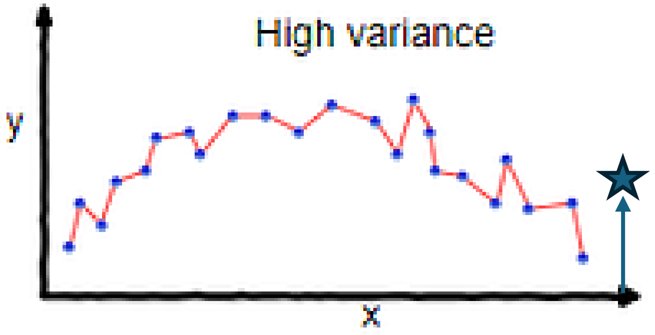
Regularization

Rina BUOY

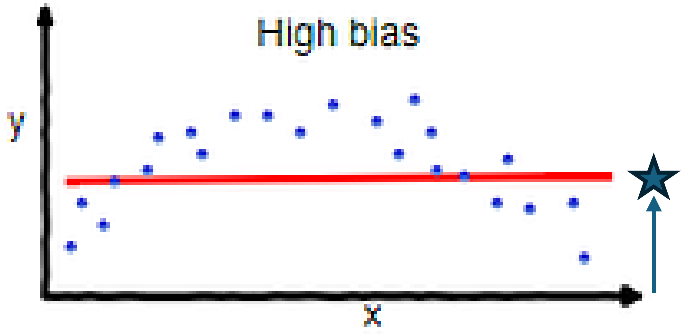
AMERICAN UNIVERSITY
OF PHNOM PENH

STUDY LOCALLY. LIVE GLOBALLY.

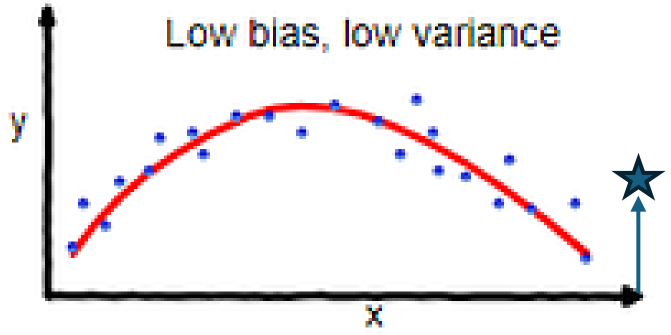
Variance-Bias Trade-off



overfitting



underfitting



Good balance

When a linear regression model exhibits high variance, it means the model is overly complex and fits the training data too closely, capturing noise and fluctuations in the data rather than the underlying trend.

Causes of High Variance

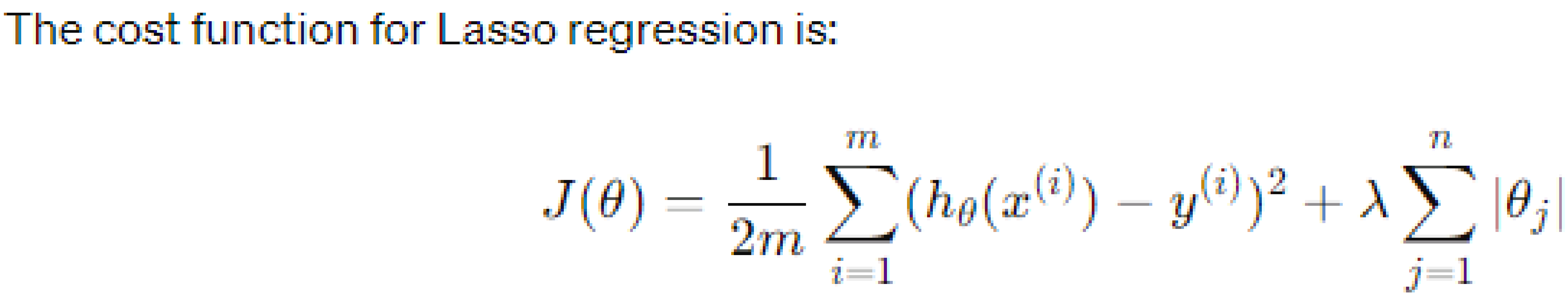
- Complex Model:
  - Using a model with a large number of features or a high polynomial degree can lead to overfitting and high variance.
- Lack of Regularization:
  - Failing to apply regularization techniques such as Ridge or Lasso regression can allow the model to overfit the training data, resulting in high variance.
- Insufficient Training Data:
  - When the training dataset is small relative to the complexity of the model, the model may memorize the training data rather than learning the underlying patterns.

Regularization

- Regularization is a technique used in machine learning to prevent overfitting by adding a penalty to the loss function.

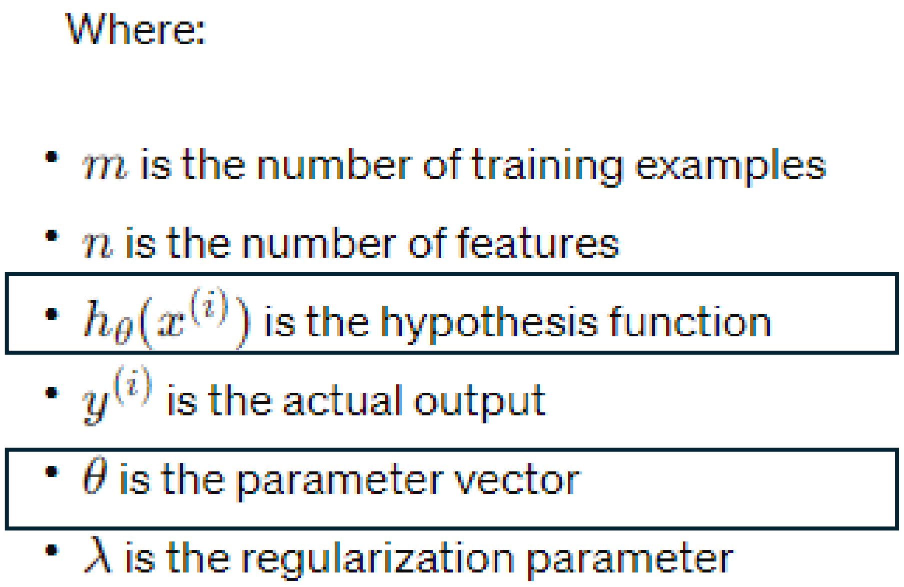
- There are two common types of regularization used in linear regression: L1 regularization (Lasso) and L2 regularization (Ridge).

L1 Regularization (Lasso Regression)



The cost function for Lasso regression is:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\theta_j|$$



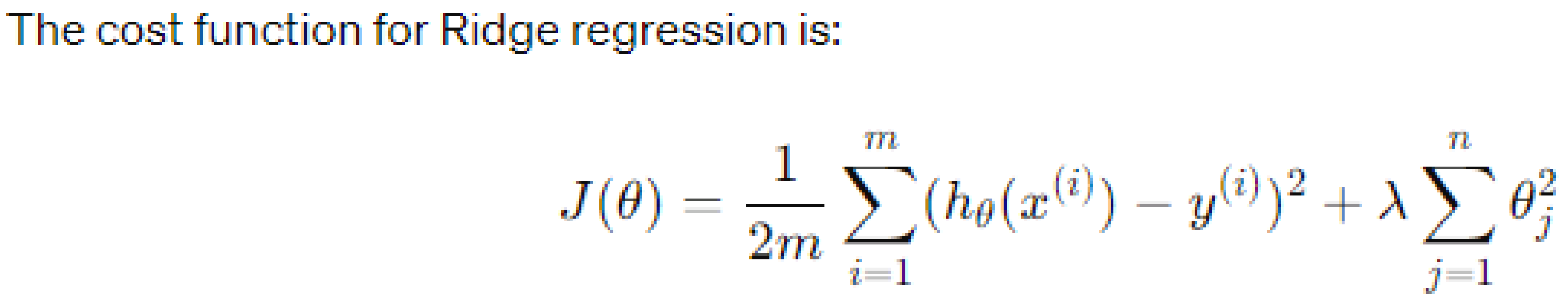
Where:

- m is the number of training examples
- n is the number of features
- $h_\theta(x^{(i)})$ is the hypothesis function
- $y^{(i)}$ is the actual output
- θ is the parameter vector
- λ is the regularization parameter

LR model

LR parameters - Beta

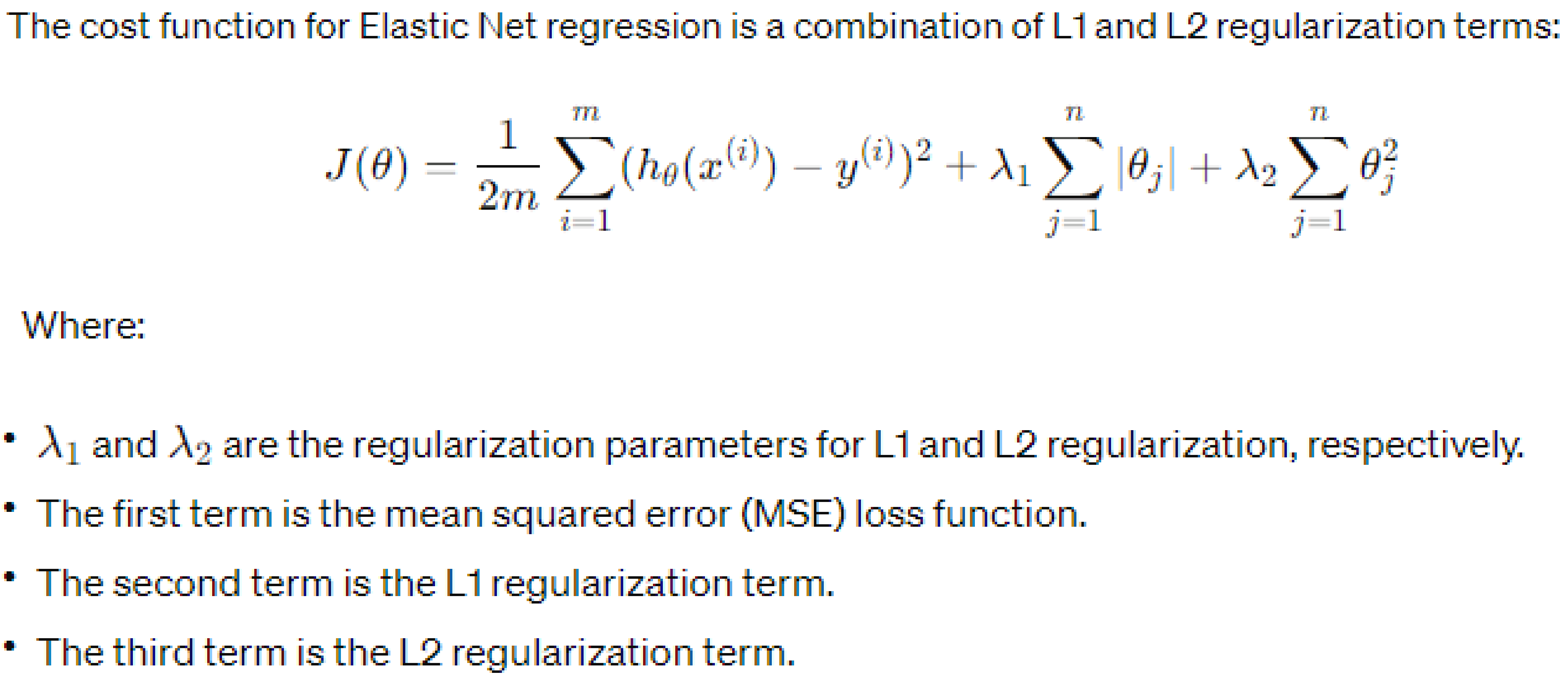
L2 Regularization (Ridge Regression)



The cost function for Ridge regression is:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2$$

Elastic Net Regression



The cost function for Elastic Net regression is a combination of L1 and L2 regularization terms:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda_1 \sum_{j=1}^{n} |\theta_j| + \lambda_2 \sum_{j=1}^{n} \theta_j^2$$

Where:

- λ_1 and λ_2 are the regularization parameters for L1 and L2 regularization, respectively.
- The first term is the mean squared error (MSE) loss function.
- The second term is the L1 regularization term.
- The third term is the L2 regularization term.

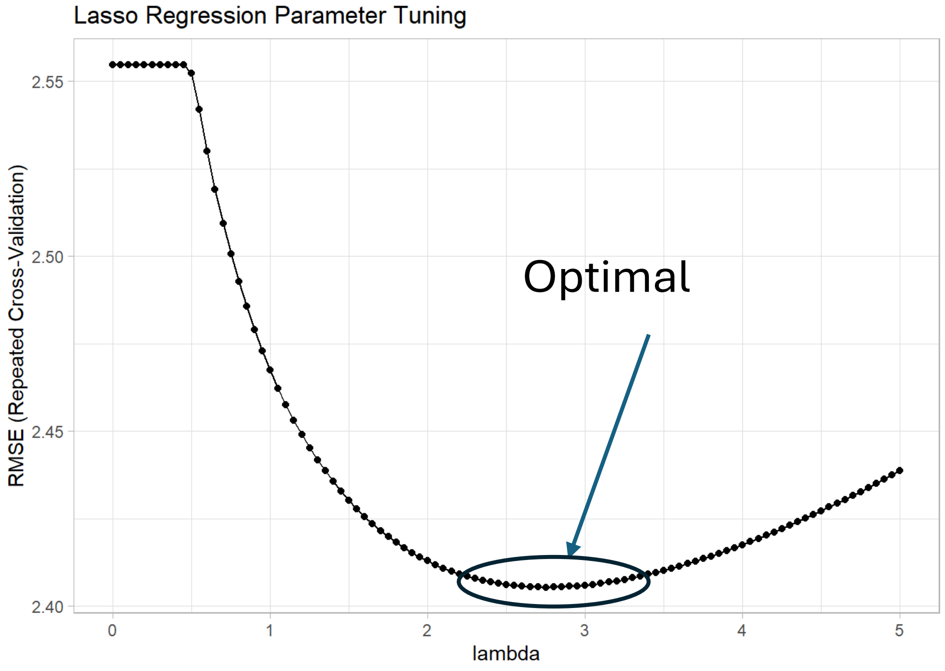
Lasso vs. Ridge Regression

- **Lasso:** Lasso tends to produce sparse models, meaning it encourages some coefficients to be exactly zero. This makes Lasso useful for feature selection as it can effectively shrink coefficients to zero and eliminate irrelevant features from the model.

- **Ridge:** Ridge regression generally does not lead to sparse models. Instead, it shrinks the coefficients towards zero but rarely forces them to be exactly zero. Ridge regression is more suitable when all features are expected to contribute to the prediction, even if some have smaller effects.

Select the Optimal Regularization Factor

1. Define a Range of Regularization Factors
2. Cross-Validation: Split your dataset into training and validation sets. Then, for each value of λ ,
   a) Fit a model on training data
   b) Evaluate performance on validation data
   c) Repeat for each λ
3. Select the optimal λ and training on the entire dataset


Lasso Regression Parameter Tuning
RMSE (Repeated Cross-Validation)
lambda
Optimal