DSC 140A - Quiz 01

Name:
PID:

Quiz 01 Instructions. All technology is prohibited. Pen/pencil and paper only. You may not discuss with any classmates. Any prohibited use of notes or technology will result in a 0 and is a violation of Academic Integrity. You will have 35 minutes to complete your quiz. The rest of the discussion time, we can go over the quiz together. If you have any questions, please raise your hand and a TA will come assist you. Make sure to write your name and PID at the top of every page. Explain/show your work.

Notation: Bold upper case letters $\mathbf{A}$ represent matrices.
Bold lower case letters $\mathbf{a}$ represent vectors.
Non-bolded lower case letters a represent scalars.
For a matrix $\mathbf{A}$, $\mathbf{a}_i$ indicates the ith row vector of matrix $\mathbf{A}$.

Problem 1.

The goal of this problem is to test your understanding of gradients of a risk function.

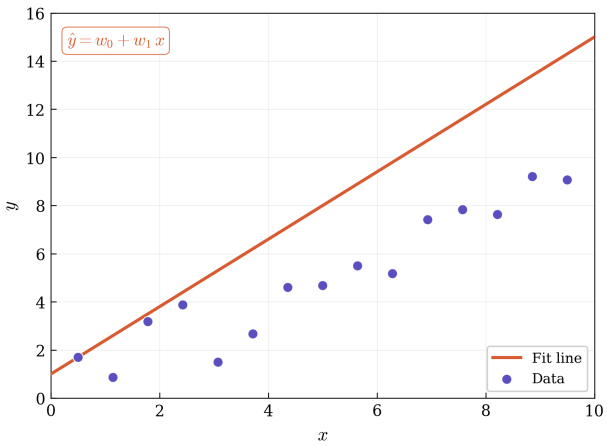
Let $R(\mathbf{w})$ be the risk with respect to the mean squared error for this data, assuming a linear prediction function of the form $f(x) = w_0 + w_1 x$. Let $\nabla_w R(\mathbf{w})$ be the gradient of the risk function with respect to the weights vector $\mathbf{w}$.

a) In one or two sentences, describe what the risk $R(\mathbf{w})$ tells us about the fit of $f(x)$ to the training data.

Solution:

The risk $R(\mathbf{w})$, calculated as the Mean Squared Error (MSE), quantifies the average squared vertical distance between the actual data points (y) and the values predicted by the model ($\hat{y}$). Essentially, it serves as a "cost" or "penalty" score: a higher risk indicates a poor fit where the line is far from the data, while a lower risk indicates that the model accurately captures the trend of the training data.

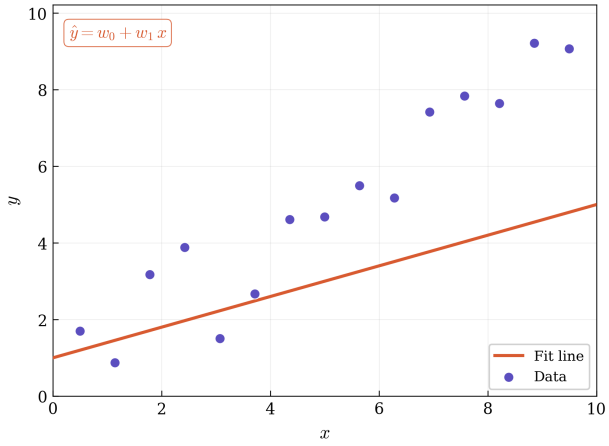
b) Consider the data and model below:



What is the sign (positive or negative) of $\frac{\partial R(\mathbf{w})}{\partial w_1}$? Explain your reasoning.

Solution: Positive. $\frac{\partial R(\mathbf{w})}{\partial w_1} = \lim_{\delta w_1 \to 0} \frac{R(w_0, w_1 + \delta w_1) - R(w_0, w_1)}{\delta w_1}$. When increasing w_1, the fitted line gets steeper, and be farther from the data. $R(\mathbf{w})$ will increase. From the definition, the partial derivative has a positive sign.

c) Now, consider this data and model below:



What is the sign (positive or negative) of $\frac{\partial R(\mathbf{w})}{\partial w_1}$? Explain your reasoning.

Solution: Negative. When increasing w_1, the fitted line will be closer to the data, $R(\mathbf{w})$ will decrease. From the definition, the partial derivative has a negative sign.

Problem 2.

Let $\mathbf{X} \in \mathbb{R}^{n \times (d+1)}$ be the augmented design matrix with n samples and $d+1$ features (including the intercept). Upon inspecting the data, you believe that using a **linear** class of models will represent the data well (Hint: linear models are of the form $f_w(x) = \mathbf{w}^T\mathbf{x}$). Using mean squared error ($R(\mathbf{w}) = \frac{1}{n}\sum_{i\in[n]}(f_w(\mathbf{x}_i) - y_i)^2 = \frac{1}{n}\|f_w(\mathbf{X}) - \mathbf{y}\|^2$) as empirical risk, calculate the optimal weights (i.e. the weights that give the lowest risk) for our linear model using vector-matrix notation (i.e. do not use summations).

Solution: 1. Set up the Risk Function The empirical risk (Mean Squared Error) is given by:

$$R(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

We can expand the squared L^2 norm using the property $\|\mathbf{a}\|^2 = \mathbf{a}^T\mathbf{a}$:

$$R(\mathbf{w}) = \frac{1}{n}(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y})$$

2. Expand the Expression Expanding the transpose and distributing the terms:

$$R(\mathbf{w}) = \frac{1}{n}(\mathbf{w}^T\mathbf{X}^T - \mathbf{y}^T)(\mathbf{X}\mathbf{w} - \mathbf{y})$$

$$R(\mathbf{w}) = \frac{1}{n}(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{y}^T\mathbf{y})$$

Since $\mathbf{w}^T\mathbf{X}^T\mathbf{y}$ is a scalar, it is equal to its own transpose ($\mathbf{y}^T\mathbf{X}\mathbf{w}$), we can combine the middle terms:

$$R(\mathbf{w}) = \frac{1}{n}(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y})$$

3. Take the Gradient To find the minimum, we take the gradient with respect to $\mathbf{w}$ and set it to zero. Using standard matrix calculus identities: $\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T\mathbf{A}\mathbf{w}) = 2\mathbf{A}\mathbf{w}$ (for symmetric $\mathbf{A}$) and $\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T\mathbf{b}) = \mathbf{b}$:

$$\nabla_{\mathbf{w}} R(\mathbf{w}) = \frac{1}{n}(2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y})$$

4. Solve for $\mathbf{w}$

Set the gradient to the zero vector $\mathbf{0}$:

$$\frac{2}{n}(\mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{y}) = \mathbf{0}$$

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$$

Assuming $\mathbf{X}^T\mathbf{X}$ is invertible (i.e., the features are linearly independent), we multiply both sides by the inverse $(\mathbf{X}^T\mathbf{X})^{-1}$:

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$