Data 100, Fall 2025

Discussion #12

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.

This discussion worksheet has been modeled from the Data 100 Summer 2025 Final and the Data 100 Spring 2024 Final. It only includes questions that are in scope for the Data 100 Fall 2025 Midterm 2.

1 Summer 2025 Final: Question 2 [16 Pts]

1. CDSS wants to predict the number of students who will enroll in its courses. The CDSS course coordinators request Data 100 staff to build a variety of linear models to assist their prediction.

   (a) [2 Pts] Jake receives enrollment numbers from various past courses. He is interested in creating a constant model ($\hat{y} = \theta_0$) to make predictions. Jake wants to decide between optimizing for MSE vs MAE. Select all statements below that are true.

   - ☐ A. Jake can use gradient descent to find a value of θ_0 that minimizes MAE.
   - ☐ B. It is possible for two different constant predictions to both result in the optimal MAE.
   - ☐ C. Since the MSE loss surface is always convex and smooth, there is always one optimal $\hat{\theta}_0$ that minimizes it.
   - ☐ D. Suppose Jake chooses to optimize the MSE. The median of the data can result in a lower training loss than the training loss associated with the mean.

   (b) [3 Pts] Wesley has $n = 30$ unique data points and 2 features. He fits the following OLS model:

   $$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 (x_{i1})^3 + \hat{\theta}_2 \ln x_{i2}$$

   Wesley correctly calculates a **unique optimal solution** for his OLS model. Select all statements about Wesley's fitted model (which uses the optimal parameters) that must be true:

   - ☐ A. It is highly likely, but not guaranteed, that the sum of the residuals is 0.
   - ☐ B. The rank of the design matrix ($\mathbb{X}$) must be less than 3.
   - ☐ C. The vector of outcomes ($\mathbb{Y}$) must be in the span of $\mathbb{X}$.
   - ☐ D. If Wesley refit the same model with $(x_1)^6$ as an additional feature, a unique solution will no longer exist.

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

☐ E. If Wesley refit the same model using L2 (Ridge) regularization and $\lambda = 1$, a unique solution will still exist.

(c) [5 Pts] Sammie wants to fit and assess a couple models.

(i) [2 Pts] First, Sammie wants to find the optimal solution for θ_0 with respect to the objective function below.

$$L(\theta_0, \theta_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(\frac{\theta_0}{4\theta_1} - \theta_1 x_i\right)\right)^2$$

→ MSE convex

Find the value of θ_0 that minimizes $L(\theta_0, \theta_1)$.

$$\frac{dL}{d\theta_0} = \frac{1}{n}\sum_{i=1}^{n} 2\left(y_i - \left(\frac{\theta_0}{4\theta_1} - \theta_1 x_i\right)\right) \cdot -\frac{1}{4\theta_1} \overset{\text{set}}{=} 0 \quad \Big| \cdot \frac{n}{2} \cdot -4\theta_1$$

$$0 = \sum_{i=1}^{n}\left(y_i - \frac{\hat{\theta}_0}{4\theta_1} + \theta_1 x_i\right)$$

$$0 = \left(\sum_i y_i\right) - \frac{n\hat{\theta}_0}{4\theta_1} + \left(\sum_i \theta_1 x_i\right)$$

$$\frac{n\hat{\theta}_0}{4\theta_1} = \sum_i (y_i + \theta_1 x_i) \quad \Big| \cdot 4\theta_1$$

$$\hat{\theta}_0 = \frac{4\theta_1 \sum_i (y_i + \theta_1 x_i)}{n}$$

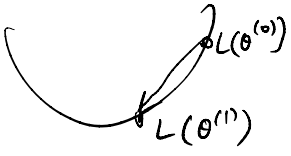
(ii) [3 Pts] Sammie chooses a new constant model specification and objective function. She calculates the gradient of her new objective function below with respect to θ_0:

$$L(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0)^2$$

$$\nabla L(\theta_0) = \left[-\frac{2}{n}\sum_{i=1}^{n} y_i - \theta_0\right]$$

Sammie asserts that $\theta_0^{(1)} = 5$, $\alpha = 1$, and $\vec{y} = \begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}y_1\\y_2\end{bmatrix}$

Using the values above, calculate $\theta_0^{(0)}$ (i.e., the initializing θ_0 value).



gradient descent:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \cdot \nabla L(\theta^{(t)})$$

$$\theta^{(t)} = \theta^{(t-1)} - \alpha \cdot \nabla L(\theta^{(t-1)})$$

$$\hookrightarrow \theta_0^{(1)} = \theta_0^{(0)} - \alpha \cdot \nabla L(\theta_0^{(0)})$$

$\theta_0^{(0)} = x$

$5 = x - 1 \cdot \left(-\frac{2}{2}\left((1-x) + (2-x)\right)\right)$

$5 = x - 1\left(-1(3-2x)\right)$

$= x - (2x - 3)$

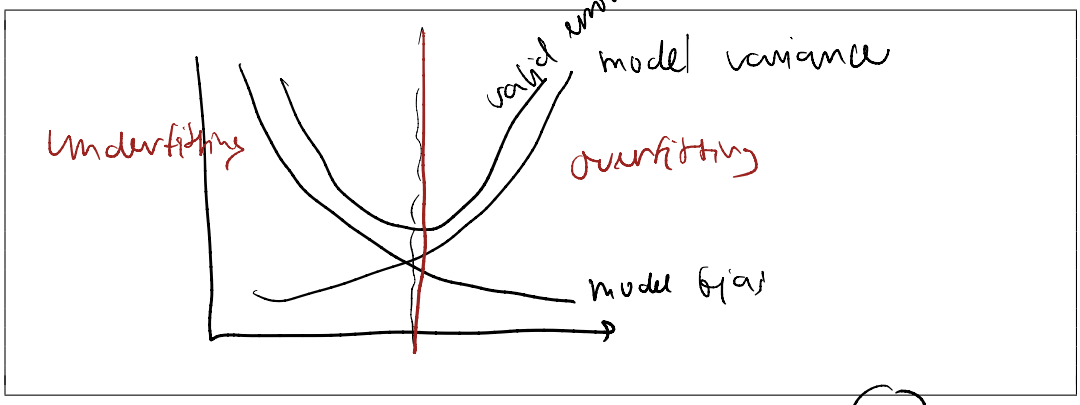
$5 = -x + 3$

$x = -2 = \theta_0^{(0)}$

(d) [6 Pts] Milena chooses a new model specification and decides to fit her chosen model with regularization.

(i) [2 Pts] Select all statements below that are true.

- ☐ A. For a fixed model specification, training dataset, and validation dataset, the optimal choice of λ hyperparameter is always the same for L1 and L2 regularization.
- ☐ B. L1 regularization typically increases model bias.
- ☐ C. L2 regularization typically increases model variance. decrease
- ☐ D. Unlike L2 regularization, L1 regression can be used to identify the most predictive features. But, L1 regularization will generally result in higher average loss. ↔ L1 worse than L2



(ii) [1 Pt] Milena chooses a special form of regularization that takes in 2 hyperparameters, λ and ρ. To pick the values for these parameters, she decides to use 7-fold cross validation. Milena tests 2 values of λ and 5 values of ρ. To estimate the optimal λ value and ρ value, how many times is each data point used to fit a model?

$2 \cdot 5 \cdot (7-1) = 60$

(iii) [2 Pt] Milena uses gradient descent to find the optimal parameters. Select all statements below that are true.

- ☐ A. With an **infinite** number of updates and a **fixed** learning rate, gradient descent will always converge to a **global** minimum.
- ☐ B. With an **infinite** number of updates and a **fixed** learning rate, gradient descent will always converge to a **local** minimum.
- ☐ C. The gradient descent algorithm can compute gradient updates using both convex and non-convex functions.
- ☐ D. Consider the average of the estimated gradients computed over one epoch of stochastic gradient descent. This value is always equal to the true gradient of the loss surface at the initial values of the model parameters.

(iv) [1 Pt] Milena decides to use mini-batch gradient descent on a dataset with 500 data points. She records that a total of 125 gradients were calculated after 5-epochs. How many data points were in each mini-batch?

SGD: take 1 point
mini batch GD: take batch of points

1 epoch = go through all n data points once

$\frac{125}{5} = 25$ gradients / epoch

$\frac{500}{25} = 20$ = size of mini batch

2 Summer 2025 Final: Question 4 [9.5 Pts]

2. A team of researchers is studying the consumption of sugary drinks of adults living in Berkeley. In the study, the team surveys 1000 adults from Berkeley at random **with replacement** and collects the follow data.

   - `drink_soda`: Whether the surveyed adult consumes at least 1 soda every day. The value is `True` if the surveyed adult consumes at least 1 soda every day, and `False` otherwise. (type: boolean)

   The team decides to estimate the proportion of adults living in Berkeley that consume at least 1 soda every day. In the problems below, you can assume that the proportion of adults living in Berkeley that consume at least 1 soda every day is p, a fixed but unknown proportion strictly between 0 and 1.

   Note: In probability notation, the `drink_soda` data collected can be viewed as i.i.d. random variables $X_1, X_2, \ldots, X_{1000}$, where each $X_i \sim \text{Bernoulli}(p)$.

   (a) [3 Pts] The team uses an estimator $\hat{\alpha}$ that is just the sample proportion. That is,

   $$\hat{\alpha} = \frac{1}{1000} \sum_{i=1}^{1000} X_i$$

   Find $\text{Bias}(\hat{\alpha})$ and $\text{Var}(\hat{\alpha})$. You answer can be in terms of p.

   The team then uses the bootstrap to analyze the estimator $\hat{\alpha}$. Recall that "a bootstrap sample" under this context means "a sample of size 1000 drawn uniformly at random with replacement from the original sample". Suppose Michael appears once in the original sample.

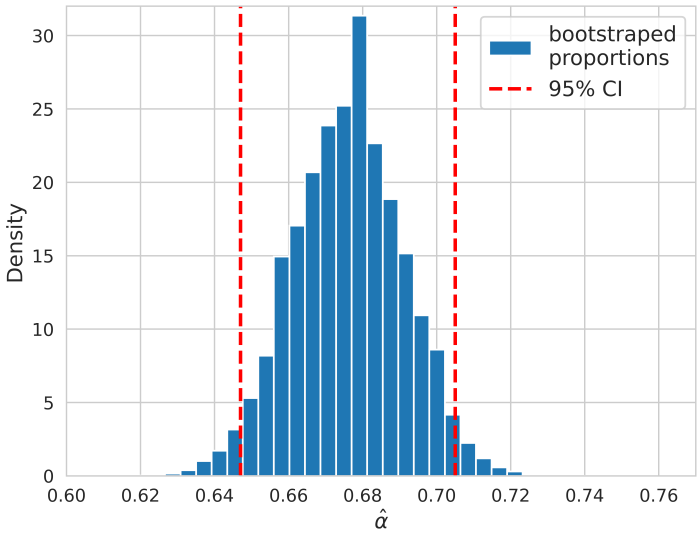
   (b) [1 Pt] The team first draws one bootstrap sample. Find the probability that Michael is drawn as the first individual of the bootstrap sample.

◯ A. 0
◯ B. $\frac{1}{1000}$
◯ C. $\frac{1}{2}$
◯ D. $\frac{999}{1000}$
◯ E. 1

(c) [1 Pt] Let m be the correct answer to (b). Find the probability that Michael does not appear in the bootstrap sample at all. Your answer can be in terms of m.

◯ A. 0
◯ B. m^{1000}
◯ C. $(1-m)^{1000}$
◯ D. $\left(\frac{m}{2}\right)^{1000}$
◯ E. 1

The team then draws 100,000 bootstrap samples and computes $\hat{\alpha}$ on each sample. The sample proportions are plotted in the histogram below. The two dashed lines represent the lower and upper bounds of the 95% confidence interval for p computed on the bootstrap samples.



(d) [1.5 Pts] Suppose $\hat{\alpha} = 0.67$ for the original sample. Using this information and the histogram, which of the following **must be true**?

◯ True ◯ False This histogram is an approximation of the distribution of X_1 (or any one of $X_1, \ldots, X_{1000}$ since they are i.i.d.).

◯ True ◯ False The two-sided null hypothesis "the proportion of adults in Berkeley that consume at least 1 soda everyday is 72%" is rejected at the 5% significance level.

◯ True ◯ False The two-sided null hypothesis "the proportion of adults in Berkeley that consume at least 1 soda everyday is 72%" is rejected at the 10% significance level.

After this background study, the team goes on to explore the relationship between consumption of sugary drinks and blood sugar level. When the team collected the data on `drink_soda`, it also collected the following data on the same individuals.

- `blood_sugar`: The pre-breakfast blood sugar level measured by the research team in the unit of millimoles per liter. (type: `np.float`)
- `t2d`: An indicator of whether the individual has type 2 diabetes, where a value of 1 indicates type 2 diabetes, and 0 indicates no type 2 diabetes. (type: `int`)

To study the relationship, the team fit the following linear regression model over the data

$$\widehat{\text{blood\_sugar}} = \theta_0 + \theta_1 \times \text{drink\_soda} + \theta_2 \times \text{t2d}$$

Suppose the optimal estimated parameters are $\hat{\theta}_0 = 5.1$, $\hat{\theta}_1 = 2.8$, and $\hat{\theta}_2 = 3.8$.

(f) [1 Pt] What is the interpretation of $\hat{\theta}_0$? Explain in one sentence.

(g) [1 Pt] What is the interpretation of $\hat{\theta}_1$? Explain in one sentence.

(h) [1 Pt] What is the interpretation of $\hat{\theta}_2$? Explain in one sentence.

3 Spring 2024 Final: Question 4b-d (adapted)

3. Jessica is a cowgirl who wants to predict how much food she should give to her cows. Every day, she records data from her cattle in a `DataFrame` called `cows`.

	date	name	weight	food
0	May 9th	Angus	2210	25
1	May 9th	Butters	2503	28
2	May 9th	C.R.E.A.M.	3024	38
3	May 8th	Angus	2207	26
4	May 8th	Butters	2501	30

```
cows.head()
```

Jessica builds a simple linear regression model to predict how much `food` to give her cows. Jessica decides to use LASSO regression to prevent overfitting her training data.

(a) [2 Pts] If Jessica has 5 candidate values of regularizer parameter λ, how many validation errors will she calculate if she runs 4-fold cross-validation.

(b) [2 Pts] Suppose Jessica has a design matrix $\mathbb{X}$ with an unknown number of rows and features and a target variable $\mathbb{Y}$ which stores the data in the `food` column. For each subpart, pick the **best** choice for each given scenario.

(i) Jessica wants to have unimportant features of $\mathbb{X}$ have a corresponding weight more likely set to 0. What should Jessica do?

◯ A. Use OLS to predict $\mathbb{Y}$ from $\mathbb{X}$

◯ B. Use ridge regression to predict $\mathbb{Y}$ from $\mathbb{X}$

◯ C. Use LASSO regression to to predict $\mathbb{Y}$ from $\mathbb{X}$

(ii) Jessica trains a model using LASSO regression and finds that the training error is low, but validation error is high. What should Jessica do?

◯ A. Increase λ

◯ B. Decrease λ

◯ C. Set $\lambda = 0$