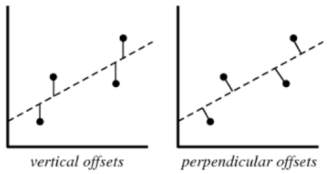
Discussion Session Problems 1

01/14/2026

1. Which of the following offsets, do we use in linear regression's least square line fit? Assume the horizontal axis is the independent variable and vertical axis is dependent variable.



   A) Vertical offset
   B) Perpendicular offset
   C) Both, depending on the situation
   D) None of above

   Answer: A) Vertical offset
   Linear regression minimizes the sum of squared vertical distances (residuals) between data points and the fitted line. This is because we're predicting Y given X, so we measure error in the Y-direction.

2. Which of the following if any is a valid cost function in a regression setting and why?

   a. $J(w) = \frac{1}{2m} \sum_{i=1}^{m} (f_w(x^{(i)}) - y^{(i)})^2$
   b. $J(w) = \frac{1}{2m} \sum_{i=1}^{m} (f_w(x^{(i)}) - y^{(i)})$
   c. $J(w) = \frac{1}{2m} \sum_{i=1}^{m} |f_w(x^{(i)}) - y^{(i)}|$

   Answer: A, C.
   Option (a) is the standard cost function for linear regression because squaring the errors ensures they are always positive and penalizes large deviations, creating a smooth convex curve that is easy to optimize. Conversely, Option (b) is invalid because it sums raw errors, allowing positive and negative values to cancel out, which could falsely report zero error. Option (c) uses absolute values; this is a valid cost function to make progress with gradient descent, but be aware of the sharp corner (non-differentiable) at zero.

3. You are building a model to set real estate prices. If the predicted price is too high no customer will buy the house, but the monetary loss is low because the price can easily be decremented. Of course it should not be too high as then the house may not be bought for a long time. On the other hand if the predicted price is too low, the house will be bought quickly without having a chance to adjust the price. In other words the learning algorithm should predict slightly higher prices which can be decremented if necessary rather than underestimating the 'good' price which will result in an immediate monetary loss. How would you design an error metric incorporating this cost asymmetry? Write your new cost function and draw a sketch of the graph where you plot the cost versus w.

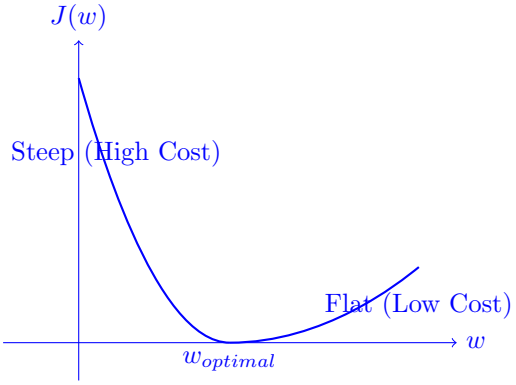
We can use a modified squared error cost function that incorporates a tuning parameter α to penalize underestimation more heavily than overestimation.

$$J(w) = \frac{1}{2m} \sum_{i=1}^{m} \left(f_w(x^{(i)}) - y^{(i)} \right)^2 \cdot \left(\text{sgn}(f_w(x^{(i)}) - y^{(i)}) + \alpha \right)^2$$

where $-1 < \alpha < 0$.

- If the prediction is too low (underestimation), the sign is -1, resulting in a penalty factor of $(-1 + \alpha)^2$. Since α is negative, this factor is large (> 1).
- If the prediction is too high (overestimation), the sign is $+1$, resulting in a penalty factor of $(1 + \alpha)^2$. Since α is negative, this factor is small (< 1).

The graph below shows the cost $J(w)$ versus w. The curve is steeper for lower w (underestimation) and shallower for higher w (overestimation), encouraging the model to "err on the side of caution" by predicting slightly higher prices.



4. An outlier in a regression model:

   a) Always improves model accuracy

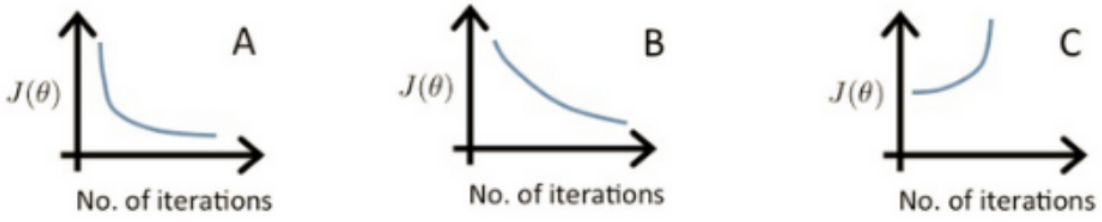
   b) Will most likely change or skew the regression line's slope

   c) Has no impact on the model

   d) Automatically indicates a data collection error

Answer: B) Will most likely change or skew the regression line's slope
Least squares regression minimizes the sum of squared errors. Because the error is squared, a single outlier (a point far from the trend) creates a massive penalty. To minimize this total cost, the regression line must tilt significantly toward the outlier to reduce that specific distance, effectively "skewing" the slope of the entire model.

5. a. What can you say about the relationship between the cost function and the number of iterations in the graphs below?



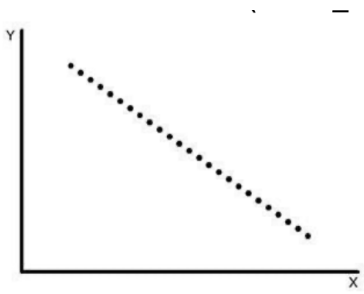
For graph A, the learning rate is well-tuned. The cost decreases rapidly at first and then levels off as it converges to the minimum.
For graph B, the learning rate is too small. The cost decreases, but the slope is shallow, meaning convergence will be very slow and computationally expensive.
For graph C, the learning rate is too large. The algorithm is overshooting the minimum, causing the cost to increase (diverge) rather than decrease.

b. Suppose l_1, l_2 and l_3 are the three learning rates for A, B, C respectively. Which of the following is true about l_1, l_2 and l_3?

A) $l_2 < l_1 < l_3$
B) $l_1 > l_2 > l_3$
C) $l_1 = l_2 = l_3$
D) None of these

Answer: A.

6. Consider the following data where one input(X) and one output(Y) is given. What would be the cost for this data if you run a Linear Regression model of the form ($Y = w_1 \cdot x_1 + b$)?



A) Less than 0
B) Greater than zero
C) Equal to 0
D) None of these

Answer: C) Equal to 0

Hyperparameter setting	Learning Rate	Number of iterations	Training Error	Validation Error
1	0.1	1000	100	110
2	0.2	600	105	105
3	0.3	400	110	110
4	0.4	300	120	130
5	0.4	250	130	150

7. Which of the following hyperparameter settings is seemingly the best?

A) 1
B) 2
C) 3
D) 4

Answer: B.
This is because it has the lowest validation error.

8. Normal equations are very slow when we have a big X. However technically they should work all the time. **True/False**

   Normal Equations require computing $(X^T X)^{-1}$. This inverse does not exist if the features are linearly dependent (redundant) or the number of features exceeds the number of training examples.

9. Why is feature scaling important in linear regression?

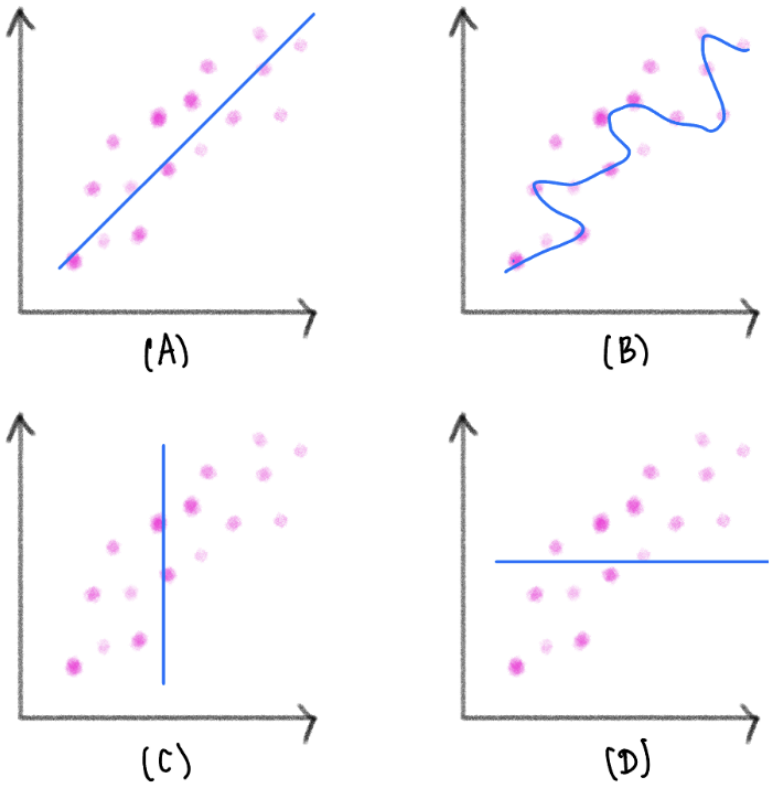
   a) It always improves model accuracy
   b) It prevents variables with larger magnitudes from dominating
   c) It guarantees perfect predictions
   d) It reduces computational time

   Answer: B.

10. Can we model non-linear relationships with a linear regression?

    Answer: Yes.
    The term "Linear" in Linear Regression refers to the model being linear in the parameters (weights, w), not necessarily in the input features (x). By transforming the input data into higher-order terms (like squares or cubes), we can fit curves while still using the standard linear regression equation.

11. Which of the following is/are a theoretically possible linear regression fit? State how or why.



Answer: A, B.
A represents the standard model $f_w(x) = w_0 + w_1 x$, which fits a straight line to the data. B represents a model with higher-order features (e.g., x^2, x^3), namely the model $f_w(x) = w_0 + w_1 x + w_2 x^2 \dots$.
Note: C is impossible (vertical line is not a function of x) and D is highly unlikely (least squares would find a positive slope, not a flat line).

12. You want to extend your linear regression approach to capture a non-linear relationship by creating polynomial features (e.g., x^2, x^3, etc.).

    (a) How can this approach still be considered "linear" regression?

    It is still considered linear regression because the model remains linear in the parameters (weights). The prediction is a linear combination of the coefficients ($y = w_0 + w_1 x + w_2 x^2 \dots$), regardless of how the input features are transformed.

    (b) What potential problem might arise when adding too many polynomial terms, and how can you mitigate it?

    Adding too many terms leads to overfitting (high variance), where the model captures noise instead of the underlying trend. This can be mitigated by applying regularization or by using cross-validation to select the optimal polynomial degree.

13. Why do we split our dataset into training and test sets when building a regression model?

    We split the dataset to evaluate the model's ability to generalize to new, unseen data. By training on one subset (Training Set) and evaluating on another (Test Set), we can detect overfitting and obtain an unbiased estimate of its real-world performance.