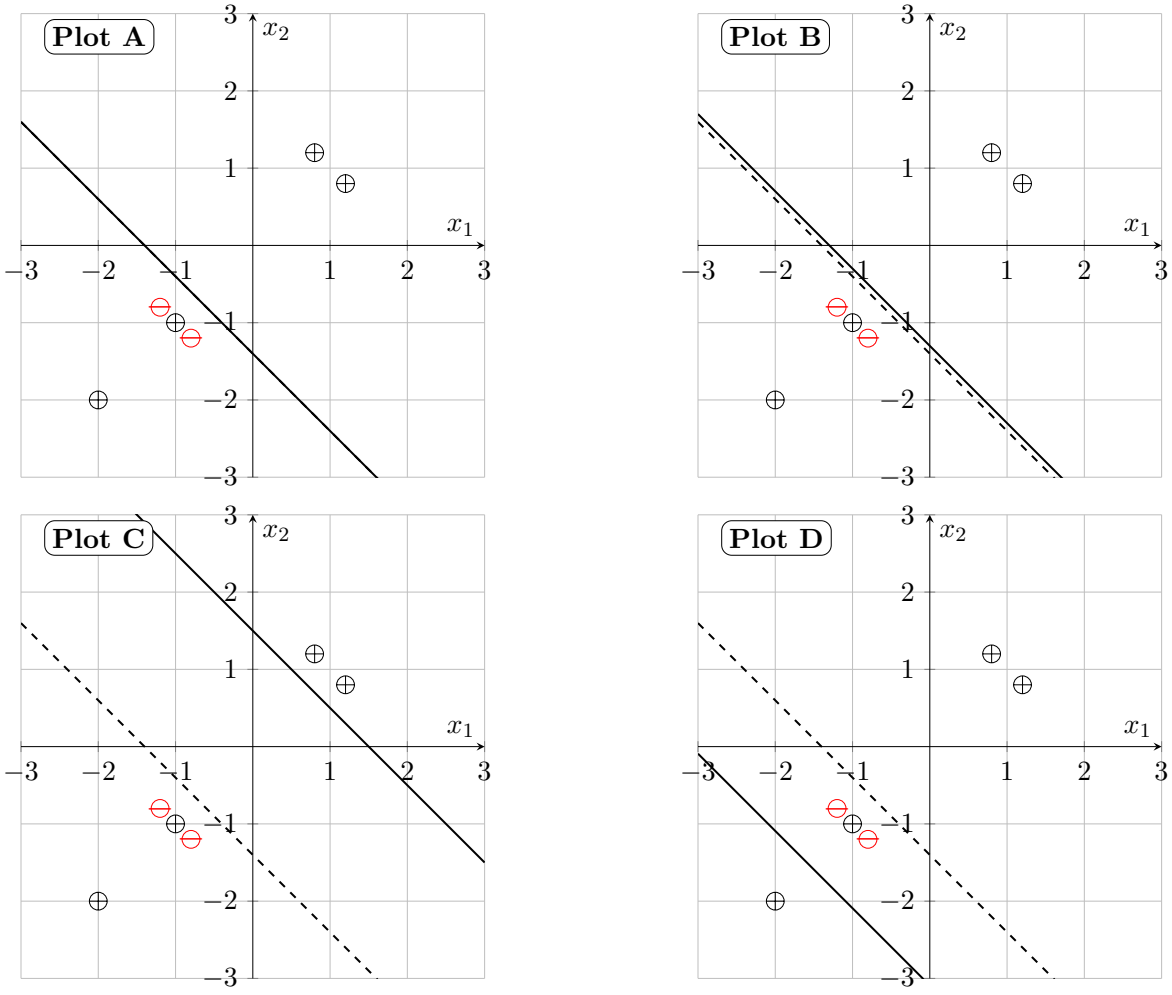
i. (4 points) Can we determine the normal vector direction? If yes, give the normal vector direction, as numerical values. If no, explain why not.

Solution:
☑ Yes ◯ No
Normal vector or why not possible to determine:
$\theta = c * [\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]^T$. for some constant $c > 0$.
Explanation: The normal vector will point in the upper-right direction. The decision boundary is further away from the data points that are the most certain, i.e., the two positively-label points in quadrant one.

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.

Solution:
◯ Yes ☑ No
We only know the direction of the normal vector but not the magnitude.
For example, check that $\theta = [1, 1]^T$ and $\theta_0 = -1.4$ gives the separator $x_2 = -x_1 - 1.4$. However, a positively scaled version of the parameters $\theta = [2, 2]^T$ and $\theta_0 = -2.8$ also gives the same separator $x_2 = -x_1 - 1.4$. Therefore, we cannot determine $h(x)$ at point $(-2, -2)$.

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.



Solution:
◯ Plot A ◯ Plot B ◯ Plot C ☑ Plot D
Brief explanation: Adding a positive data point at $(-1, -1)$ will pull the decision boundary toward this new point, meaning that the dashed line should be moving towards the new point.
This is because: right now, the new point is in the negative region according to the original boundary, meaning that $h(-1, -1) = \sigma(-1.0 \cdot \theta_1 - 1.0 \cdot \theta_2 + \theta_0) < \sigma(0)$ (Again, we do not know the magnitude of the exact values of $\theta_1, \theta_2, \theta_0$, but we know the direction).
In order to improve the NLL loss, we need to make $h(-1, -1)$ bigger. Only Plot D satisfies this condition.

(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?

Solution:
Predicted class for binary classifier: ◯ Positive class ☑ Negative class
Predicted class for 3-class classifier: ◯ Class 1 ◯ Class 2 ☑ Class 3

Brief explanation: For binary classifier: $h(-2) = \sigma(1 \cdot (-2) + (-1)) = \sigma(-3) < 0.5$, so negative class. For 3-class: logits $= [1, 2, 0] \cdot (-2) + [0, -1, 1] = [-2, -5, 1]$, so Class 3 has the highest logit value (1), so Class 3 is predicted.

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 $x = -2$?

Solution:
☑ Positive class ◯ Negative class

Brief explanation: $h(-2) = \sigma(1 \cdot (-2) + 9) = \sigma(7) > 0.5$, so positive class. Adding 10 to the offset shifts the decision boundary significantly.

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 $x = -2$?

Solution:
◯ Class 1 ◯ Class 2 ☑ Class 3

Brief explanation: New logits $= [1, 2, 0] \cdot (-2) + [10, 9, 11] = [8, 5, 11]$. Class 3 has the highest logit value (11), so Class 3 is still predicted. Adding the same constant to all offsets doesn't change the relative ordering of logits.

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.

Solution:
For binary classifier: We need $\sigma(x - 1)$ and $\sigma(x + 9)$ to give the same prediction (both < 0.5 or both > 0.5). This happens when $x - 1$ and $x + 9$ have the same sign, i.e., $(x - 1)(x + 9) \geq 0$. Solving: $x \leq -9$ or $x \geq 1$.
For 3-class classifier: Adding the same constant to all offsets doesn't change the relative ordering of logits, so the prediction is unchanged for all x.
Therefore, the range is $x \leq -9$ or $x \geq 1$ for binary classifier, and all x for 3-class classifier.

This is the end of the exam.