# Manoj Kumar

**GATE AIR - 13**

**M.Tech in Data Science From IIT Guwahati**

**Expertise in Machine Learning, Deep Learning, Artificial Intelligence, Probability and Statistics**

Telegram: @Manoj_Gate_DSAI

Unacademy

CLICK

## Question 1

**What is the primary objective in determining the optimal regression line in Simple Linear Regression?**

**A.** To ensure the line passes through every data point.

**B.** To minimize the total squared deviations between the observed and predicted values.

**C.** To maximize the strength of the linear relationship as measured by the correlation coefficient.

**D.** To align the line horizontally with the x-axis.

## Question 2

**Consider a regression line $y = ax + b$, fitted from a set of numbers, where $a$ is the slope and $b$ is the intercept. If we know the value of the slope $a$ then by using which option can we always find the value of the intercept $b$?**

A) Put the value (0,0) in the regression line.

B) Put any value from the points used to fit the regression line and compute the value of $b$.

C) Put the mean values of $x$ and $y$ in the equation along with the value $a$ to get $b$.

D) None of the above.

## Question 3



Consider a dataset with values $X = [1, 2, 3, 4]^T$ and $Y = [3, 4, 8, 11]^T$. The prediction model is defined as $\hat{y} = 2x + 1$. Calculate the mean squared error (MSE) for this model and choose the correct answer from the options below:

A) 0.75

B) 1.5

C) 2.5

D) 3.0



https://t.me/gatedsai2025

## Question 4

https://t.me/gatedsai2025

Consider the dataset from a study investigating the correlation between the number of hours spent driving and the risk of developing acute back pain. The dataset includes pairs of driving hours (x) and corresponding risk scores (y) on a scale of 0-100. Fit a best-fit line to this data and using the best-fit line, estimate the risk score for 20 hours of driving. Which of the following values is closest to the predicted risk score?

| Number of Hours Spent Driving (x) | Risk Score (y) |
|---|---|
| 10 | 95 |
| 9 | 80 |
| 2 | 10 |
| 15 | 50 |
| 10 | 45 |
| 16 | 98 |
| 11 | 38 |
| 16 | 93 |

A) 79.42

B) 108.38

C) 104.38

D) 82.38

# Question 5



**Choose the correct Statements:**

1. We can get multiple local optimum solutions if we solve a linear regression problem by minimizing the sum of squared errors using gradient descent.
2. For convex loss functions (i.e., with a bowl shape), stochastic gradient descent is guaranteed to eventually converge to the global optimum.
3. For convex loss functions (i.e., with a bowl shape), batch gradient descent is guaranteed to eventually converge to the global optimum while stochastic gradient descent is not.



**A.** (1,2)
**B.** (1,3)
**C.** (1)
**D.** (3)

https://t.me/gatedsai2025

# Question 6



Given the line equation $y = \theta_0 + \theta_1 \times x$ in linear regression, what is the correct sequence of steps for applying the gradient descent algorithm to find the optimal values of theta ($\theta$)?

1. Calculate the error between the predicted values (using the line equation) and the actual data points.
2. Update the parameters theta ($\theta_0$ and $\theta_1$) based on the gradient of the error.
3. Initialize random values for the parameters theta ($\theta_0$ and $\theta_1$).
4. Repeat the process until the changes in the parameters are minimal or a specified number of iterations is reached.



Options:

A) 1, 3, 2, 4

B) 1, 2, 3, 4

C) 3, 1, 2, 4

D) 3, 2, 1, 4

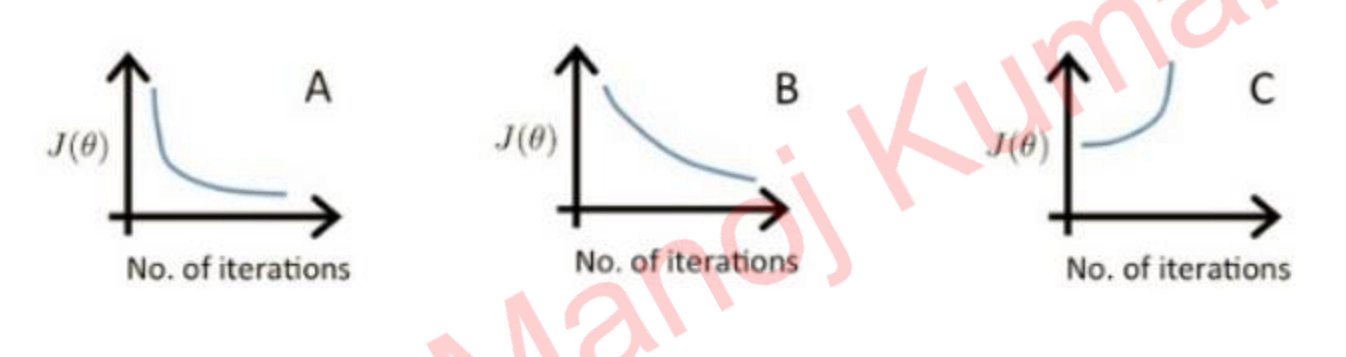
https://t.me/gatedsai2025

## Question 7



Given the three graphs labeled A, B, and C, which depict the relationship between the cost function and the number of iterations for different learning rates l1, l2, and l3 respectively, how do these learning rates compare?

A) l2 < l1 < l3
B) l1 < l2 < l3
C) l1 = l2 < l3
D) l1 = l2 = l3



**Consider the statements about the Normal Equations method for determining the coefficients in linear regression. Which of the following statements are correct?**

1. There is no need to choose a learning rate when using Normal Equations.
2. Normal Equations may become computationally intensive with a large number of features.
3. Iteration is not required when using Normal Equations.

**Options:**

**A)** Statements 1 and 2 are true.

**B)** Statements 1 and 3 are true.

**C)** Statements 2 and 3 are true.

**D)** All statements (1, 2, and 3) are true.

# Question 9

**Which of the following statements regarding the gradient descent optimization algorithm are correct? (Choose all that apply.)**

1. The weight vector is adjusted in the direction of the gradient after each iteration.
2. It is necessary to choose a learning rate that does not change during the optimization process.
3. The weight vector is adjusted in the direction opposite to the gradient after each iteration.
4. Every update to the weight vector within the gradient descent algorithm is influenced by all the training examples.

**Options:**

A) Statements 1 and 2 are correct.

B) Statements 1 and 3 are correct.

C) Statements 2 and 4 are correct.

D) Statements 3 and 4 are correct.

# Question 10[MSQ]



Let $f$ be a smooth function such that $f(\theta_0, \theta_1)$ outputs a number. Consider the use of gradient descent to minimize $f(\theta_0, \theta_1)$ with respect to $\theta_0$ and $\theta_1$, which of the following statements are true?

1. Even if the learning rate is very large, every iteration of gradient descent will decrease the value of $f(\theta_0, \theta_1)$.
2. If $\theta_0, \theta_1$ are initialized so that $\theta_0 = \theta_1$, then by symmetry (because we perform simultaneous updates to the two parameters), after one iteration of gradient descent, we will still have $\theta_0 = \theta_1$.
3. If $\theta_0, \theta_1$ are initialized at a local minimum, one iteration of gradient descent will not change their values.
4. Regardless of the initial values of $\theta_0, \theta_1$ as long as the learning rate is sufficiently small, gradient descent will converge to the same solution.

## Question 11



Given a dataset with three training instances, each instance consisting of three features and a corresponding label.

| example id | input features (x) | label (y) |
|---|---|---|
| 0 | [1, 0, 1] | 1 |
| 1 | [0, 1, 1] | 0 |
| 2 | [1, 0, 0] | 1 |

Apply the gradient descent algorithm to minimize the error, starting with the initial weight vector $w = [0, 0, 0]$. The error function $E$ is defined as:

$$E = \frac{1}{2}\sum_{i=1}^{n} E_i^2$$

where $n$ is the number of data points, and $E_i$ is the error for each instance, calculated as the difference between the true value and the predicted value. The prediction is given by $w^T x_i$ (the transpose of the weight vector $w$ multiplied by the input vector $x_i$ for that data point).

If a learning rate of 1 is used, what is the first update to the weight vector $w$ after the first iteration of the gradient descent algorithm?

A) [2, 0, 1]
B) [-1, 0, 1]
C) [3, 1, 3]
D) [1, 0, 2]

# Question 12



Given a linear regression model where the predicted output $\hat{y}$ for each instance is calculated as $\hat{y} = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n$, where $x_1, \ldots, x_n$ are the feature values of the instance, $\theta_0, \theta_1, \ldots, \theta_n$ are the model parameters, and $y$ is the true value for each instance. The loss function used is the log-cosh loss, defined as $\ell(\hat{y}, y) = \log(\cosh(\hat{y} - y))$. Here, $N$ is the total number of data points in the dataset.

How should the parameter $\theta_j$ be updated using gradient descent based on this loss function?

A) $\theta_j := \theta_j - \alpha \sum_{i=1}^{N} \tanh((\hat{y}^{(i)} - \theta^T x^{(i)}) \cdot x_j^{(i)})$

B) $\theta_j := \theta_j - \alpha \sum_{i=1}^{N} (\theta^T x^{(i)} - y^{(i)}) \cdot x_j^{(i)}$

C) $\theta_j := \theta_j - \alpha \sum_{i=1}^{N} \sinh(\theta^T x^{(i)} - y^{(i)}) \cdot x_j^{(i)}$

D) $\theta_j := \theta_j - \alpha \sum_{i=1}^{N} \tanh(\theta^T x^{(i)} - y^{(i)}) \cdot x_j^{(i)}$

## Question 13



Suppose we have a regularized linear regression model: $\mathrm{argmin}_w \|Y - Xw\|^2 + \lambda\|w\|_1$. What is the effect of increasing $\lambda$ on bias and variance?

(a) Increases bias, increases variance

(b) Increases bias, decreases variance

(c) Decreases bias, increases variance

(d) Decreases bias, decreases variance

## Question 14

Imagine you have a machine learning model that is too sensitive to small fluctuations in the training data. Which of the following strategy is needed to overcome this issue?

A) Reduce the size of the training dataset.

B) Make the model more complex.

C) Implement L1 or L2 regularization.

D) Eliminate any regularization.

## Question 15



Suppose we have a regularized linear regression model: $\text{argmin}_w \|Y - Xw\|^2 + \|w\|^p$. What is the effect of increasing $p$ (where $p \geq 1$) on bias and variance if the weights are all larger than 1?

(a) Increases bias, increases variance

(b) Increases bias, decreases variance

(c) Decreases bias, increases variance

(d) Decreases bias, decreases variance

## Question 16

Which of the following regularization methods are scale-invariant?

A. Ridge Regression

B. Lasso Regression

C. Both (a) and (b)

D. Neither (a) nor (b)

Which of the following correctly represents the formulas for bias and variance, where $\mathbb{E}$ denotes the expected value, $f(x)$ is the true function, and $\hat{f}(x)$ is the predicted function?

A. Bias = $\mathbb{E}[\hat{f}(x)] - f(x)$, Variance = $\mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]$
B. Bias = $\mathbb{E}[(\hat{f}(x) - f(x))^2]$, Variance = $\mathbb{E}[\hat{f}(x)] - f(x)$
C. Bias = $\mathbb{E}[f(x) - \hat{f}(x)]^2$, Variance = $(\mathbb{E}[\hat{f}(x)] - f(x))^3$
D. Bias = $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, Variance = $\frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$

## Question 18



Which of the following statements are true about regularization techniques such as Ridge and Lasso?

1. Regularization exclusively minimizes overfitting by significantly reducing the number of features in the model.
2. Lasso regression uniquely performs feature selection by potentially shrinking some coefficients to zero.
3. The effectiveness of regularization techniques is measured by evaluating the model's performance with the regularization term still applied to the cost function.
4. Regularization techniques lead to a decrease in model bias.

## Question 19



As the number of training examples goes to infinity, your model trained on that data will have:

A. Lower variance, high bias

B. Higher variance, lower bias

C. Lower variance, same bias

D. Same variance, same bias

## Question 20[MSQ]



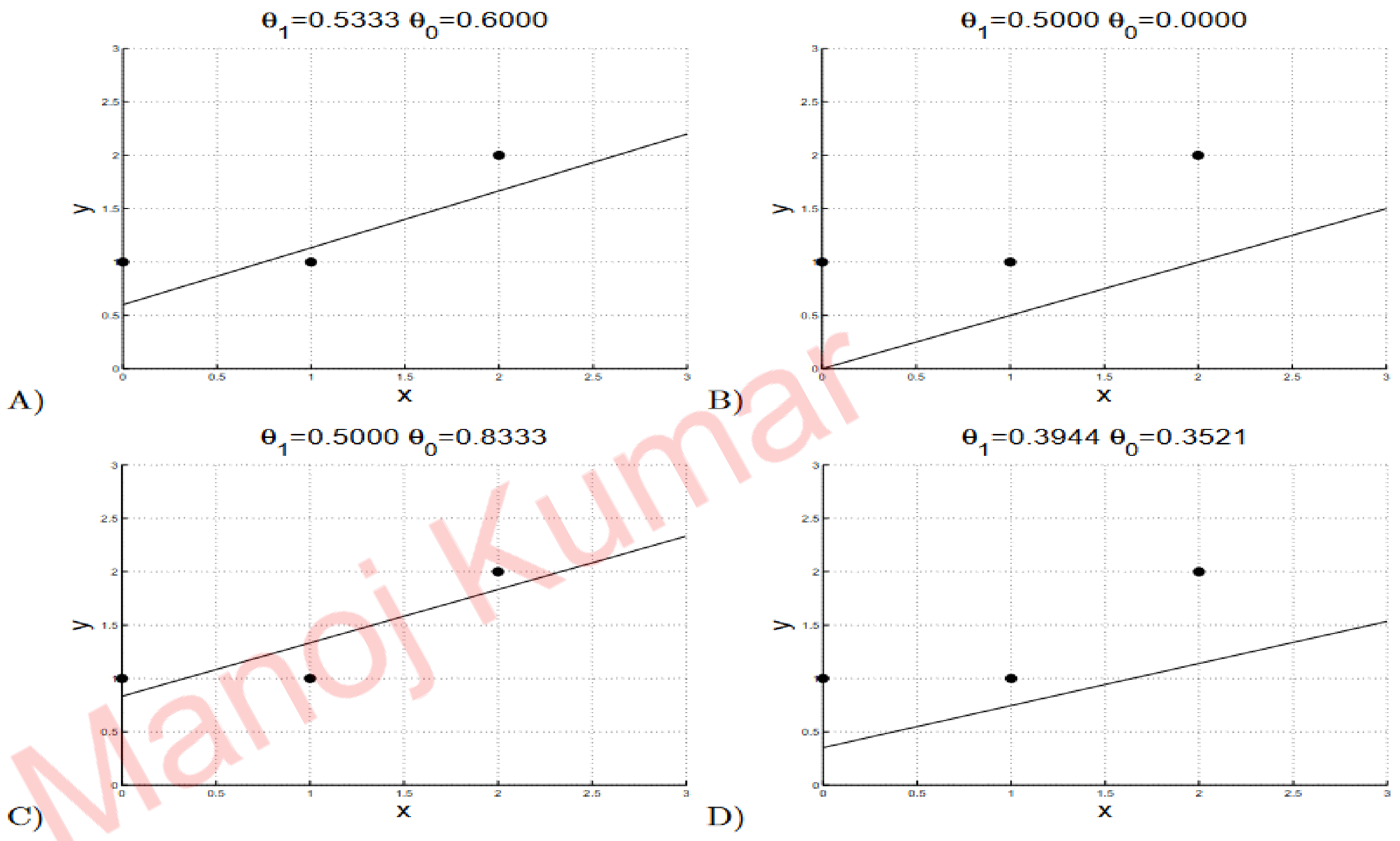
Consider a least-squares linear regression problem. Which of the following statements are true regarding the impact of adding an L2 regularization penalty?

1. Adding an L2 regularization penalty always decreases the sum of squares error of the solution on unseen test data.
2. Adding an L2 regularization penalty can decrease the sum of squares error of the solution on the training data.
3. Adding an L2 regularization penalty cannot decrease the sum of squares error of the solution on the training data.
4. Adding an L2 regularization penalty never decreases the expected sum of squares error of the solution on unseen test data.



https://t.me/gatedsai2025

## Question 6




In this problem, we are working on linear regression with regularization on points in a 2-D space. The figure below plots linear regression results based on three data points: (0,1), (1,1), and (2,2), with different regularization penalties.

**Instructions:** Match each plot in the figure to one and only one of the appropriate regularization methods listed below.

Regularization Methods:

- **1.** No regularization (or regularization parameter equals to 0):

$$\sum_{i=1}^{3}(y_i - (\theta_1 x_i + \theta_0))^2$$

- **2.** L1 regularization with $\lambda$ being 5:

$$\sum_{i=1}^{3}(y_i - (\theta_1 x_i + \theta_0))^2 + \lambda(|\theta_1| + |\theta_0|) \quad \text{where} \quad \lambda = 5$$

- **3.** L2 regularization with $\lambda$ being 1:

$$\sum_{i=1}^{3}(y_i - (\theta_1 x_i + \theta_0))^2 + \lambda(\theta_1^2 + \theta_0^2) \quad \text{where} \quad \lambda = 1$$

- **4.** L2 regularization with $\lambda$ being 5:

$$\sum_{i=1}^{3}(y_i - (\theta_1 x_i + \theta_0))^2 + \lambda(\theta_1^2 + \theta_0^2) \quad \text{where} \quad \lambda = 5$$

Consider a dataset $D_1 = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ with a linear regression model $y = w_1 x + b_1$ that minimizes the mean squared error. A second dataset $D_2$ is created by transforming $D_1$ as follows: $D_2 = \{(x^{(1)} + \alpha, y^{(1)} + \beta), \ldots, (x^{(N)} + \alpha, y^{(N)} + \beta)\}$, where $\alpha, \beta > 0$ and $w_1 \alpha \neq \beta$. A new model $y = w_2 x + b_2$ is fitted to $D_2$. Which of the following statements correctly describes the relationship between the parameters of the models trained on $D_1$ and $D_2$?

A) $w_1 = w_2, b_1 = b_2$

B) $w_1 \neq w_2, b_1 = b_2$

C) $w_1 = w_2, b_1 \neq b_2$

D) $w_1 \neq w_2, b_1 \neq b_2$



https://t.me/gatedsai2025

Question 22



Consider a dataset $D_1 = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\}$. Suppose a subset of rows from $D_1$ is duplicated to form a new dataset $D_1'$. How will the parameters $w_1$ and $b_1$ of the linear regression model $y = w_1 x + b_1$ change when fitted on the duplicated dataset $D_1'$ compared to the original dataset $D_1$?



Options:

A) $w_1 = w_2, b_1 = b_2$

B) $w_1 \neq w_2, b_1 = b_2$

C) $w_1 = w_2, b_1 \neq b_2$

D) $w_1 \neq w_2, b_1 \neq b_2$

## Question 23,24

Consider a disease that affects 1% of the population. If someone has the disease, the test correctly identifies it 95% of the time. If someone does not have the disease, the test still shows a positive result 5% of the time.

What is the chance that a test shows a positive result?

A) 2.5%

B) 5.9%

C) 10.3%

D) 25.2%

What is the chance that someone actually has the disease if their test result is positive?

A) 1%

B) 5.2%

C) 16.1%

D) 50.5%