Midterm Exam, Fall 2026
ESE 577
SOLUTIONS

Instructor: Jorge Mendez-Mendez

Date: Thursday October 8th, 2026

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, is 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	26.5
2	27
3	22.5
4	24
Total:	100

Name: _____

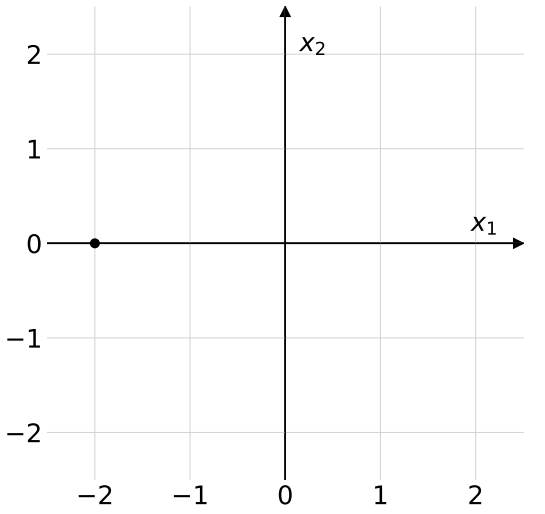
Classification

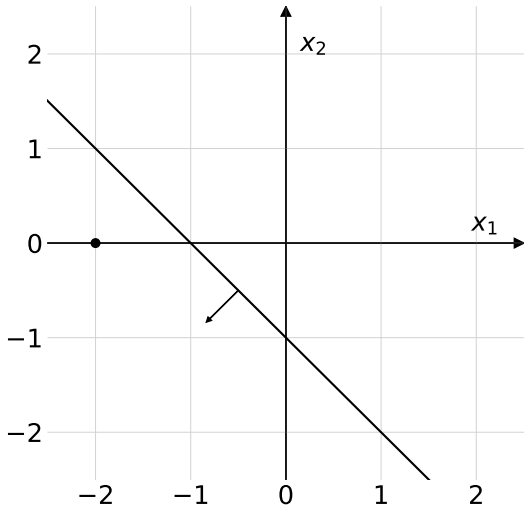
1. (26.5 points) Recall that a linear logistic classifier is characterized by

$$h(x;\theta,\theta_0) = \sigma(\theta^\top x + \theta_0) \ ,$$

where $\sigma(\cdot)$ is the standard sigmoid function, $\sigma(z) = 1/(1+\exp(-z))$. Define the argument of the sigmoid function in $h(\cdot)$ to be $z = \theta^\top x + \theta_0$.

(a) i. On the graph below, draw the linear separator defined by the parameters $\theta = [-3,-3]^\top, \theta_0 = -3$. Be sure to include a normal vector pointing in the direction of the positive class.





ANSWER:

ii. Would the corresponding linear logistic classifier assign the indicated point at $(x_1, x_2) = (-2.0, 0.0)$ as positive or negative?

Name: _____

Circle one: Positive Negative

ANSWER: The indicated point is classified as positive.

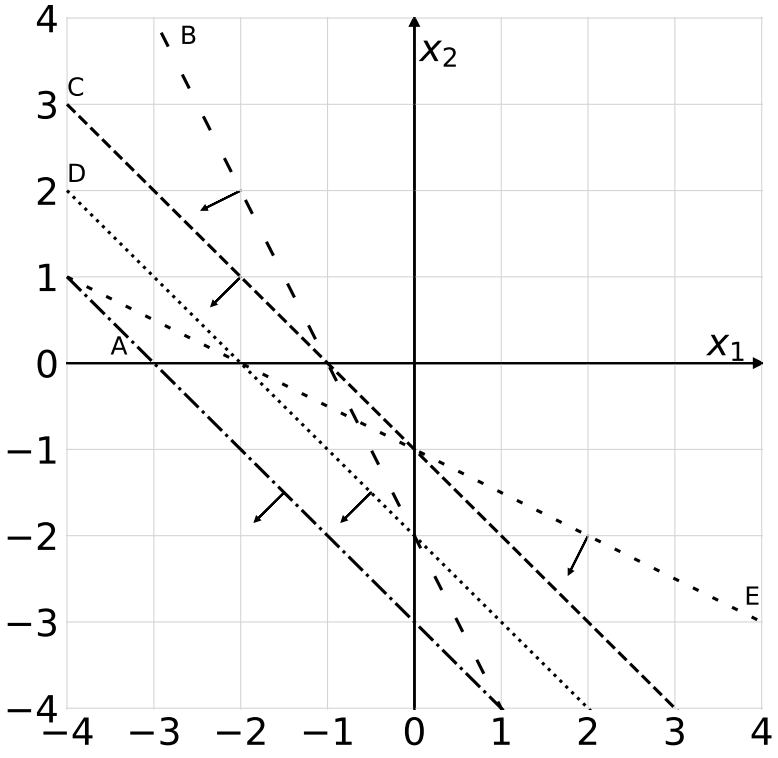
iii. What value of z does the classifier assign to the indicated point? (Your answer should be a numerical integer value.)

$z =$
ANSWER: $z = -3(-2) - 3(0) - 3 = 3$.

iv. What is the numerical probability output by the classifier? (You can leave the mathematical expression involving e, but your solution must not have matrix operations left unsolved.)

$\sigma(z) =$
ANSWER: $\sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-3}}$.

(b) The following plot represents a two-dimensional space into which five separating hyperplanes for classifiers have been drawn, each with an associated normal vector intended to point toward the positive examples.



For each of the hypotheses parameterized by the values (θ, θ_0) below, identify the matching hyperplane from the plot above.

Name: _____

i. $\theta = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $\theta_0 = -2$

Circle one: A B C D E

ANSWER: The correct answer is D.

ii. $\theta = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $\theta_0 = -3$

Circle one: A B C D E

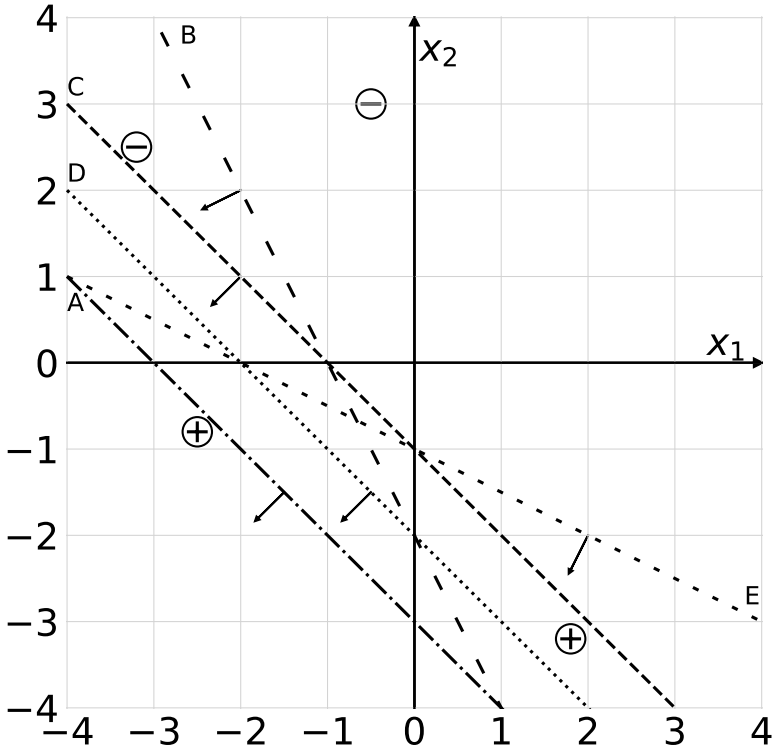
ANSWER: The correct answer is A.

iii. $\theta = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$, $\theta_0 = -2$

Circle one: A B C D E

ANSWER: The correct answer is E.

(c) The plot below is the same as in the previous part, but now with some "held-out" data points that were not used in training the models. Each data point is labeled as positive (+) or negative (−).



Imagine the linear logistic classifier model with parameter $|\theta_1| = 1$ that corresponds to each of

the drawn separators.

i. What is the accuracy of each of the models (A–E) on the held-out dataset shown in the plot?

Accuracy of A:

Accuracy of B:

Accuracy of C:

Accuracy of D:

Accuracy of E:

ANSWER:
Accuracy of A: 75%
Accuracy of B: 50%
Accuracy of C: 100%
Accuracy of D: 75%
Accuracy of E: 100%

ii. Using the NLL as the loss, which model has the minimum loss on the held-out data? Justify your response. Hint: you do not need to explicitly compute the NLL value.

Circle one: A B C D E

Justification:

ANSWER: E. Both C and E achieve 100% accuracy and should have lower loss than models with lower accuracy. Since the parameters are in the same scale (we are told that $|\theta_1| = 1$), the distance from the separator contributes similarly to each model's NLL. Comparing models C and E, two data points are very close to the separator for classifier C and contribute high loss. Instead, for model E, all points are relatively far.

(d) Now we would like to analyze a multiclass linear logistic classifier with $K = 3$ classes ('1', '2', and '3') in a feature space with 2 dimensions. Our θ will be a 2×3 matrix and our θ_0 will be a 1×3 matrix. With that, we will obtain:

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \theta^\top \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \theta_0^\top \ ,$$

and the output of the model will be:

$$g = \text{softmax}(z) = \begin{bmatrix} \exp(z_1)/\sum_{i=1}^3 \exp(z_i) \\ \exp(z_2)/\sum_{i=1}^3 \exp(z_i) \\ \exp(z_3)/\sum_{i=1}^3 \exp(z_i) \end{bmatrix} \ .$$

Recall that the vector g represents the likelihood assigned to each of the three classes, and that the predicted class is the one corresponding to the largest element of g.

Name: _____

i. Suppose that we have a model defined by the following parameters:

$$\theta = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad , \theta_0 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \quad .$$

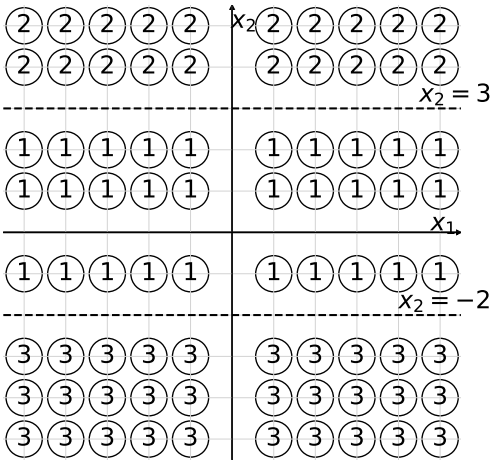
Consider the data point $x = [-1, 1]^\top$. Compute $z = \theta^\top x + \theta_0^\top$ and determine which class will be assigned to x.

$z =$

Class assigned:

ANSWER: $z = [2, -1, 0]^\top$ and so the assigned class will be '1'.

ii. Examine the classification problem represented by the plot below, where points are labeled with their class: '1', '2', or '3'. Also given below is a model with parameters θ, θ_0.



$$\theta = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix} \quad , \theta_0 = \begin{bmatrix} 0 & -3 & -2 \end{bmatrix}$$

Does the provided model given by θ, θ_0 perfectly classify the data in the plot? If yes, show your reasoning. If not, identify all the data points that are misclassified.

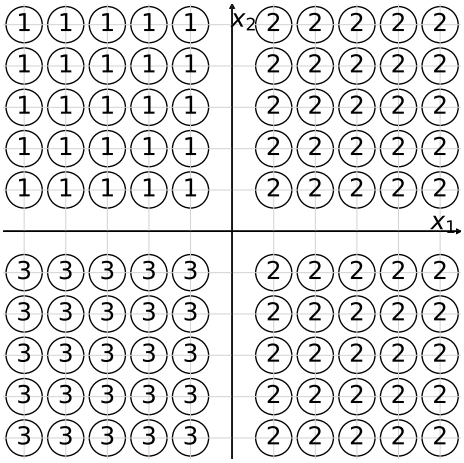
Name: _____

Circle one: Yes No

Explanation:

ANSWER: Yes. We have $z = [0, x_2 - 3, -x_2 - 2]^\top$. This implies that we get a label of '2' when $x_2 > 3$, a label of '3' when $x_2 < -2$, and a label of '1' in-between.

iii. Examine the classification problem represented by the plot below, where points are labeled with their class: '1', '2', or '3'. Also given below is a model with parameters θ, θ_0.



$$\theta = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}, \theta_0 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

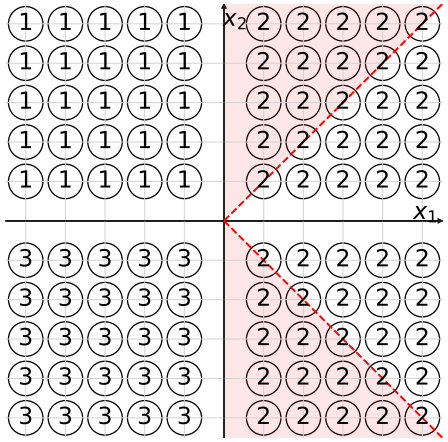
Does the provided model given by θ, θ_0 perfectly classify the data in the plot? If yes, show your reasoning. If not, identify all the data points that are misclassified.

Name: _____

Circle one: Yes No

Explanation:

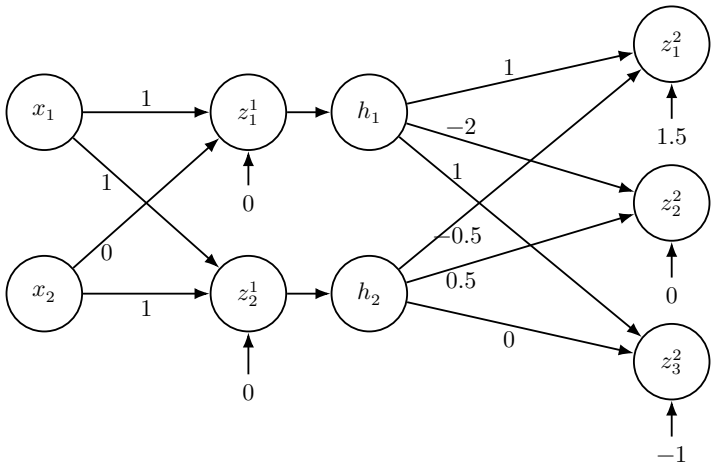
ANSWER: No. We have $z = [x_2, x_1, -x_2]^\top$. This implies that the model predicts '1' when $x_2 > 0$ and $x_1 < 0$ or $x_2 > x_1$ and $x_1 > 0$; the model predicts '3' when $x2 < 0$ and $x_1 < 0$ or $x_2 < -x_1$ and $x_1 > 0$. The misclassified regions are shown in the figure below.



Name: _____

All-the-Way Backprop

2. (27 points) The diagram below shows a fully-connected neural network with one hidden layer using ReLU activation and one output layer with softmax activation for 3-class classification.



We may also write the network's computation algebraically as:

$$h = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \text{ReLU}\left(W_1 \cdot x\right) = \text{ReLU}\left(\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$$

$$z^2 = \begin{bmatrix} z_1^2 \\ z_2^2 \\ z_3^2 \end{bmatrix} = W_2 \cdot h + b_2 = \begin{bmatrix} 1 & -0.5 \\ -2 & 0.5 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} + \begin{bmatrix} 1.5 \\ 0 \\ -1 \end{bmatrix}$$

$$g = \text{softmax}(z^2) = \begin{bmatrix} \exp(z_1^2)/\sum_{i=1}^3 \exp(z_i^2) \\ \exp(z_2^2)/\sum_{i=1}^3 \exp(z_i^2) \\ \exp(z_3^2)/\sum_{i=1}^3 \exp(z_i^2) \end{bmatrix}$$

Throughout all parts of this question, assume that the input is $x = [0.5, 1.5]^\top$, unless otherwise noted.

(a) We will now perform forward propagation.

i. Compute the value of the hidden activations h.

$h =$

ANSWER:
$h = \text{ReLU}([0.5, 0.5 + 1.5]^\top) = [0.5, 2]^\top$

ii. Compute the value of the final layer's pre-activations, z^2.

$z^2 =$

ANSWER:
$z^2 = [0.5 - 1 + 1.5, -1 + 1 + 0, 0.5 + 0 - 1]^\top = [1, 0, -0.5]^\top$

iii. What class is predicted by the network for this data point? You do not need to compute the softmax values.

Circle one: 1 2 3

ANSWER:
The predicted label is '1'.

(b) Now, we will compute backpropagation to measure how changes in the *input* x modify the pre-activation (or logit) for class '2'. In particular, we will compute the gradient of z_2^2 with respect to x via backpropagation.

i. Compute the gradient of z_2^2 with respect to h. *Hint:* think about which entries in W_2 impact the computation of z_2^2.

$\nabla_h z_2^2 =$
ANSWER: $\nabla_h z_2^2 = W_2[2]^\top = [-2, 0.5]^\top$

ii. Compute the gradient of z_2^2 with respect to z^1. Your answer should be a numerical vector. *Hint:* use the chain rule and use $\frac{\partial \text{ReLU(w)}}{\partial w} = \begin{cases} 1 & \text{if } w \geq 0 \\ 0 & \text{otherwise} \end{cases}$

$\nabla_{z^1} z_2^2 =$
ANSWER: $\nabla_{z^1} z_2^2 = \nabla_{z^1} h \odot \nabla_h z_2^2 = [1, 1]^\top \odot [-2, 0.5]^\top = [-2, 0.5]^\top$.

iii. Compute the gradient of z_2^2 with respect to x using the chain rule. *Hint:* For $v = Ay$ and scalar $f(v)$, $\nabla_y f = A^\top \nabla_v f$.

$\nabla_x z_2^2 =$
ANSWER: $\nabla_x z_2^2 = \nabla_x z^1 \cdot \nabla_{z^1} z_2^2 = W_1^\top \cdot [-2, 0.5]^\top = [-1.5, 0.5]^\top$

(c) We will now use gradient ascent on the *input* x to nudge our network toward predicting class '2'. We will maximize the predicted z_2^2 while holding all network weights and biases fixed.

i. Perform one step of gradient *ascent* using a step-size of $\eta = 1$. What is the resulting input x'?

$x' =$
ANSWER: $x' = x + 1 \cdot \nabla_x z_2^2 = [0.5, 1.5]^\top + [-1.5, 0.5]^\top = [-1, 2]^\top$.

ii. Repeat the forward propagation, backpropagation process above for one step: compute the pre-activation output for the updated input x', compute the gradient of z_2^2 with respect to x', and perform one step of gradient ascent to obtain a new input x''. Was the predicted class for x' changed? What was the new value of x''? Show your work.

Name: _____

What was the predicted label for x'? Circle one: 1 2 3

$x'' =$

ANSWER:

$$h = \text{ReLU}\left(\begin{bmatrix}-1\\1\end{bmatrix}\right) = \begin{bmatrix}0\\1\end{bmatrix}$$

$$z^2 = \begin{bmatrix}-0.5+1.5\\0.5+0\\-1\end{bmatrix} = \begin{bmatrix}1\\0.5\\-1\end{bmatrix}$$

That gives us our predicted class '1'.
Then for backpropagation we have:

$$\begin{aligned}\nabla_x z_2^2 &= \nabla_x z^1 \cdot (\nabla_{z^1} h \odot \nabla_h z_2^2)\\ &= W_1^\top \cdot (\begin{bmatrix}0\\1\end{bmatrix} \odot W_2[2]^\top)\\ &= \begin{bmatrix}1&1\\0&1\end{bmatrix} \cdot \begin{bmatrix}0\\0.5\end{bmatrix}\\ &= \begin{bmatrix}0.5\\0.5\end{bmatrix}\end{aligned}$$

And with that, our updated $x'' = [-0.5, 2.5]^\top$.

iii. Repeat the forward propagation for another step: compute the pre-activation output for the newly updated input x''. Was the predicted class for x'' changed? Show your work.

What was the predicted label for x''? Circle one: 1 2 3

ANSWER:

$$h = \text{ReLU}\left(\begin{bmatrix}-0.5\\2\end{bmatrix}\right) = \begin{bmatrix}0\\2\end{bmatrix}$$

$$z^2 = \begin{bmatrix}-1+1.5\\1\\-1\end{bmatrix} = \begin{bmatrix}0.5\\1\\-1\end{bmatrix}$$

That gives us our predicted class '2', showing that we changed the network's prediction by backpropagating on the input!

Name: _____

Buy 'n' Large

3. (22.5 points) We have been hired by Buy 'n' Large to deliver a predictor of change in sales volume from last year, for each of their stores. We have a machine-learning algorithm that can be used with regularization parameter λ. Our overall objective is to deliver a predictor that minimizes squared loss on predictions when actually used by the company. There are three relevant data sets: training dataset $\mathcal{D}_{\text{train}}$, validation dataset $\mathcal{D}_{\text{val}}$, and test dataset $\mathcal{D}_{\text{test}}$, each of size n. You have $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$ is owned by the company and not shared with you.

   We will use a linear predictor with parameters θ (without the offset parameter θ_0 for simplicity) and use regularizer $\lambda\|\theta\|^2$. There are several phases of the training process:

   1) First, we find the optimal hypothesis for any hyperparameter.
   2) Then, we find the optimal hyperparameter.
   3) Then, we find the best hypothesis to send to Buy 'n' Large.
   4) Lastly, Buy 'n' Large evaluates the hypothesis we have sent to them.

   For each of these phases, we need to select the appropriate objective for each of those tasks. Each question provides the template of an answer, and asks you to fill in choices for different "slots" in the template that specify:

   - A — what the minimization is over,
   - B — the dataset used,
   - C — the predictor used, and
   - D — whether regularization is added.

   Fill in the slots (A, B, C, D) for each phase by choosing the expressions for the indicated slots.

   Note that θ and λ are variables that range over d-dimensional column vectors and positive reals, respectively. $\theta_{\text{best}}(\lambda)$ is a value of θ that is a function of λ. θ^*, λ^* are specific values found as described below.

   (a) Training — Selecting the best hypothesis (parameters θ) for some fixed value of the regularization parameter λ. Call this $\theta_{\text{best}}(\lambda)$.

$$\theta_{\text{best}}(\lambda) = \arg\min_{A} \frac{1}{|B|} \sum_{(x,y)\in B} (C^\top x - y)^2/2 + D \ .$$

   i. What should we minimize over? This is denoted by the variable A.

   Circle one: θ $\quad$ $\theta_{\text{best}}(\lambda)$ $\quad$ θ^*
   ANSWER: θ. To obtain our value of $\theta_{\text{best}}(\lambda)$ for a given λ, we need to find the value of θ that minimizes the loss function while λ is fixed. This means that the parameter that we will be using for arg min is θ.

   ii. What dataset should we use? This is denoted by the variable B.

   Circle one: $\mathcal{D}_{\text{train}}$ $\quad$ $\mathcal{D}_{\text{val}}$ $\quad$ $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$ $\quad$ $\mathcal{D}_{\text{test}}$
   ANSWER: $\mathcal{D}_{\text{train}}$. We are trying to find the best θ for a given λ so that we can later compare optimal models with each other using other data, so to obtain this θ and in order to have unseen data to use later, we must use $\mathcal{D}_{\text{train}}$ to minimize loss.

iii. What predictor should we use? This is denoted by the variable C.

Circle one: θ $\theta_{\text{best}}(\lambda)$ θ^*
ANSWER: θ. We are optimizing over values of θ in order to minimize the loss function, so we must use θ inside of our loss function.

iv. What regularizer should we use? This is denoted by the variable D.

Circle one: 0 λ $\lambda\|\theta\|^2$
ANSWER: $\lambda\|\theta\|^2$. Lastly, we need to regularize with the magnitude of θ weighted by the regularization hyperparameter λ, so our regularization term is $\lambda\|\theta\|^2$. Note that we cannot use λ alone as it is a constant and will not affect the search for the parameters θ.

(b) Hyperparameter selection — Selecting the best value of the regularization hyperparameter λ. We will call this best value λ^*.

$$\lambda^* = \arg\min_A \frac{1}{|B|} \sum_{(x,y)\in B} (C^\top x - y)^2/2 + D \ .$$

i. What should we minimize over? This is denoted by the variable A.

Circle one: λ λ^* $\lambda\|\theta\|^2$ $\lambda^*\|\theta\|^2$
ANSWER: λ. In the previous part of this problem, we were working on finding an optimal model θ for a given hyperparameter λ. In this part, we are trying to optimize our hyperparameter given multiple models, each one corresponding to an optimal model found with a value of λ we are considering. Therefore, to obtain our value of λ^*, we are optimizing the loss over values of λ. This means that the parameter that we will be using for $\arg\min$ is λ.

ii. What dataset should we use? This is denoted by the variable B.

Circle one: $\mathcal{D}_{\text{train}}$ $\mathcal{D}_{\text{val}}$ $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$ $\mathcal{D}_{\text{test}}$
ANSWER: $\mathcal{D}_{\text{val}}$. To determine the best value of λ given an optimal model, we will need to observe the performance of the model with unseen data. Thus, it is necessary to use $\mathcal{D}_{\text{val}}$ as the data with which to minimize loss.

iii. What predictor should we use? This is denoted by the variable C.

Circle one: θ $\theta_{\text{best}}(\lambda)$ θ^*
ANSWER: $\theta_{\text{best}}(\lambda)$. As noted before, we are trying to optimize the λ hyperparameter based on optimal models found with each one, so the model used while minimizing loss must be $\theta_{\text{best}}(\lambda)$.

iv. What regularizer should we use? This is denoted by the variable D.

Circle one: 0 λ λ^* $\lambda\|\theta\|^2$ $\lambda^*\|\theta\|^2$
ANSWER: 0. Lastly, there is no need to regularize with anything as no training is being done (we are only comparing performance—MSE—on the validation data).

(c) Shipping a predictor — Selecting the hypothesis (parameters θ) to deliver to the company. Call

this θ^*.

$$\theta^* = \arg\min_A \frac{1}{|B|} \sum_{(x,y)\in B} (C^\top x - y)^2/2 + D \ .$$

i. What should we minimize over? This is denoted by the variable A.

Circle one: θ $\theta_{\text{best}}(\lambda)$ θ^*
ANSWER: θ. To find the value θ^* that minimizes our objective function, we want to search over all θ values.

ii. What dataset should we use? This is denoted by the variable B.

Circle one: $\mathcal{D}_{\text{train}}$ $\mathcal{D}_{\text{val}}$ $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$ $\mathcal{D}_{\text{test}}$
ANSWER: $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$. Before delivering the regressor to the company, we'd like to train on all the data which we have and which represents the underlying distribution that we are trying to capture. $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$ allows us to use all the data we have.
You may argue that $\mathcal{D}_{\text{train}}$ is the actual training data that was used to find the $\theta_{\text{best}}(\lambda)$ values in the first place. But technically, having more data (i.e., including $\mathcal{D}_{\text{val}}$) should not cause problems—-it should only help the hypothesis get better.

iii. What predictor should we use? This is denoted by the variable C.

Circle one: θ $\theta_{\text{best}}(\lambda)$ θ^*
ANSWER: θ. We are optimizing over values of θ in order to minimize the loss function, so we must use θ.

iv. What regularizer should we use? This is denoted by the variable D.

Circle one: 0 λ λ^* $\lambda\|\theta\|^2$ $\lambda^*\|\theta\|^2$
ANSWER: $\lambda^*\|\theta\|^2$. In the regularization term, we should use our best regularization parameter λ^*, which we found in the previous step.

(d) Evaluation in practice — Evaluating the actual on-the-job performance ϵ^* of the selected hypothesis from validation.

$$\epsilon^* = \frac{1}{|B|} \sum_{(x,y)\in B} (C^\top x - y)^2/2 + D \ .$$

i. What dataset should we use? This is denoted by the variable B.

Circle one: $\mathcal{D}_{\text{train}}$ $\mathcal{D}_{\text{val}}$ $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$ $\mathcal{D}_{\text{test}}$
ANSWER: $\mathcal{D}_{\text{test}}$. To evaluate actual performance, we must run our model on real deployment data, or $\mathcal{D}_{\text{test}}$.

ii. What predictor should we use? This is denoted by the variable C.

Circle one: θ $\theta_{\text{best}}(\lambda)$ θ^*
ANSWER: θ^*. We want to use our selected hypothesis, θ^*, for our predictor.

iii. What regularizer should we use? This is denoted by the variable D.

Circle one: 0 λ λ^* $\lambda\|\theta\|^2$ $\lambda^*\|\theta\|^2$
ANSWER: 0. Because we are performing evaluation rather than training, we have no need for regularization and want only the MSE loss, so D is 0.

Name: _____

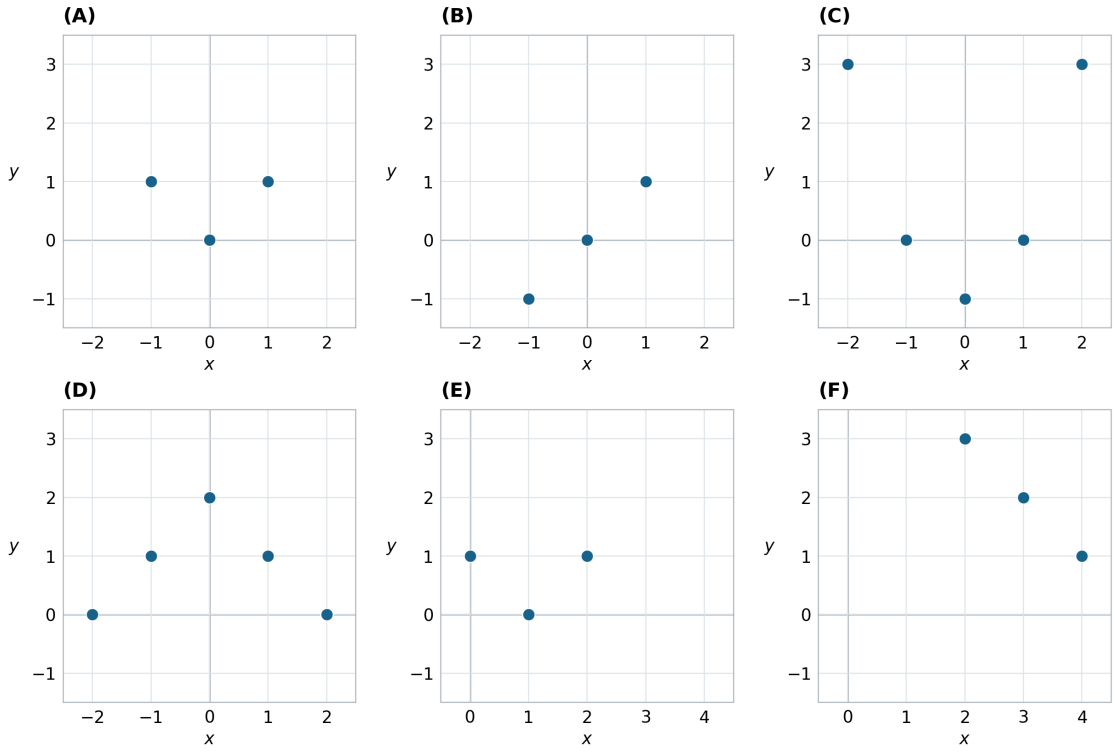
Regression features

4. (24 points) Feature transformations allow us to empower linear regression models with the ability to make nonlinear predictions on input features. For this problem, we will consider feature transformations from 1D to 1D $\phi : \mathbb{R} \mapsto \mathbb{R}$. We will deal with linear regressors of the following form:

$$f(x) = \theta \cdot \phi(x) + \theta_0 \ .$$

Each scatter plot below illustrates a dataset of 1D features and corresponding real-valued labels y. For each dataset, circle all the transformations that can fit the data exactly (i.e., achieve 0 training error).

Name: _____



(a) Plot A: $\phi(x) = x$ $\phi(x) = |x|$ $\phi(x) = x^2$ $\phi(x) = |x - 1|$

(b) Plot B: $\phi(x) = x$ $\phi(x) = |x|$ $\phi(x) = x^2$ $\phi(x) = |x - 1|$

(c) Plot C: $\phi(x) = x$ $\phi(x) = |x|$ $\phi(x) = x^2$ $\phi(x) = |x - 1|$

(d) Plot D: $\phi(x) = x$ $\phi(x) = |x|$ $\phi(x) = x^2$ $\phi(x) = |x - 1|$

(e) Plot E: $\phi(x) = x$ $\phi(x) = |x|$ $\phi(x) = x^2$ $\phi(x) = |x - 1|$

(f) Plot F: $\phi(x) = x$ $\phi(x) = |x|$ $\phi(x) = x^2$ $\phi(x) = |x - 1|$

ANSWER:

(a) Plot A: $\phi(x) = |x|$ $\phi(x) = x^2$

(b) Plot B: $\phi(x) = x$ $\phi(x) = |x - 1|$

(c) Plot C: $\phi(x) = x^2$

(d) Plot D: $\phi(x) = |x|$

(e) Plot E: $\phi(x) = |x - 1|$

(f) Plot F: $\phi(x) = x$ $\phi(x) = |x|$ $\phi(x) = |x - 1|$