CS 189/289A Introduction to Machine Learning
Fall 2024 Jennifer Listgarten, Saeed Saremi

Midterm

- Please do not open the exam before you are instructed to do so.
- **Electronic devices are forbidden on your person**, including cell phones, tablets, headphones, and laptops. Leave your cell phone off and in a bag; it should not be visible during the exam.
- The exam is closed book and closed notes except for your one-page 8.5×11 inch cheat sheet.
- You have 1 hour and 50 minutes (unless you are in the DSP program and have a larger time allowance).
- Please write your initials at the top right of each page after this one (e.g., write "JD" if you are John Doe). Finish this by the end of your 1 hour and 50 minutes.
- Mark your answers on the exam itself in the space provided. Do **not** attach any extra sheets. We will not grade any work outside of the space provided.
- For multiple choice questions, fill in the bubble for the single best choice.
- For short and long answer questions, write within the boxes provided. If you run out of space, you may use the last four pages to continue showing your work.
- **The last question is for CS289A students only**. Students enrolled in CS189 will **not** receive any credit for answering this question.

Your Name	
Your SID	
Name and SID of student to your left	
Name and SID of student to your right	

○ CS 189

○ CS 289A

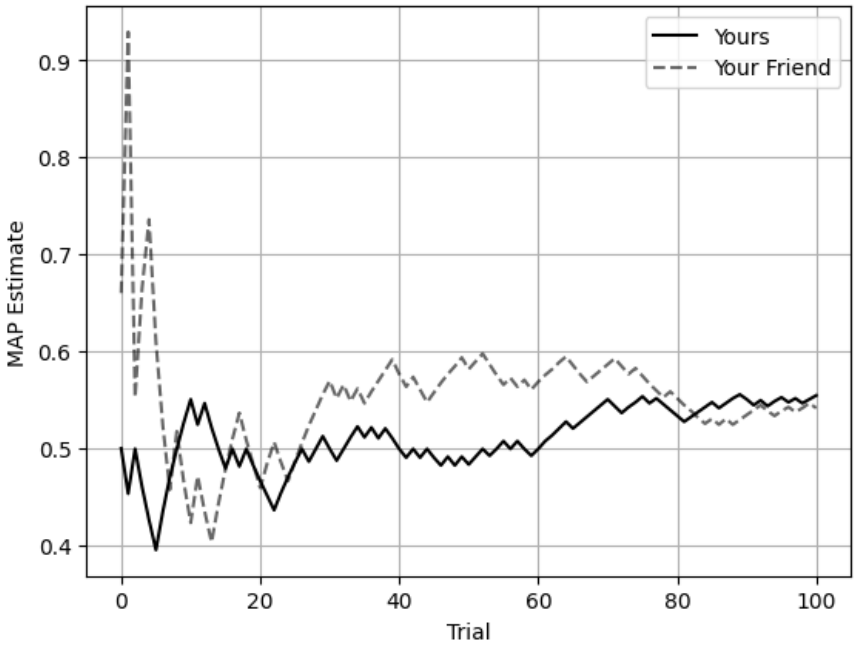
This page intentionally left blank.

initial here

1 Multiple Choice

For the following questions, select the **single best response**.

1. You and your friend are trying to estimate the probability of heads p of a coin. To do this, you both perform a MAP estimate with a binomial likelihood, and you each hand-select a Gaussian prior. You are pretty confident the coin is fair, so you use the prior $p \sim \mathcal{N}(0.5, 0.04^2)$. Your friend does not share your confidence and uses the prior $p \sim \mathcal{N}(\mu, \sigma^2)$. You each flip the coin 100 times and plot your estimate over time starting from $n = 0$ to $n = 100$ trials.



(0.5 points) What is most likely to be true about the mean of your friend's prior μ?

- ◯ $\mu > 0.5$
- ◯ $\mu < 0.5$
- ◯ $\mu = 0.5$

(0.5 points) What is most likely to be true about the variance of your friend's prior σ^2?

- ◯ $\sigma^2 > 0.04^2$
- ◯ $\sigma^2 < 0.04^2$
- ◯ $\sigma^2 = 0.04^2$

(0.5 points) What is most likely to be true about the true probability of heads p?

- ◯ $p = 0.6$
- ◯ $p = 0.55$
- ◯ $p = 0.5$

initial here

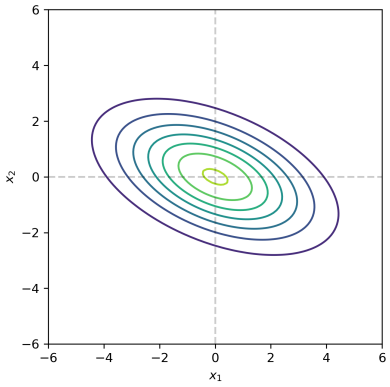
2. (1.5 points) Suppose we observe the following sequence of values:

$$\mathcal{D} = (2, 2, \pi, 5, \frac{100}{3}, \pi^2, 52, 3, 5, 6)$$

We hypothesize that our data is generated i.i.d. from a continuous uniform distribution $\mathcal{U}[a, b]$. What is the joint MLE of a, b?

- ◯ $a^* = 2, b^* = 52$.
- ◯ $a^* = b^*$, and a^* is the mean of $\mathcal{D}$.
- ◯ $a^* = -48, b^* = 102$.
- ◯ $a^* = \mu - 2\sigma, b^* = \mu + 2\sigma$, where μ, σ are the mean and variance of $\mathcal{D}$, respectively.

3. (1.5 points) Let $X = [x_1, x_2]^T$ be a 2-dimensional Gaussian random variable with mean zero whose probability density function is illustrated by the below contour plot.



Which of the following is most likely to be the correct covariance matrix for X:

- ◯ $\begin{bmatrix} 2 & 1 \\ -1 & 5 \end{bmatrix}$
- ◯ $\begin{bmatrix} 2 & -1 \\ -1 & 5 \end{bmatrix}$
- ◯ $\begin{bmatrix} 5 & -1 \\ -1 & 2 \end{bmatrix}$
- ◯ $\begin{bmatrix} 5 & 1 \\ -1 & 2 \end{bmatrix}$

4. (1.5 points) Which of the following statements about regularization is **false**?

- ◯ The LASSO objective is convex.
- ◯ You can combine different forms of regularization.
- ◯ It is not possible to add an L2-regularization term to the cost function of a neural network.
- ◯ Weight sharing in convolutional neural networks can be viewed as regularization.

initial here

5. (1.5 points) Which of the following statements about generative vs discriminative models for classification is **true**?

   ◯ Logistic regression is a generative approach.

   ◯ Consider a method where we use MLE to fit Gaussian distributions to features, conditioned on each class, then choose the most likely class. This approach is generative.

   ◯ Generative models directly model $p(y \mid x)$, where x denotes input features and y denotes output labels.

   ◯ Suppose our data is linearly separable, and we choose a separating hyperplane by maximizing the distance from the nearest data points of each (binary) class. This is a generative approach.

6. (1.5 points) Which of the following statements about stochastic gradient descent (SGD) is **false**?

   ◯ The computation time it takes for SGD to converge is always greater than that of standard gradient descent.

   ◯ SGD can be used to find approximate solutions to non-convex problems.

   ◯ SGD is more memory efficient than standard gradient descent.

   ◯ Due to its stochastic nature, SGD can escape local minima.

7. (1.5 points) Which of the following statements about the backpropagation algorithm is **true**?

   ◯ Backpropagation cannot be used to compute the gradients for self-attention layers.

   ◯ Backpropagation requires neural networks to use activation functions that are differentiable everywhere.

   ◯ Backpropagation requires a forward pass to be run through a neural network before the backward pass can be called.

   ◯ The gradients computed during backpropagation can only be used for gradient descent.

8. (1.5 points) Suppose a convolutional layer has the following specifications:

   - Input dimensions: [height = 10, width = 10, channels = 5]
   - Output dimensions: [height = 10, width = 10, channels = 20]
   - Kernel size: [height = 2, width = 2]
   - Each kernel also contains an additional bias term.

   How many trainable parameters are there in this convolutional layer?

   ◯ 80

   ◯ 100

   ◯ 400

   ◯ 420

initial here

9. (1.5 points) Which of the following statements about Transformers is **false**?

   ◯ Transformers use attention to make the loss convex.

   ◯ Masked attention is only necessary in the decoder.

   ◯ A constant positional encoding is equivalent to no positional encoding.

   ◯ None of the above, they are all true.

10. (1.5 points) The method t-SNE is used to visualize high-dimensional data in a lower dimensional space. Which of the following statements regarding t-SNE is **false**?

    ◯ The method is stochastic.

    ◯ The method tries to match neighborhood probabilities in the high-dimensional space and the low-dimensional space.

    ◯ The method uses total variation distance in its loss function, hence the name "total variation SNE".

    ◯ The method uses a t-distribution to prevent crowding in the low dimensional space.

initial here

2 Short Answer

1. (2 points) Let Y be a multivariate Gaussian that can be expressed as AX, where $A \in \mathbb{R}^{3\times 3}$ is a matrix and $X \in \mathbb{R}^3$ is a collection of i.i.d. standard normal RVs. Assume that Y has zero mean. What is Σ_Y, the covariance of Y?

2. Suppose we have a dataset of independent and identically distributed data points

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

where $x_i \in \mathbb{R}$ and $y_i \in \{0, 1\}$. Further assume that that data from the two classes fall on different sides of the origin (i.e. $y_i = 0$ when $x_i < 0$ and $y_i = 1$ when $x_i > 0$). We train a logistic regression model (without an intercept term) defined by

$$p(y_i = 1 \mid x_i) = \frac{1}{1 + e^{-\beta x_i}}.$$

(a) (1 point) For what values of $\beta \in \mathbb{R}$ will the model obtain a perfect accuracy on the dataset? Recall that for binary-classification, perfect accuracy is achieved if $p(y_i = 1 \mid x_i) > 0.5$ when $y_i = 1$ and $p(y_i = 1 \mid x_i) < 0.5$ when $y_i = 0$.

(b) (1 point) Say that there is some $\hat{\beta}$ that obtains perfect accuracy. Find a β^* that also obtains perfect accuracy but increases the log-likelihood (or equivalently, lowers the cross-entropy loss) compared to $\hat{\beta}$.

initial here

3. (2 points) The SiLU (sigmoid linear unit) function is defined as follows

$$\text{SiLU}(x) = x \cdot \sigma(x)$$

where $\sigma(x)$ is the sigmoid function. For this question, we will only consider scalar inputs $x \in \mathbb{R}$. The derivative of the SiLU function can be written as

$$\frac{d}{dx}\text{SiLU}(x) = b[1 + a \cdot (1 - b)]$$

What are the missing terms a and b?

4. (2 points) Suppose we have a convolutional layer composed of a single filter where the bias parameter is zero and the weight parameter is

$$W = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

Consider the following input X containing a single channel

$$X = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 3 & 4 \end{bmatrix}$$

Calculate the (pre-activation) output of the convolutional layer applied on X, with no padding and a stride of 1.

initial here

5. (2 points) Consider a convolutional layer with the following description:

   - Number of filters: 128
   - Kernel size: 4x4
   - Stride: 2
   - Padding size: 1

   Calculate the output shape $[n, c, h, w]$ after applying the convolutional layer to an input of shape $[32, 3, 28, 28]$. The shapes are ordered as [batch size, number of channels, height, and width].

6. (2 points) Suppose we are applying self-attention (with one head) on n tokens. Let $q_1, \ldots, q_n$ be the query vectors, $k_1, \ldots, k_n$ be the key vectors, and $v_1, \ldots, v_n$ be the value vectors.

   Assume that q_i is orthogonal to the key vectors for all tokens (including that for token i). What will the output of self-attention be for token i?

initial here

7. (2 points) Consider the following design matrix containing sample points $X_i \in R^2$.

$$X = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -2 & 0 \end{bmatrix}$$

Using PCA, what is the first principal component and its explained variance?

8. (2 points) Describe one advantage of t-SNE over PCA and one advantage of PCA over t-SNE.

initial here

3 A Classifier for Count Data

We consider a binary classification problem in which the observed features are counts. Specifically, for the i-th observation, let $Y_i \in \{0, 1\}$ be the class label, and $X_i \in \mathbb{N}$ be the count features.

We assume that the Y_i are i.i.d. Bernoulli random variables with parameter θ. We also assume that the X_i are i.i.d. random variables with the following distribution

$$\begin{cases} X_i \sim \text{Poisson}(\lambda_1) & \text{if} \quad Y_i = 1 \\ X_i \sim \text{Poisson}(\lambda_0) & \text{if} \quad Y_i = 0. \end{cases}$$

Recall if Z is a Poisson distribution of parameter λ, its probability mass function is given by

$$P(Z = k) = \frac{\exp\{-\lambda\}\lambda^k}{k!}.$$

We hope to learn $(\theta, \lambda_0, \lambda_1)$ using maximum likelihood estimation as follows:

$$(\hat{\theta}, \hat{\lambda}_0, \hat{\lambda}_1) = \underset{\theta, \lambda_0, \lambda_1}{\text{argmax}} \ \mathcal{L}(\theta, \lambda_0, \lambda_1),$$

where $\mathcal{L}(\theta, \lambda_0, \lambda_1) = \sum_{i=1}^{n} \log P(Y_i, X_i; \theta, \lambda_0, \lambda_1)$ is the log-likelihood of the data. Here, $P(Y_i, X_i; \theta, \lambda_0, \lambda_1)$ is the joint probability of observing Y_i and X_i.

(a) (3 points) Show that log-likelihood $\mathcal{L}(\theta, \lambda_0, \lambda_1)$ can be written, up to a constant, as

$$\sum_{i=1}^{n} Y_i \left[X_i \log \lambda_1 - \lambda_1 + \log \theta\right] + (1 - Y_i) \left[X_i \log \lambda_0 - \lambda_0 + \log(1 - \theta)\right]$$

initial here

(b) (3 points) Is $\theta, \lambda_0, \lambda_1 \mapsto \mathcal{L}(\theta, \lambda_0, \lambda_1)$ concave? Justify your answer by computing $\nabla^2 \mathcal{L}(\theta, \lambda_0, \lambda_1)$. Assume the Y_i are not all the same value.

initial here

(c) (3 points) Compute the maximum likelihood estimates $\hat{\theta}, \hat{\lambda}_0, \hat{\lambda}_1$.

(d) (1 point) Is this model generative or discriminative? Why?

initial here

4 Matrix Decomposition on Ridge Regression

Consider the following common set-up: we have a dataset $X \in \mathbb{R}^{n \times d}$ that contains n data points and d features per data point, and a target dataset $Y \in \mathbb{R}^n$ of target outputs. We parameterize a linear regression model $f(x) = w^T x$, where w is a learned weight vector and x is a data point.

To train our model we use ridge regression, where we solve the following optimization problem for some fixed value of λ

$$w_{\text{opt}} = \underset{w}{\text{argmin}} \ ||Xw - Y||^2 + \lambda ||w||^2$$

The optimal solution to this problem is

$$w_{\text{opt}} = (X^T X + \lambda I)^{-1} X^T Y$$

(a) (1 point) As λ approaches ∞, what does w_{opt} approach? Provide a brief justification.

(b) (2 points) Suppose we know the SVD of X, where $X = U \Sigma V^T$. Recall that $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{d \times d}$ are orthonormal matrices, and $\Sigma \in \mathbb{R}^{n \times d}$ is defined as $\text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$.

Consider the matrix $(X^T X + \lambda I)$ that is computed to produce w_{opt}. Write the spectral decomposition of $(X^T X + \lambda I)$.

initial here

(c) (3 points) When $\lambda \gg \sigma_1^2$, what matrix does $(X^T X + \lambda I)^{-1}$ approach? You may find your answer from part (b) useful.

(d) (3 points) Now, we want to consider how the direction of w_{opt} changes as we vary λ. To do this, we define $w_{\text{norm}} = \frac{w_{\text{opt}}}{\|w_{\text{opt}}\|}$. Write an expression for w_{norm} in terms of X, Y, λ as λ approaches ∞. You may find your answer from part (c) useful.

initial here

5 Residuals vs Errors in Linear Regression

In linear regression, we hold the following hypotheses:

i. $Y = Xw + e$

ii. $e \sim \mathcal{N}(0, \sigma^2 I)$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$, $w \in \mathbb{R}^d$, and $e \in \mathbb{R}^n$. We call e the vector of "errors."

The errors are a theoretical quantity that we don't know as they depend on knowledge of the true (often unobserved) value. What we can do is look at the residuals defined by $\epsilon = Y - \hat{Y}$, the difference between our observed Y and our predicted $\hat{Y}$. In this problem, you will show that the residuals do not share the same distribution as the errors.

(a) (3 points) From the above assumptions, we can derive that $Y \mid X \sim \mathcal{N}(Xw, \sigma^2 I)$. You may use this fact without proof.

Show that the MLE of w (keeping σ fixed) is the same as the minimizer of the sum of squared residuals. Precisely, demonstrate the following:

$$\operatorname*{argmax}_w \mathcal{L}(w; Y \mid X) = \operatorname*{argmin}_w \|Y - Xw\|_2^2.$$

initial here

(b) (3 points) Suppose we minimize the above loss function (sum of squared residuals) and attain our optimal weight vector w^*. Prove that, when our data matrix contains a **bias term** (that is, one column of X is the vector with every entry being 1), then the sum of residuals is 0. Precisely, prove that

$$\sum_{i=1}^{n} \epsilon_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0,$$

where ϵ is our vector of residuals defined as $Y - \hat{Y}$, and $\hat{Y} = Xw^*$. You do **not** need to prove convexity of $\|Y - Xw\|_2^2$.

Hint: It may be helpful to rewrite $Y_i - \hat{Y}_i$ as $Y_i - (X'w')_i - w_0$, where X' is the data matrix with the bias column removed, w' is the weight vector with the bias weight removed, and w_0 is the bias weight.

initial here

(c) (3 points) Using the result of the previous part, prove that the residuals ϵ do not share the same distribution as the errors e. Precisely, prove that

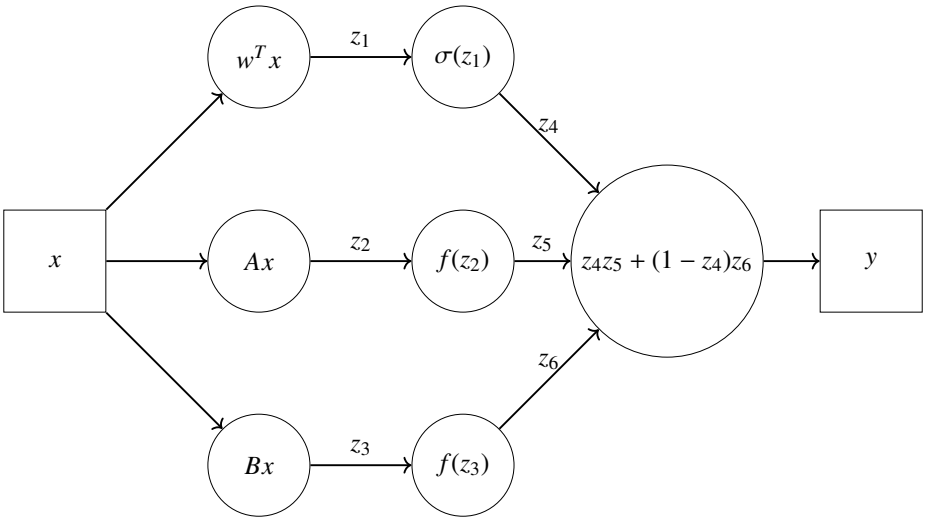
$$\epsilon \nsim \mathcal{N}(0, \sigma^2 I).$$

Hint: A multivariate Gaussian with a diagonal covariance matrix has (mutually) independent components. You may use this fact without proof.

initial here

6 Computational Graph Analysis

You are building a small neural network which takes in a vector $x \in \mathbb{R}^d$ and outputs another vector $y \in \mathbb{R}^d$. The computational graph of your network looks like this



where $w \in \mathbb{R}^d$, $A, B \in \mathbb{R}^{d \times d}$ are the parameters of your layers, σ is the sigmoid function, and $f : \mathbb{R}^d \to \mathbb{R}^d$ is some activation function.

(a) Suppose $f(x) = x$ is the identity function. Compute the following derivatives in terms of the variables defined in the computational graph.

i. (1.5 points) $\partial y / \partial z_1$

ii. (1 point) $\partial y / \partial z_2$

initial here

iii. (1 point) $\partial y / \partial A_{*j}$, where A_{*j} is the j-th column of A

iv. (1.5 points) $\partial y / \partial x$

(b) (2 points) Suppose $f(x) = \text{ReLU}(x)$. Briefly identify when and how each term of $\partial y / \partial x$ would change from your answer in part (a).

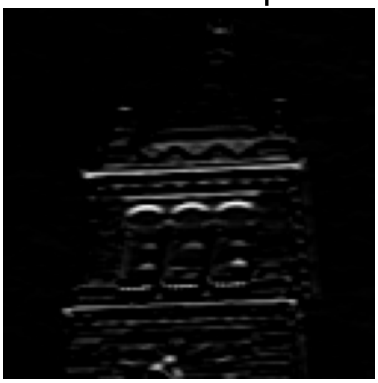
initial here

7 Reverse-Engineering CNN Filters

You have an image and the outputs from two different convolutional filters, and you want to reverse-engineer what filters were applied. You know that the filters are 3x3 with no bias term, applied with no padding and a stride of 1, followed by a ReLU activation. The image and the two outputs are normalized such that all values are $\in [0, 1]$, where 0 is black and 1 is white.

Input Image

Filter 1 Output



Filter 2 Output

(a) (3 points) Fill in weights for each filter. All weights are a value in the set $\{-1, 0, 1\}$. Some values have already been filled in for you.

Filter 1
$$\begin{bmatrix} 1 & \_\_ & \_\_ \\ \_\_ & 0 & \_\_ \\ \_\_ & \_\_ & -1 \end{bmatrix}$$

Filter 2
$$\begin{bmatrix} 1 & \_\_ & \_\_ \\ \_\_ & 0 & \_\_ \\ \_\_ & \_\_ & -1 \end{bmatrix}$$

(b) (1 point) Qualitatively, what types of features are being identified by each filter?

initial here

8 Gradient Descent for Linear Regression (CS289A Only)

Only complete this problem if you are enrolled in CS289A.
Do **not** complete this problem if you are enrolled in CS189.

Suppose we have a dataset of n samples $D = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We can stack the data into an $n \times d$ design matrix X and the labels into a vector $y \in \mathbb{R}^n$. Recall that the objective of the ordinary least squares problem is to find a weight vector $w \in \mathbb{R}^d$ that minimizes the squared error $J(w) = \|Xw - y\|_2^2$. You may use without proof that $J(w)$ is convex.

(a) (2 points) Show that any $w \in \mathbb{R}^d$ that satisfies the *normal equation* $X^\top Xw = X^\top y$ will be a solution to the OLS problem. When does a unique solution exist, and what is that solution?

(b) (2 points) In Big-O notation, what's the computational cost of finding a solution to the normal equation $X^\top Xw = X^\top y$? We will measure the cost in FLOPs (floating point operations), i.e., the total number of additions, divisions, subtractions and multiplications between two scalar numbers required to perform a computation.
Hint: Solving a linear system, represented by an $m \times m$ matrix, using Gaussian elimination takes $O(m^3)$ FLOPs.

initial here

(c) (1 point) We learned about an iterative algorithm called gradient descent in lecture for training logistic regression models and neural networks. However, gradient descent is a very general algorithm that can be applied to many optimization problems, including least squares. Write down the gradient descent update for w. Denote the learning rate by η.

(d) (2 points) In Big-O notation, what is the computational cost of performing gradient descent for L iterations? Once again, we will measure the cost in FLOPs. Different sequences of computations will yield different costs. Pick the lowest cost.

(e) (1 point) For an appropriately chosen learning rate, gradient descent will converge as $L \to \infty$. However, we can only run a finite number of gradient descent iterations on a real computer so this solution ends up being an approximate minimizer of the OLS objective, unlike the solution returned by the normal equation, which will be an exact minimizer.

Regardless, it is still used as one of the main methods for solving OLS in practice. In what situations might the gradient descent approach be preferable over attempting to solve the normal equation directly?

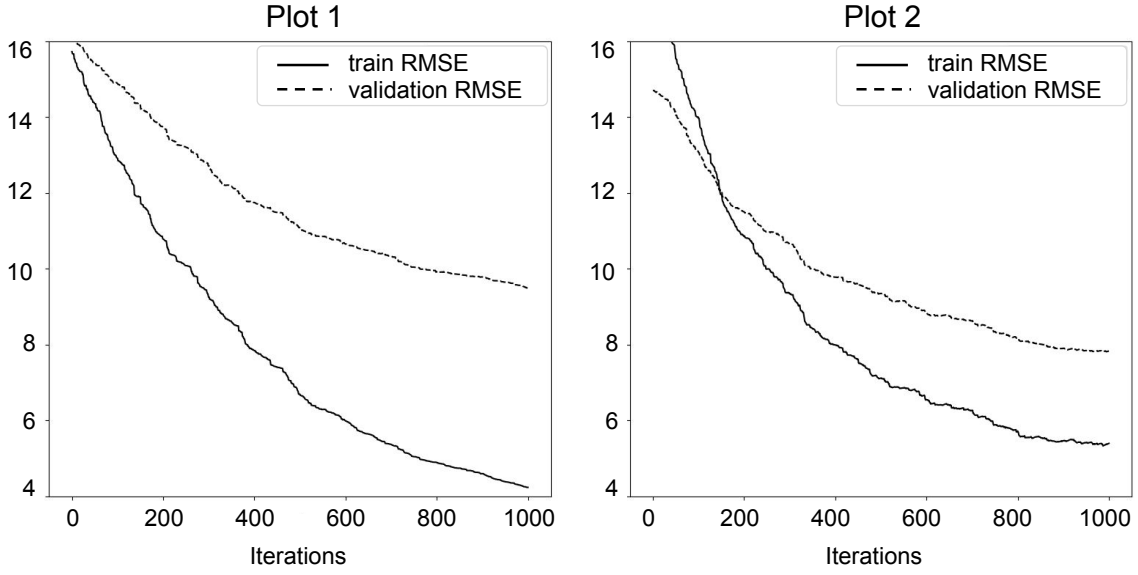
initial here

(f) (2 points) Let $J_\lambda(w) = \|Xw-y\|_2^2+\lambda\|w\|_2^2$, for $\lambda > 0$, be the objective function for ordinary least squares with an l2-penalty, i.e., ridge regression.

Suppose we split the dataset into a training and validation split, and run stochastic gradient descent (SGD), using only the training set, for 1000 iterations to optimize $J(w)$ versus $J_\lambda(w)$, for some $\lambda > 0$. We also plot the *root mean squared error* (RMSE) of the parameter vector $w^{(t)}$ at iteration t for each $t = 0, \ldots, 1000$, on the entire training and validation splits, for each experiment. The root mean squared error is defined as

$$\text{RMSE}(X, y, w) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i^\top w - y_i)^2}$$

for $X \in \mathbb{R}^{m\times d}$ and $y \in \mathbb{R}^m$, where (X, y) is either the train or validation split with m samples.



We have plotted the training and validation RMSE curves from running SGD to minimize $J(w)$ and $J_\lambda(w)$, but we don't know which plot corresponds to which objective! Based on just the figure above, identify this correspondence and briefly justify your reasoning (1-2 sentences is enough).

initial here

You may use this page to show extra work. Clearly mark your work with the problem number here, and also mention in the problem-specific box that your work is continued here.

initial here

You may use this page to show extra work. Clearly mark your work with the problem number here, and also mention in the problem-specific box that your work is continued here.

initial here

You may use this page to show extra work. Clearly mark your work with the problem number here, and also mention in the problem-specific box that your work is continued here.

initial here

You may use this page to show extra work. Clearly mark your work with the problem number here, and also mention in the problem-specific box that your work is continued here.