

CSE 446/546 Spring 2026 Midterm Exam

May 1, 2026

Name _____

UW NetID (not the numbers) _____

Please **wait** to open the exam until you are instructed to begin, and please take out your Husky Card and have it accessible when you turn in your exam.

Instructions: This exam consists of a set of short questions (True/False, multiple choice, short answer).

- For multiple choice questions, fill in the circle(s) to indicate your answers. Do not write your answer to the side. Example:



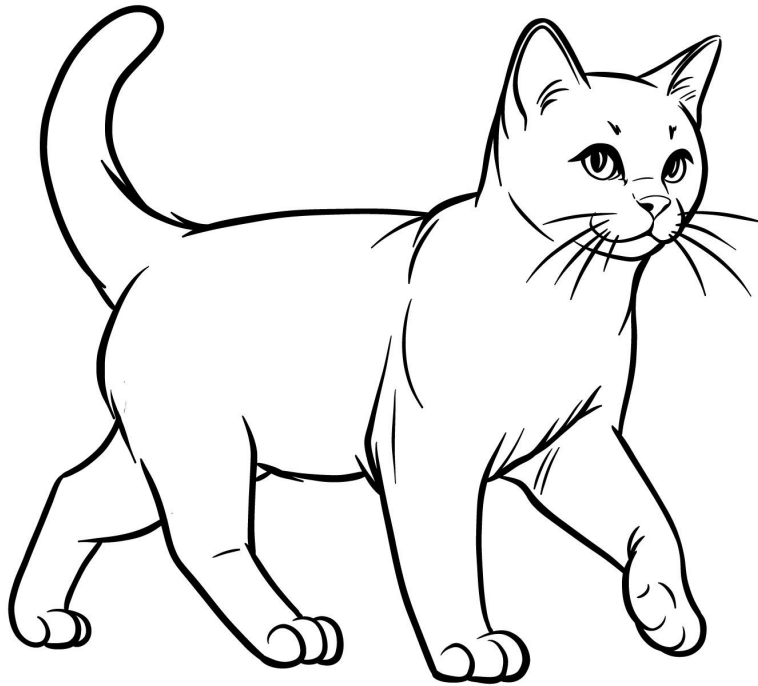
Selected answer.



Unselected answer.

- Multiple choice questions marked with One Answer should only be marked with one answer. All other multiple choice questions are Select All That Apply, in which case any number of answers may be selected (**including none, one, or more**).
- For Select All That Apply questions, we will give partial credit.
- For each short answer question, please write your answer in the provided space.
- If you need to change an answer or run out of space, clearly indicate what you would like graded as your final answer. If we cannot determine your final answer, we cannot give credit.
- If you complete the exam before the last 5 minutes, you may turn in your exam to course staff. Otherwise, please remain in your seats for the last 5 minutes of the exam.
- If you need to use the restroom, please first leave your exam with an instructor at the front of the room. You can then collect and resume your exam when you return.

This image is included only to cover the back of this page. It has no relation to the exam.



1. 1 points

Which of the following transformations of a non-negative likelihood function L preserve the MLE estimate $\hat{w}$, i.e, yield the same maximizer? Select All That Apply

☐ a $\log(L)$

☐ b $c * L$ where $(c \in \mathbb{R})$

☐ c $\sqrt{L}$

☐ d L^2

Solution: a, c, and d. This is because these functions are monotonically increasing transformations of L .

2. 1 points

You are training a linear regression model (with an offset term) to predict house prices given the following features: number of bathrooms, living room size in ft^2 , backyard size in ft^2 , total square footage in ft^2 . What is the minimum number of samples needed to learn a unique predictor? Briefly justify your answer.

Solution: At least 5 samples are needed since we have 4 features plus an offset term to learn, and need $n \geq d$ for $X^T X$ to be invertible.

3. 1 points

You have a dataset where each input vector $x \in \mathbb{R}^3$ has a label y , which is generated according to the true underlying function $y = 7 + x_1 + 2x_2^2 + 3x_3^3$. Let w denote the vector of parameters that are to be learned. Which of the following basis functions $h(x)$ could be used within a linear regression framework to achieve 100% training accuracy on this dataset without noise?

Select All That Apply

☐ a $h(x) = [7, x_1, x_2, x_3]^T$

☐ b $h(x) = [x_1, 2x_2^2, 3x_3^3]^T$

(c) $h(x) = [1, x_1, 2x_2^2, x_3^3]^T$

(d) $h(x) = [7, w_1x_1, 2w_2x_2^2, 3w_3x_3^3]^T$

(e) $h(x) = [3, x_1, x_2^2, x_3^5, x_3^3]^T$

Solution: Correct options: C, E.

Option A is wrong since there aren't any nonlinear basis functions for feature expansion, meaning the model is restricted to a linear function of x . Option B is wrong since it does not have an offset term; since y has an offset, we cannot achieve 100% train accuracy without an offset. Option D is wrong, as the features contain w ; a transformation containing w would mean that the relation between y and w is non-linear, meaning it cannot be used within a LR framework. Options C and E are correct, as they can perfectly fit the labels (e.g. consider coefficients of $[7, 1, 1, 3]$ and $[\frac{7}{3}, 1, 2, 0, 3]$, respectively).

4. 1 points

You are working with a dataset and decide to use a set of p basis functions $h(x) : \mathbb{R} \rightarrow \mathbb{R}^p$ for training. Once you're done training and obtain a learned weight vector w , you want to test the model on an unseen data point a . Write an equation for how you would calculate $\hat{y}(x) \in \mathbb{R}^d$ for this data point $a \in \mathbb{R}$.

Equation: _____

Solution: $\hat{y} = w^T h(a)$; We first transform the data point a using the same basis functions we used for training, $h(\cdot)$. We then apply the learned weights w to the transformed feature vector to obtain $\hat{y}$; we do so by taking the inner product of $h(a)$ and w .

5. 1 points

Suppose you are trying to predict $Y \in \mathbb{R}$ given $X \in \mathbb{R}^d$ and $(X, Y) \sim P_{XY}$. Which of the following predictors minimizes squared error loss? One Answer

- ☐ a $f(x) = \mathbb{E}[Y]$
- ☐ b $f(x) = \mathbb{E}[Y|X = x]$
- ☐ c $f(x) = \text{Var}(Y|X = x)$
- ☐ d $f(x) = \arg \max_y P(Y = y|X = x)$

Solution:

Correct option: B.

A is incorrect because it ignores X , which means that it's a constant prediction for all inputs.

B is correct, because under squared error, the best prediction is the conditional mean of Y given $X = x$.

C is incorrect because it answers how uncertain Y is instead of what Y is likely to be. You can rule this option out because Y could be negative, but variance is always non-negative.

D is incorrect because it's the mode, which is not optimal for squared error loss. You can rule this option out intuitively because squared error is a distance-based metric. The ideal prediction should fall somewhere in the middle of the data, which the mode is not guaranteed to be.

6. 1 points

The reason we solve the LASSO objective using gradient descent is: One Answer

- ☐ (a) To reduce overfitting
- ☐ (b) To encourage sparsity
- ☐ (c) There is no way to solve for the optimal weights analytically
- ☒ (d) **Solving for the optimal weights analytically is possible but computationally infeasible**

Solution: D: According to lecture 9 slide 25, LASSO can be solved analytically in 2^d time, which is computationally infeasible for large d .

7. 1 points

Suppose you have two datasets and you'd like to use cross-validation to finalize a model for both of them. Dataset A has 1 million samples with hundreds of features each, and Dataset B has 50 samples with a small number of hand-selected features.

For each dataset, you have the option of using leave-one-out (LOO) or k -fold cross-validation. Suppose $k = 5$.

- (a) For Dataset A, would you use LOO or k -fold cross-validation? Circle your answer below and give a brief explanation.

LOO

k -fold

- (b) For Dataset B, suppose the model is being used for a safety-critical healthcare application. Would you use LOO or k -fold cross-validation? Circle your answer below and give a brief explanation.

LOO

k -fold

Solution:

- (a) k -fold. LOO is extremely expensive with this large of a dataset.
- (b) **LOO**. Because the dataset is very small ($n = 50$) and the application is safety-critical, we want to use as much training data as possible in each fold to get the best estimate of performance. LOO provides an almost unbiased estimate of the true error because

each model is trained on $n - 1$ samples, which is particularly important when data is scarce and accuracy is paramount.

8. 1.5 points

Suppose you are examining the validation error for several different models trained on the same dataset. For the following candidate models, you observe:

1. Model A: Train MSE = 3.1, Validation MSE = 3.4
2. Model B: Train MSE = 0.5, Validation MSE = 7.8
3. Model C: Train MSE = 9.5, Validation MSE = 9.7

(a) Which model appears to be underfitting relative to the other models? Circle your answer below and give a brief explanation.

Model A

Model B

Model C

(b) Which model appears to be overfitting relative to the other models? Circle your answer below and give a brief explanation.

Model A

Model B

Model C

(c) Which model appears generalize best relative to the other models? Circle your answer below and give a brief explanation.

Model A

Model B

Model C

Solution:

- (a) Model C. Both training and validation error are high.
- (b) Model B. Validation error appears to be significantly higher than the training error.
- (c) Model A. Validation error is similar to the training error, and it has lower error overall than Model C.

9. 1 points

For $(X, Y) \sim P_{XY}$, let η minimize $\mathbb{E}_{XY}[(Y - \eta(X))^2]$. Given training dataset $\mathcal{D} \sim P_{XY}$ of size n , and test data set $\mathcal{T} \sim P_{XY}$ of size m , we fit a predictor $\hat{f}_{\mathcal{D}}$ from a set of functions $\mathcal{F}$ as follows:

$$\hat{f}_{\mathcal{D}} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{(x,y) \in \mathcal{D}} (y - f(x))^2.$$

Consider the following quantities A, B, C, D, and E. Match each quantity to its corresponding definition below. Write the letter corresponding to the quantity.

A. $\mathbb{E}_{Y|X} \left[\mathbb{E}_{\mathcal{D}} \left[(Y - \hat{f}_{\mathcal{D}}(x))^2 \right] \mid X = x \right]$

B. $\frac{1}{m} \sum_{(x,y) \in \mathcal{T}} (y - \hat{f}_{\mathcal{D}}(x))^2$

C. $\mathbb{E}_{\mathcal{D}} \left[\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] \right)^2 \right]$

D. $\frac{1}{n} \sum_{(x,y) \in \mathcal{D}} (y - \hat{f}_{\mathcal{D}}(x))^2$

E. $\mathbb{E}_{Y|X} \left[(Y - \eta(x))^2 \mid X = x \right]$

Training Error: _____

Testing Error: _____

Learning Error: _____

Generalization Error: _____

Irreducible Error: _____

Solution:

D represents the training error - the mean squared error over all the points in the training set.

B represents the test error - the mean squared error over all the points in the test set.

C represents the learning error - the expected squared difference between the optimal predictor and a fixed model trained on dataset D over all possible datasets.

A represents the generalization error - the expected squared error over all possible X and Y sampled from the true distribution P_{XY} .

E represents the irreducible error - the error that is inherent in our data (due to noise or measurement error) and cannot be captured even by the optimal predictor η

10. 1 points

How do the bias and variance of the ridge regression model typically change compared to the standard Ordinary Least Squares (OLS) estimator? One Answer

- (a) Both bias and variance decrease, because the ℓ_2 penalty reduces overfitting while also bringing estimates closer to the true parameters.
- (b) Bias decreases and variance increases, because the model becomes more flexible and fits the training data more closely.
- (c) Bias increases and variance decreases, because the ℓ_2 penalty shrinks coefficients toward zero, introducing systematic error but stabilizing the model.
- (d) Both bias and variance increase, because adding a penalty makes the model both more constrained and more sensitive to noise.

Solution: C Ridge regression introduces a penalty on the magnitude of the coefficients (shrinkage). By constraining the weights, the model becomes less sensitive to noise in the training data, which reduces variance. However, this intentional "simplification" of the model prevents it from fitting the training data perfectly, leading to an increase in bias.

11. 1 points

After running Lasso with regularization strength λ , a practitioner identifies two features with non-negligible weights and discards the rest. They then consider two strategies:

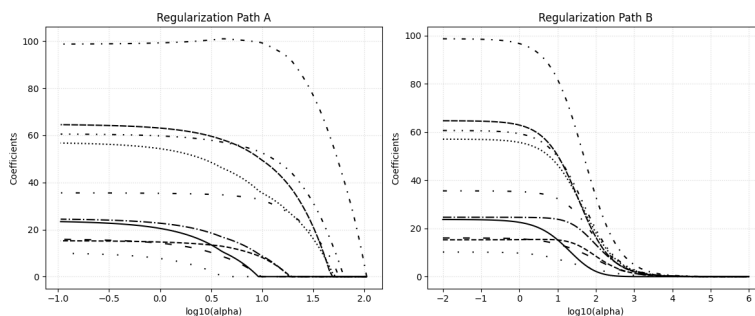
- **Strategy A:** Use the Lasso weights directly for the selected features
- **Strategy B:** Retrain an unpenalized linear regression using only the selected features

State which strategy is better and why.

Solution: Strategy B: Retraining without the ℓ_1 penalty removes the shrinkage bias introduced by Lasso, yielding coefficient estimates that better reflect the true relationship between the selected features and the target. This was discussed in detail in lecture 8.

12. 1.5 points

The following two plots (Path A and Path B) display the “path” of feature coefficients as the regularization strength parameter (λ) increases from left to right.



- (a) Identify which plot corresponds to LASSO (L_1) regularization and which corresponds to Ridge (L_2) regularization.

Left: _____

Right: _____

- (b) Explain the specific characteristic in the plots that allowed you to differentiate between the two.

Solution:

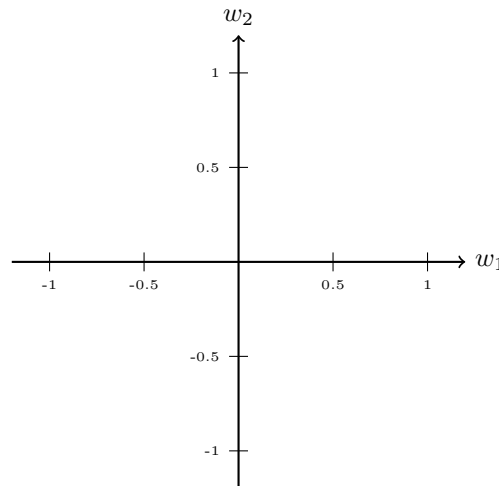
- (a) Left: Lasso, Right: Ridge
- (b) The weights in the plot on the left go to zero more quickly, whereas in the plot on the right the magnitude of the weights all gets smaller, there are fewer large weights, but many hover near zero.

13. 2 points

Suppose you have trained an unconstrained Ordinary Least Squares (OLS) model and obtained the weight vector $\hat{w}_{OLS} = [0.8, 0.4]^\top$. You are now considering applying regularization to this model.

Definition: A constraint is active if the unconstrained solution ($\hat{w}_{OLS}$) violates the constraint, forcing the regularized solution to change. It is inactive if the unconstrained solution already satisfies the constraint.

- (a) On the axes below, draw the ℓ_1 unit ball (the set where $\|w\|_1 \leq 1$). Then, plot and label the point $\hat{w}_{OLS}$ on the same axes.



- (b) Consider a LASSO-constrained problem: $\min \mathcal{L}(w)$ subject to $\|w\|_1 \leq 1$.

Based on your drawing and the definition of the ℓ_1 norm, is the constraint **Active** or **Inactive**? Justify your answer.

Solution:

- (a) The l_1 unit ball is a diamond passing through the points $(0,1)$, $(1,0)$, $(0,-1)$, $(-1,-1)$
- (b) The constraint is active because the unconstrained solution $[0.8, 0.4]$, lies outside the constraint ball, forcing the constrained solution to change relative to the unconstrained solution.

14. 1 points

Modern gradient descent algorithms use an adaptive step size to change the learning rate based on the current estimate of the magnitude of the gradient. When the gradient is larger, the step size is decreased, and when the gradient is smaller, the step size is increased. Give a conceptual explanation for this procedure which draws from our theoretical analysis of gradient descent, specifically relating it to the Lipschitz constant L and the learning rate η . (You do not have to prove anything in your answer, but please support your answer with relevant concepts or facts learned in class).

Solution: If $\eta = \frac{1}{L}$, then gradient descent is guaranteed to converge, i.e. the optimal learning rate that guarantees convergence depends on the value of L . The Lipschitz constant L bounds how quickly the gradient can change, so large gradients suggest large L and warrant a smaller step. Thus adaptive step size, which uses the current estimate of the gradient to set η , is a theoretically justified procedure because it enables taking smaller steps when gradients are large, however when gradients are small, it can take larger steps. This mimics the theoretically optimal $\eta = \frac{1}{L}$ locally.

15. 1 points

Which of the following is true about gradient descent: One Answer

- ☐ (a) If the loss function is convex, we are guaranteed convergence in a finite number of iterations.
- ☐ (b) If the loss gradient is Lipschitz continuous, we are guaranteed convergence in a finite number of iterations.
- ☐ (c) The loss is guaranteed to change at every iteration before convergence.
- ☒ (d) The weights are guaranteed to change at every iteration before convergence.

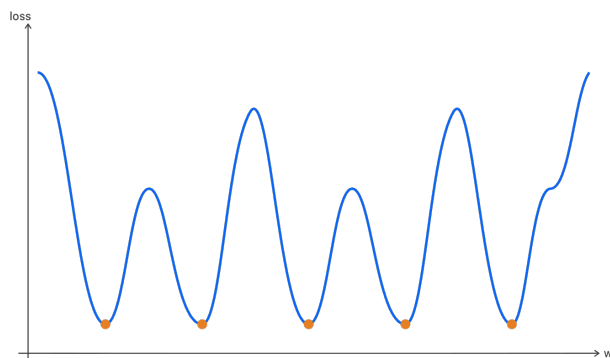
Solution: D: Observe that if the weights did not change, i.e., $w_{t+1} = w_t$, plugging in $w_{t+1} = w_t - \eta \nabla f(w_t)$ we find that

$$w_t - \eta \nabla f(w_t) = w_t \implies \eta \nabla f(w_t) = 0 \implies \nabla f(w_t) = 0,$$

and therefore, gradient descent must have already converged.

16. 1 points

The figure below plots a loss function for different values of the parameter w . Show that the function either is or is not convex, citing one of the definitions covered in class. You may draw on the figure.



Solution: The definitions of convexity are covered on slide 18 here <https://courses.cs.washington.edu/courses/cse446/26sp/schedule/lectures/10-convexity.pdf>. They can use:

Line above function: draw a line connecting any of the minima and show it's below the function

Epigraph: show that the points above the function are not a convex set

1st order Taylor expansion: show there are points where the tangent line goes above the function itself

Hessian: show the second derivative goes below 0

17. 1 points

Suppose we have a data matrix $X \in \mathbb{R}^{n \times 2}$ where the first feature is the square footage of a house and the second feature is the number of bedrooms, and labels $Y \in \mathbb{R}^n$ containing the house prices of each house. After fitting a linear regression model (with no data pre-processing

or regularization), we find that the resulting weights are

$$\hat{w} = \begin{bmatrix} 100 \\ 50000 \end{bmatrix}$$

Which feature is more important to the house price? One Answer

- ☐ (a) Square footage, because the variance of square footage is larger than the variance of the number of bedrooms.
- ☐ (b) The number of bedrooms, since its weight is much larger
- ☐ (c) Square footage, since a smaller weight leads to a smaller loss.
- ☒ (d) **We cannot infer from the weights which feature is more important.**

Solution: (d)

Here “important” can be synonymous with “predictive power”. B is incorrect. Because square footage is a much larger number than the number of bedrooms, its weight is going to be much smaller to compensate.

We cannot interpret the weights if we did not normalize the dataset before training.

18. 1.5 points

Suppose we are fitting a linear regression model using L1 penalty term (LASSO regression). To do this, we decide to use minibatch stochastic gradient descent (SGD) until convergence. In other words, for all k until convergence,

$$w_{k+1} = w_k - \eta \left(\frac{1}{m} \sum_{i=1}^m \nabla \ell_{I_i}(w_k) \right)$$

- (a) How does the variance of the gradient estimate change as we increase m from 1 to n , where n is the size of the dataset?

Solution: The variance decreases as you increase m

- (b) How does the bias of the gradient estimate change as we increase m from 1 to n ?
-

Solution: Regardless of the size of the batch, SGD is an unbiased estimator of gradient descent

19. 1 points

Using the same house prices data matrix from question 17, we now normalize each feature in the data to have 0 mean and unit variance. We then train a LASSO regression model on this data and find that feature i has a weight of 0 (i.e., $w_i = 0$). What can we say about how important feature i is in determining house prices? Explain your answer.

Solution: We cannot say anything about this features importance. It is possible that it is important to determining house prices, but is redundant with respect to feature j .

20. **Bonus question: answer this question correctly for an additional 0.5 points, but no points will be deducted if you do not answer.**

Given the linear regression model $y_i = x_i^\top w + v_i$, and assuming **uniform noise** (v_i is drawn i.i.d. from $\text{Uniform}(-\epsilon, \epsilon)$), write the expression for the likelihood function that is maximized by the learned parameter $\hat{w}$. Recall the PDF of a Uniform distribution on $[-\epsilon, \epsilon]$:

$$p(v_i) = \begin{cases} \frac{1}{2\epsilon} & \text{if } v_i \in [-\epsilon, \epsilon] \\ 0 & \text{otherwise} \end{cases}$$

In other words, find $L(w) = p(y_1, \dots, y_n \mid x_1, \dots, x_n, w)$ and simplify it to an expression without a $\prod$ or $\sum$ operator.

Solution:

Since $y_i = x_i^\top w + v_i$ and $v_i = y_i - x_i^\top w$:

$$p(y_i \mid x_i, w) = \begin{cases} \frac{1}{2\epsilon} & \text{if } |y_i - x_i^\top w| \leq \epsilon \\ 0 & \text{otherwise} \end{cases}$$

Since the v_i are i.i.d., the full likelihood is:

$$L(y_i) = \prod_{i=1}^n \begin{cases} \frac{1}{2\epsilon} & \text{if } |y_i - x_i^\top w| \leq \epsilon \\ 0 & \text{otherwise} \end{cases}$$

which simplifies to:

$$L(y_i) = \begin{cases} \left(\frac{1}{2\epsilon}\right)^n & \text{if all } |y_i - x_i^\top w| \leq \epsilon \\ 0 & \text{otherwise} \end{cases}$$

