CARNEGIE MELLON UNIVERSITY 07-280

HOMEWORK 4

DUE: Thursday, Sep. 24, 2026

https://www.cs.cmu.edu/~07280/

INSTRUCTIONS

- **Format:** Use the provided LaTeX template to write your answers in the appropriate locations within the *.tex files and then compile a pdf for submission. We try to mark these areas with STUDENT SOLUTION HERE comments. Make sure that you don't change the size or location of any of the answer boxes and that your answers are within the dedicated regions for each question/part. If you do not follow this format, we may deduct points.

  You may also digitally annotate the pdf. Illegible handwriting will lead to lost points. However, we suggest that try to do at least some of your work directly in LaTeX.

- **How to submit written component:** Submit to Gradescope a pdf with your answers. Again, make sure your answer boxes are aligned with the original pdf template.

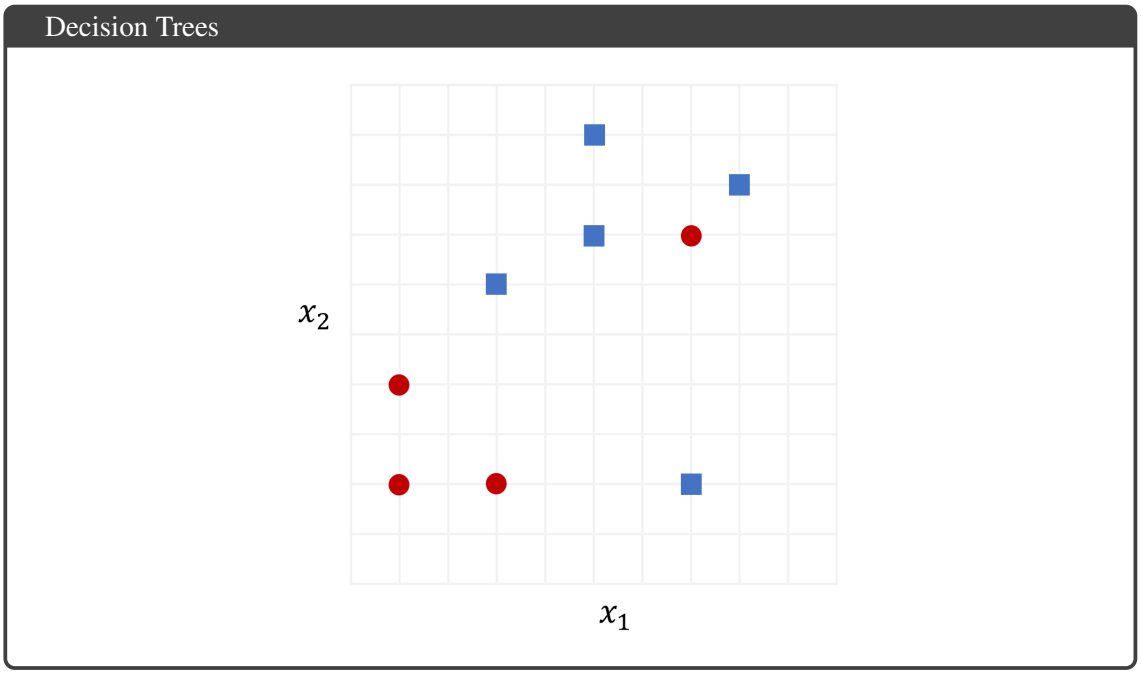
- **How to submit programming component:** See section Programming Submission for details on how to submit to the Gradescope autograder.

- **Policy:** See the course website for homework policies, including late policy, and academic integrity policies.

Name	
Andrew ID	
Hours to complete all components (nearest hour)	

1 [4 pts] Decision Boundaries

1. **[4 pts]** Draw a valid decision boundary that a decision tree may create in order to perfectly classify the training set. You do not need to calculate entropy for this question.



2 [24 pts] Linear Regression

We are given the following four-point data set, $\mathcal{D} = \left\{ (x_1^{(i)}, x_2^{(i)}, y^{(i)}) \right\}_{i=1}^{4}$:

x_1	x_2	y
1	-1	0
0	3	-1
-3	0	0
-1	3	2

We are going to use a linear model with no offset/bias term for our prediction function:

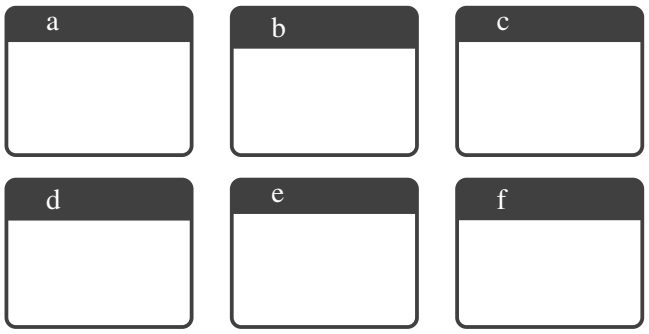
$$\hat{y} = w_1 x_1 + w_2 x_2$$

Our goal is to find the parameters that minimize a mean squared error objective function given our data:

$$J(w_1, w_2; \mathcal{D}) = \frac{1}{4} \sum_{i=1}^{4} \left(y^{(i)} - h\left(\mathbf{x}^{(i)} \right) \right)^2$$

1. **[12 pts]** Prove that this objective function can fit into the following form by computing values for a, b, c, d, e, f given our dataset. Showing your work in these questions is optional, but it is recommended to help us understand where any misconceptions may occur. Answer these questions in fraction form.

$$J(w_1, w_2; \mathcal{D}) = a w_1^2 + b w_2^2 + c w_1 w_2 + d w_1 + e w_2 + f$$



Work (optional)

2. **[4 pts]** What shape is function $J(w_1, w_2; \mathcal{D})$ when you plug in the values for a, b, c, d, e, f that you found above? We are looking for a one word answer for the shape. Be specific, e.g. don't say rectangle if it is in fact a square.

Your Answer

3. **[8 pts]** Write equations for the partial derivatives of our objective with respect to the parameters.

   To avoid potential error propagation from the previous part, write the derivatives in terms of symbols a, b, c, d, e, f rather than the values you computed above.

$\frac{\partial J}{\partial w_1}$

$\frac{\partial J}{\partial w_2}$

3 [16 pts] Linear Regression Data Set Creation

Recall that we define the objective function in linear regression as:

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - h\left(\mathbf{x}^{(i)}\right)\right)^2$$

In order to find a line that best fits our data, we want to minimize this objective function and find an optimal value, $J^*(\theta)$.

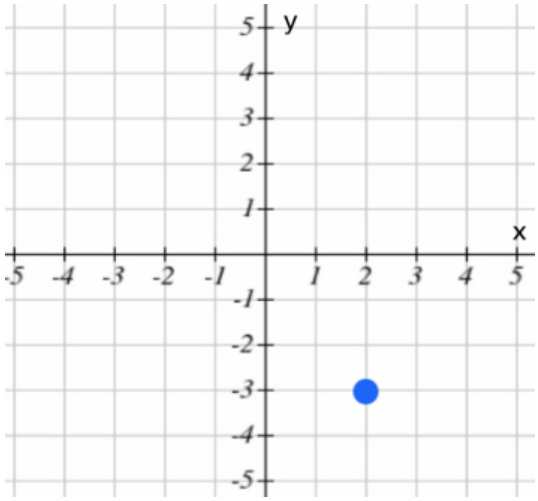
In this question, you will be asked to create data sets, $\mathcal{D} \subseteq \mathbb{Z}^2$, such that there are a specified number of solutions with an objective value equal to $J^*(\theta)$. For each subpart, if it is possible to create $\mathcal{D}$ such that the optimization problem has the specified number of solutions, draw the points in $\mathcal{D}$ on the left plot. If it is not, select "Not Possible".

Notes For all Sub-Parts:

- The data set that you create must have $J^*(\theta) > 0$.
- The entries you select must have integer coordinates.

1. **[4 pts]** Add exactly 2 more points, (x_1, y_1) and (x_2, y_2), to the plot below such that the set of points has **no solution** or identify that this is not possible.

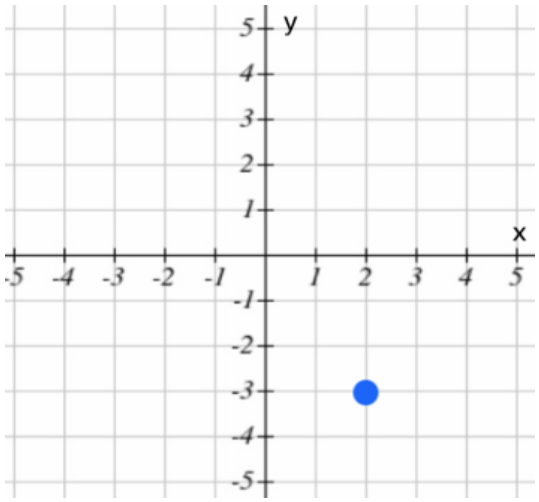
Your Answer



◯ Not Possible

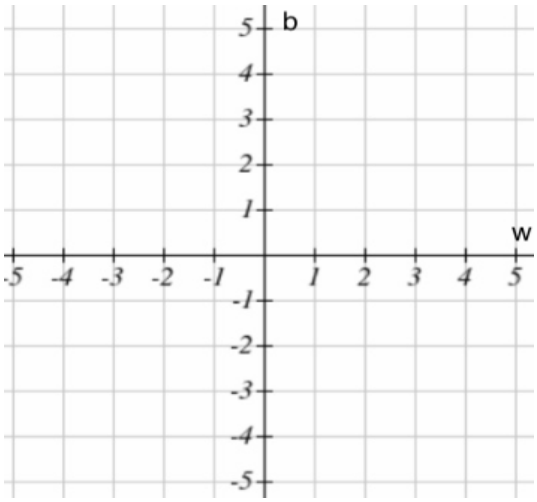
2. **[4 pts]** Add **exactly 2** more points, (x_1, y_1) and (x_2, y_2), to the plot below such that the set of points has **exactly one solution** or identify that this is not possible. If possible, draw the optimal solution as a line ($\mathbf{y = wx + b}$) on the left plot and draw the solution as a point, (w, b) on the right plot.

Your Answer



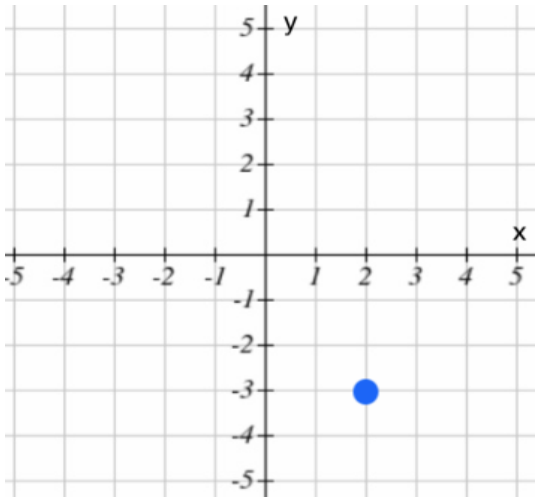
◯ Not Possible

Your Answer



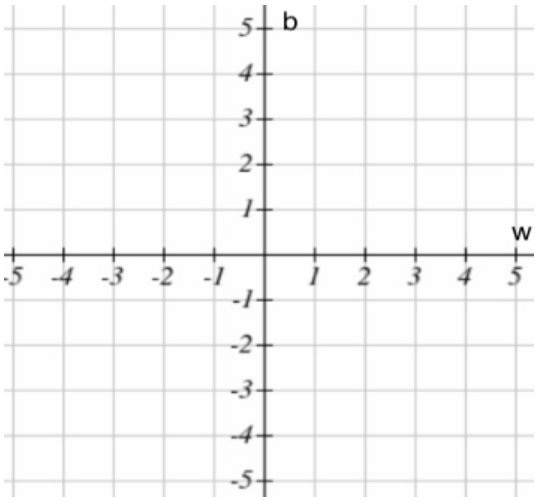
3. **[4 pts]** Add **exactly 2** more points, (x_1, y_1) and (x_2, y_2), to the plot below such that the set of points has **exactly two solutions** or identify that this is not possible. If possible, draw the two optimal solutions as lines ($\mathbf{y = wx+b}$) on the left plot and draw each solution as a point, (w_i, b_i), on the right plot.

Your Answer



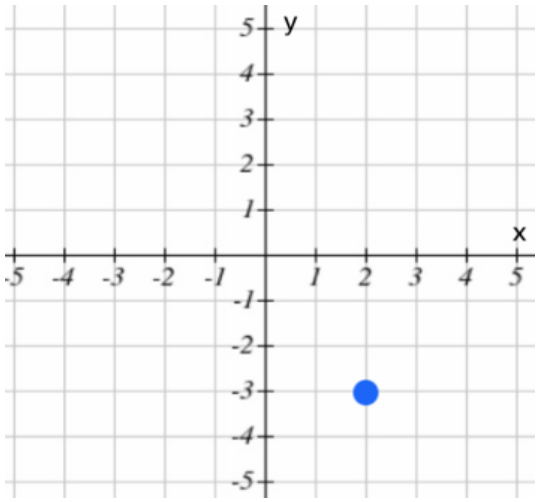
◯ Not Possible

Your Answer



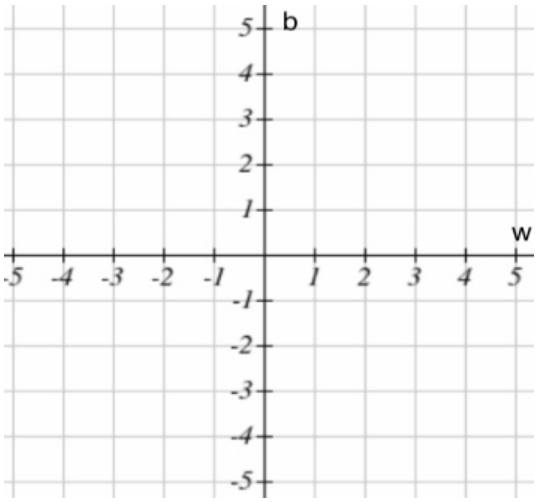
4. **[4 pts]** Add **exactly 2** more points, (x_1, y_1) and (x_2, y_2), on the plot below such that the set of points has **infinite solutions** or identify that this is not possible. If possible, draw 3 possible solutions as lines ($\mathbf{y = wx + b}$) on the left plot. Also, draw these three solutions as points, (w_i, b_i), on the plot on the right.

Your Answer



◯ Not Possible

Your Answer



4 [16 pts] Linear Regression with More Weights

In our original model for linear least squares regression, we made the assumption that each data point provided equally precise information in determining our estimate for the output. However, this may not always be the case. For example, if we were trying to predict net worth from age, we would probably see a greater variance as age increases. In situations like this, when it may not be reasonable to assume that every observation should be treated equally, weighted least squares can often be used to maximize the efficiency of parameter estimation. This is done by attempting to give each data point its proper amount of influence over the parameter estimates.

With this in mind, consider observing a data set $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ where $\mathbf{x}^{(i)} \in \mathbb{R}^M$ denotes the i-th example, and $y^{(i)}$ denotes the target response value. Assume that each data point $(\mathbf{x}^{(i)}, y^{(i)})$ comes with a weighting factor $r_i > 0$, in which case our error function becomes:

$$J_{\mathcal{D}}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} r_i \left(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)} \right)^2$$

We can also write this with linear algebra notation without the summation, in which case we get:

$$J_{\mathcal{D}}(\mathbf{w}) = \frac{1}{N} (\mathbf{y} - X\mathbf{w})^\top R (\mathbf{y} - X\mathbf{w})$$

Here, $\mathbf{w}$ is the weight matrix denoting the model parameters for our linear model, R is a diagonal matrix where $R_{i,i} = r_i$, and X is the design matrix where the i-th row contains the features of $\mathbf{x}^{(i)}$.

In this question, we will derive the solution $\mathbf{w}^*$ that minimizes the weighted sum of squares error defined above. Throughout this problem, you must show your work. Reasonable steps should be taken between lines in the derivation. Providing an explanation per line is optional.

1. **[10 pts]** Derive the derivative of the objective function with respect to $\mathbf{w}$.

   Do NOT use summations, and instead use a linear algebra representation. Also, make sure to define the dimensions of any variables you use.

   To receive full credit, you must show your work.

Your Answer

2. **[6 pts]** Set the derivative you calculated in the previous part equal to zero and solve for $\mathbf{w}^*$.

Your Answer

5 [16 pts] Matrix Calculus

Given $C = f(A, B) = AB$ with $A \in \mathbb{R}^{M \times K}$ and $B \in \mathbb{R}^{K \times N}$, derive the partial derivatives below.

Hint: Write the equation for $c_{m,n}$.

Note: It can be helpful to work this out on scratch paper for matrix multiplication with very small example matrices.

1. **[4 pts]** Derive $\partial c_{m,n} / \partial a_{m,k}$.

Your Answer

2. **[4 pts]** Derive $\partial c_{m,n} / \partial a_{i,k}$, where $m \neq i$

Your Answer

3. **[4 pts]** Derive $\partial c_{m,n} / \partial b_{k,n}$.

Your Answer

4. **[4 pts]** Derive $\partial c_{m,n}/\partial b_{k,j}$, where $n \neq j$

Your Answer

6 [24 pts] More Matrix Calculus

Let $y = f(A) = \|A\|_F^2$ where $y \in \mathbb{R}$ and $A \in \mathbb{R}^{N \times K}$. Note that this is the Frobenius norm squared.

Let $Z = g(\Theta) = X\Theta$ and $t = J(\Theta) = f(g(\Theta))$ where $X \in \mathbb{R}^{N \times M}$, $\Theta \in \mathbb{R}^{M \times K}$, and $Z \in \mathbb{R}^{N \times K}$

Note: In this question, we'll be using denominator form.

1. **[4 pts]** Derive $\partial y / \partial a_{i,j}$. Specifically, write this partial derivative in terms of the elements of A.

   Your Answer

2. **[4 pts]** Based on patterns that you can recognize from the previous part, write the $\partial y / \partial A$ in denominator form. Specifically, write this partial derivative in terms of A.

   Your Answer

3. **[4 pts]** What are the dimensions of the denominator form of $\partial t / \partial \Theta$?

   Your Answer

4. **[12 pts]** In this next problem, we'll take steps to prove the following (denominator form) partial derivative:

   $$\frac{\partial}{\partial \Theta} \|X\Theta\|_F^2 = 2X^\top X\Theta$$

   Using our variables t and Z, we can write this as:

   $$\frac{\partial t}{\partial \Theta} = 2X^\top Z = 2X^\top X\Theta$$

   To simplify the notation in the proof a bit, let's assume $N = M = K = 2$, and $X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$, $\Theta = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix}$, and $Z = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix}$.

Prove that $\partial t / \partial \theta_{1,1}$ equals the 1,1 element of $2X^\top X\Theta$.

To get full credit, you must structure your proof as follows:

(a) Expand $J(\Theta)$, writing it in terms of the elements of X and Θ. Note: To save space/time, you may omit terms that will not affect the proof of $\partial t / \partial \theta_{1,1}$.

(b) Write $\partial t / \partial \theta_{1,1}$ in terms of the elements of X and Θ.

(c) Expand $2X^\top X\Theta$ as necessary to complete the proof.

Your Answer

Additional room from the previous page if needed

Note, this problem is our attempt to help you gain deeper understanding of where the extra X comes from when solving linear regression where there are multiple output values, $\mathbf{y}^i \in \mathbb{R}^K$:

$$J(\Theta) = \frac{1}{N}\|Y - X\Theta\|_F^2$$
$$\frac{\partial J}{\partial \Theta} = -\frac{2}{N}X^\top(Y - X\Theta)$$

7 Collaboration Policy

After you have completed all other components of this assignment, report your answers to the following collaboration questions.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details including names of people who helped you and the exact nature of help you received.

Your Answer

2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details including names of people you helped and the exact nature of help you offered.

Your Answer

3. Did you find or come across code that implements any part of this assignment? If so, include full details including the source of the code and how you used it in the assignment.

Your Answer