

Note: Your TA will probably not cover all the problems. This is totally fine, the discussion worksheets are not designed to be finished within an hour. They are deliberately made slightly longer so they can serve as resources you can use to practice, reinforce, and build upon concepts discussed in lectures, labs, and homework.

Cross Validation \rightarrow pick optimal hyperparameter like λ

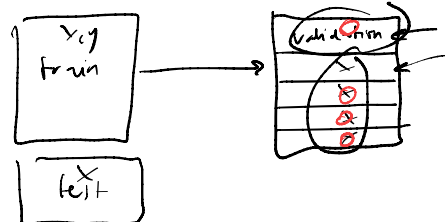
1. (a) After running 5-fold cross-validation, we get the following mean squared errors for each fold and value of λ when using Ridge regularization:

Fold Num.	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	Row Avg
1	80.2	70.2	91.2	91.8	83.4
2	76.8	66.8	88.8	98.8	82.8
3	81.5	71.5	86.5	88.5	82.0
4	79.4	68.4	92.3	92.4	83.1
5	77.3	67.3	93.4	94.3	83.0
Col Avg	79.0	68.8	90.4	93.2	

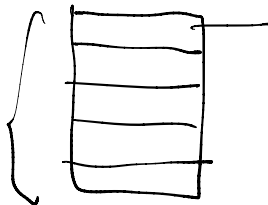
Suppose we wish to use the results of this 5-fold cross-validation to choose our hyperparameter λ , among the following four choices in the table. Using the information in the table, which λ would you choose? Why?

$$\lambda = 0.2$$

min avg MSE
across folds



- (b) You build a model with two hyperparameters, the coefficient for the regularization term (λ) and our learning rate (α). You have 4 good candidate values for λ and 3 possible values for α , and you are wondering which λ, α pair will be the best choice. If you were to perform 5-fold cross-validation, how many validation errors would you need to calculate?



$$4 \times 3 = 12 \text{ setups (combinations)}$$

$$12 \times 5 = 60$$

Ridge and LASSO Regression

Handwritten notes at the top of the page:

$$y = X\theta$$

Diagram showing matrix dimensions: $n \times (p+1)$ matrix X multiplied by $(p+1) \times 1$ vector θ equals $n \times 1$ vector y .

2. The goal of linear regression is to find the θ value that minimizes the average squared loss. In other words, we want to find $\hat{\theta}$ that satisfies the equation below:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} L(\theta) = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \|\mathbb{Y} - \mathbb{X}\theta\|_2^2$$

Here, \mathbb{X} is a $n \times (p+1)$ matrix, θ is a $(p+1) \times 1$ vector and \mathbb{Y} is a $n \times 1$ vector. Recall that the extra 1 in $(p+1)$ comes from the intercept term. As we saw in lecture, the optimal $\hat{\theta}$ is given by the closed-form expression $\hat{\theta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}$.

Normal eq

$$\hat{\theta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}$$

To prevent overfitting, we saw that we can instead minimize the sum of the average squared loss plus a regularization term $\lambda g(\theta)$. The optimization problem for such a *regularized* loss function then becomes:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} L(\theta) = \underset{\theta}{\operatorname{argmin}} \left[\frac{1}{n} \|\mathbb{Y} - \mathbb{X}\theta\|_2^2 + \lambda g(\theta) \right]$$

Handwritten note: $\hat{\theta} = (\mathbb{X}^T \mathbb{X} + n\lambda \mathbf{I})^{-1} \mathbb{X}^T \mathbb{Y}$

- If we use the function $g(\theta) = \sum_{j=1}^p \theta_j^2 = \|\theta\|_2^2$, we have “Ridge regression”. Recall that g is the ℓ_2 norm of θ , so this is also referred to as “ ℓ_2/L_2 regularization”.
- If we use the function $g(\theta) = \sum_{j=1}^p |\theta_j| = \|\theta\|_1$, we have “LASSO regression”. Recall that g is the ℓ_1 norm of θ , so this is also referred to as “ ℓ_1/L_1 regularization”.

↓ sparse θ

In this question, we intentionally choose to regularize also on the intercept term to simplify the mathematical formulation of the Ridge and LASSO regression. In practice, we would not actually want to regularize the intercept term (and you should always assume that there should **not** be a regularization on the intercept term).

For example, if we choose $g(\theta) = \|\theta\|_2^2$, our goal is to find $\hat{\theta}$ that satisfies the equation below:

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmin}} L_2(\theta) = \underset{\theta}{\operatorname{argmin}} \left[\frac{1}{n} \|\mathbb{Y} - \mathbb{X}\theta\|_2^2 + \lambda \|\theta\|_2^2 \right] \\ &= \underset{\theta}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^n (y_i - \mathbb{X}_{i,\cdot}^T \theta)^2 + \lambda \sum_{j=0}^d \theta_j^2 \right] \end{aligned}$$

Recall that λ is a hyperparameter that determines the impact of the regularization term. Like ordinary least squares, we can also find a closed-form solution to Ridge regression: $\hat{\theta} = (\mathbb{X}^T \mathbb{X} + n\lambda \mathbf{I})^{-1} \mathbb{X}^T \mathbb{Y}$. For LASSO regression, there is no such closed-form expression.

Handwritten notes at the bottom of the page:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots$$

Diagram showing the vector θ with components $\theta_0, \theta_1, \theta_2, \dots$ and the vector x with components x_1, x_2, x_3, \dots . A red circle highlights $\theta_1 x_1$ in the equation above.

- (a) Suppose we are dealing with just the OLS case (i.e., don't worry about regularization yet). We increase the complexity of the model until test error stops decreasing. If we continue to increase model complexity, what do we expect to happen to the training error of the model trained using OLS? What about the test error?

- ☒ Training error decreases ☐ Test error decreases
☐ Training error increases ☒ Test error increases

- (b) Now suppose we choose one of the above regularization methods, either L1 or L2, for some regularization parameter $\lambda > 0$ then we solve for our optimum. In terms of variance, how does a regularized model compare to ordinary least squares regression (assuming the same features between both models)?

lower model variance

- (c) Suppose we have a large number of features (10,000+), and we suspect that only a handful of features are useful. Would LASSO or Ridge regression be more helpful in interpreting useful features? Why?

LASSO \rightarrow set parameter to 0

- (d) What are the two benefits of using Ridge regression over OLS?

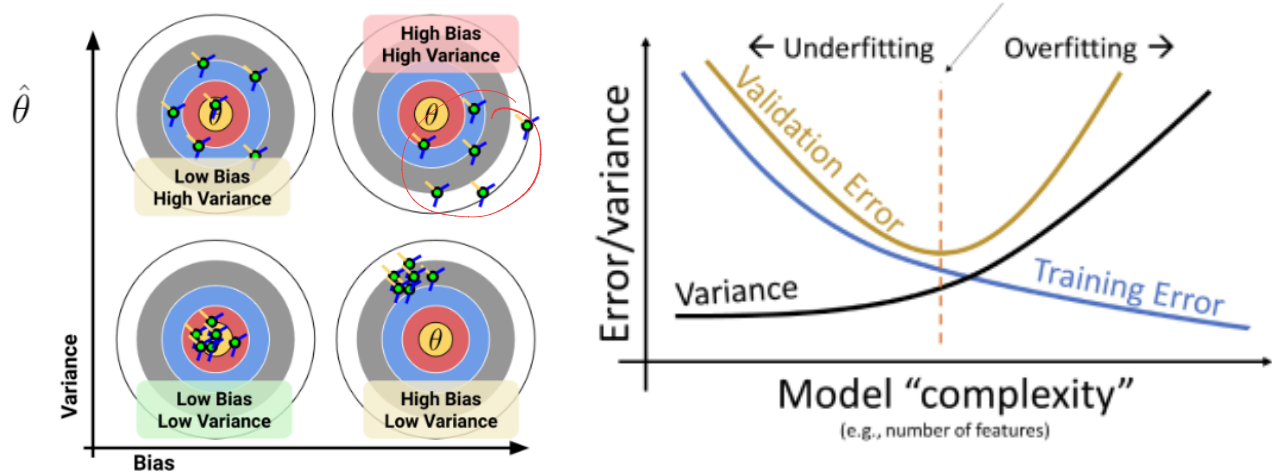
- ① $X^T X$ doesn't have to be invertible
 ② prevent overfitting

- (e) In Ridge regression, what happens to $\hat{\theta}$ if we set $\lambda = 0$? What happens as λ approaches ∞ ?

$$\hat{\theta} = (X^T X + n\lambda I)^{-1} X^T Y$$

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

$$L(\theta) = \frac{1}{2} \sum_{i=1}^n (y_i - x_i^T \theta)^2 + \lambda \sum_{i=1}^d \theta_i^2 \rightarrow \theta = \vec{0}$$



- R : The number of answers the student gets right.
 - W : The number of answers the student does not get right.
 - S : The student's score on the test.
- (a) What is the distribution of R ? Provide the name and parameters of the appropriate distribution. Explain your answer.
- (b) Find $\mathbb{E}[R]$.
- (c) True or False: $\text{SD}(R) = \text{SD}(W)$? Remember that $\text{Var}(X) = \text{SD}(X)^2$.
- (d) Find $\mathbb{E}[S]$, the student's expected score on the test.
- (e) Find $\text{SD}(S)$.

Guessing at Random

3. A multiple choice test has 100 questions, each with five answer choices. Assume for each question that there is only one correct choice. The grading scheme is as follows:

- 4 points are awarded for each right answer.
- For each other answer (wrong, missing, etc), one point is taken off; that is, -1 point is awarded.

A student hasn't studied at all and therefore selects each question's answer uniformly at random, independently of all the other questions.

Define the following random variables:

- R : The number of answers the student gets right.
- W : The number of answers the student does not get right.
- S : The student's score on the test.

(a) What is the distribution of R ? Provide the name and parameters of the appropriate distribution. Explain your answer.

$$R \sim \text{Binomial}(100, \frac{1}{5})$$

(b) Find $\mathbb{E}[R]$. $\sum x P(X=x)$

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

$$np = 100 \cdot \frac{1}{5} = \boxed{20}$$

(c) True or False: $\text{SD}(R) = \text{SD}(W)$? Remember that $\text{Var}(X) = \text{SD}(X)^2$.

$$R \sim \text{Binomial}(100, \frac{1}{5}) \quad np(1-p)$$

$$W \sim \text{Binomial}(100, \frac{4}{5}) \quad n(1-p)p$$

X, Y independent

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

(d) Find $\mathbb{E}[S]$, the student's expected score on the test.

$$R = \# \text{ Qs correct} \quad S = 4R - W$$

$$W = \# \text{ Qs incorrect}$$

$$\begin{aligned} \mathbb{E}[4R - W] \\ \mathbb{E}[4R] - \mathbb{E}[W] \\ 4\mathbb{E}[R] - \mathbb{E}[W] \\ 4 \cdot 20 - 80 \\ 80 - 80 = 0 \end{aligned}$$

(e) Find $\text{SD}(S)$.

$$W = 100 - R$$

$$\begin{aligned} \text{Var}(4R - (100 - R)) &= \text{Var}(5R) \\ \text{Var}(4R - 100 + R) &= 25 \text{Var}(R) \end{aligned}$$

$$100 \cdot \frac{1}{5} \cdot \frac{4}{5} = \left(\frac{4}{25} \cdot 100 \right)$$

$$\frac{4}{25} \cdot 100 \cdot 25 = \frac{4 \cdot 100}{1} = 400$$