UNIVERSITY OF ALBERTA
CMPUT 267 Fall 2024

Midterm Exam 2

Do Not Distribute

Duration: 75 minutes

Last Name: _____

First Name: _____

Carefully read all of the instructions and questions. Good luck!

---

1. **Do not turn this page** until instructed to begin.
2. Verify that your exam package includes 21 pages, along with a formula sheet and a blank page at the end.
3. **Only the scantron will be marked**. All of your answers must be clearly marked on the scantron.
4. Use **pencil only** to fill out the scantron (preferably an HB or #2 pencil).
5. **Erase mistakes completely** on the scantron to avoid misreading by the scanner.
6. **Mark answers firmly and darkly**, filling in the bubbles completely.
7. This exam consists of **40 questions**. Each question is worth **1 mark**. The exam is worth a total of **40 marks**.
8. Some questions may have **multiple correct answers**. To receive **full marks**, you must select **all correct answers**. If you select only **some** of the correct answers, you will receive **partial marks**. Selecting an incorrect option will cancel out a correct one. For example, if you select two answers—one correct and one incorrect—you will receive zero points for that question. If the number of incorrect answers exceeds the correct ones, your score for that question will be zero. **No negative marks** will be given.

Question 1. [1 MARK]

Consider the predictor $f(x) = xw$, where $w \in \mathbb{R}$ is a one-dimensional parameter, and x represents the feature with no bias term. Suppose you are given a dataset of n data points $\mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n))$, where each y_i is the target variable corresponding to feature x_i. You define a new regularized estimated loss as follows:

$$\hat{L}_\lambda(w) = \frac{1}{n}\sum_{i=1}^{n} c_i(x_i w - y_i)^4 + \frac{\lambda}{n}w^4$$

where $\lambda \geq 0$ is the regularization parameter and $c_i \geq 0$ weights the importance of each data point. In this question, we are interested in finding $\hat{w} = \arg\min_{w\in\mathbb{R}} \hat{L}_\lambda(w)$ using first-order gradient descent. Which of the following statements are true?

A. The first-order gradient update rule with a fixed step size is

$$w^{(t+1)} = w^{(t)} - \frac{2}{n}\eta^{(t)}\left[\sum_{i=1}^{n} c_i x_i(x_i w^{(t)} - y_i) + \lambda w^{(t)}\right].$$

B. The first-order gradient update rule with a fixed step size is

$$w^{(t+1)} = w^{(t)} - \frac{4}{n}\eta^{(t)}\left[\sum_{i=1}^{n} c_i x_i(x_i w^{(t)} - y_i)^3 + \lambda (w^{(t)})^3\right].$$

C. If you run first-order gradient descent for T epochs, you are not guaranteed that $w^{(T)} = \hat{w}$.

D. If you use a fixed step size that is too large, then first-order gradient descent can diverge. This means that as the number of epochs increases, the value of $\hat{L}(w^{(t)})$ will increase.

Solution 1. Correct answer(s): A, C, D

Explanation:

A. **True.** The gradient of the regularized loss function $\hat{L}_\lambda(w)$ with respect to w is:

$$\nabla \hat{L}_\lambda(w) = \frac{2}{n}\left(\sum_{i=1}^{n} x_i(x_i w - y_i) + \lambda w\right)$$

The gradient descent update rule is:

$$w^{(t+1)} = w^{(t)} - \eta^{(t)}\nabla \hat{L}_\lambda(w^{(t)})$$

Substituting the gradient:

$$w^{(t+1)} = w^{(t)} - \frac{2}{n}\eta^{(t)}\left[\sum_{i=1}^{n} x_i(x_i w^{(t)} - y_i) + \lambda w^{(t)}\right]$$

B. **False.** This update rule omits the regularization term $\lambda w^{(t)}$ in the gradient. Therefore, it corresponds to the gradient of the unregularized loss function.

C. **True.** Gradient descent is an iterative optimization algorithm that approaches the minimum $\hat{w}$ asymptotically, that is as t goes to infinity.

D. **True.** A step size (learning rate) that is too large can cause the updates to overshoot the minimum, leading to divergence. In such cases, the loss function $\hat{L}(w^{(t)})$ can indeed increase with each iteration.

Question 2. [1 MARK]

Let everything be defined as in the previous question. Suppose that we are now interested in using second-order gradient descent to find $\hat{w}$. Which of the following statements is true?

A. The second derivative of the regularized loss function $\hat{L}_\lambda(w)$ with respect to w is:

$$\hat{L}''_\lambda(w) = \frac{2}{n}\sum_{i=1}^{n} x_i^2$$

B. The second derivative of the regularized loss function $\hat{L}_\lambda(w)$ with respect to w is:

$$\hat{L}''_\lambda(w) = \frac{2}{n}\sum_{i=1}^{n} x_i^2 + \lambda$$

C. If we run second order gradient descent for $T = 1000$ epochs, we are guaranteed that $w^{(1000)} = \hat{w}$.

D. The second-order gradient descent update rule is the same as the first-order gradient descent update rule if the step size is $\eta^{(t)} = \frac{1}{\hat{L}''_\lambda(w^{(t)})}$.

Solution 2. Correct answer(s): D

Explanation:

A. **False.** The second derivative of the regularized loss function $\hat{L}_\lambda(w)$ is:

$$\hat{L}''_\lambda(w) = \frac{2}{n}\sum_{i=1}^{n} x_i^2 + \frac{2\lambda}{n}$$

This includes the contribution from the regularization term $\frac{\lambda}{n}w^2$, whose second derivative is $\frac{2\lambda}{n}$. Therefore, the expression in option A is incomplete as it omits the regularization term.

B. **False.** While option B includes the regularization term, it incorrectly adds λ instead of $\frac{2\lambda}{n}$. The correct second derivative should be:

$$\hat{L}''_\lambda(w) = \frac{2}{n}\sum_{i=1}^{n} x_i^2 + \frac{2\lambda}{n}$$

Thus, the expression in option B has an incorrect coefficient for the regularization term.

C. **False.** Running second-order gradient descent for a finite number of epochs $T = 1000$ does not guarantee that $w^{(1000)} = \hat{w}$. While second-order methods can converge faster than first-order methods, especially near the optimum, exact convergence within a finite number of steps is not guaranteed unless specific conditions are met, which are not stated here.

D. **True.** In second-order gradient descent, the update rule is:

$$w^{(t+1)} = w^{(t)} - \eta^{(t)} \left(\hat{L}'_\lambda(w^{(t)}) \right)$$

where $\eta^{(t)} = \frac{1}{\hat{L}''_\lambda(w^{(t)})}$. This effectively scales the gradient by the inverse of the second derivative, making the update rule analogous to first-order gradient descent update rule.

Question 3. [1 MARK]

Let everything be defined as in the previous two questions. Suppose that we are now interested in finding a closed-form solution for $\hat{w}$. Which of the following statements is true?

A. The closed-form solution for $\hat{w}$ is $\frac{2}{n} \left(\sum_{i=1}^n x_i^2 + \lambda \right)^{-1} \sum_{i=1}^n x_i y_i$

B. We have not learned in class how to check if the estimated loss $\hat{L}$ is convex.

C. The estimated loss $\hat{L}$ is not convex.

D. The closed-form solution for $\hat{w}$ is $\left(\sum_{i=1}^n x_i^2 + \lambda \right)^{-1} \sum_{i=1}^n x_i y_i$

Solution 3. Correct answer(s): D

Explanation:

A. **False.** The correct closed-form solution for ridge regression does not include the factor $\frac{2}{n}$. The standard closed-form solution is derived by minimizing the regularized loss function, which leads to:

$$\hat{w} = \left(\sum_{i=1}^n x_i^2 + \lambda \right)^{-1} \sum_{i=1}^n x_i y_i$$

Therefore, the expression in statement A is incorrect due to the presence of the extra factor $\frac{2}{n}$.

B. **False.** We can check if it is convex by checking if the second derivative is non-negative.

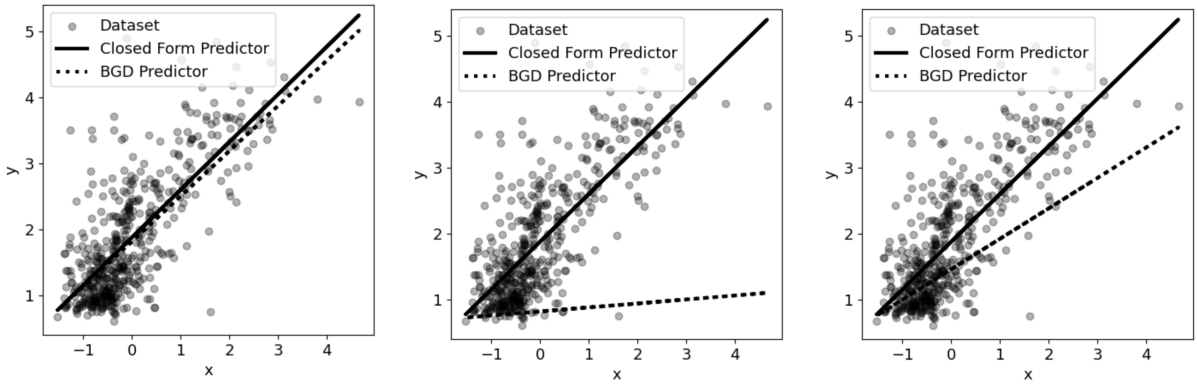
C. **False.** The estimated loss $\hat{L}$ in ridge regression is convex because the second derivative is non-negative.

D. **True.** This statement correctly presents the standard closed-form solution for ridge regression. By setting the derivative of the regularized loss function to zero and solving for w, we obtain:

$$\hat{w} = \left(\sum_{i=1}^n x_i^2 + \lambda \right)^{-1} \sum_{i=1}^n x_i y_i$$

Question 4. [1 MARK]

You are working on a linear regression problem with $d = 1$ feature. You obtain the closed-form predictor by using a closed-form learner. You also obtain a batch gradient descent (BGD) predictor by using a BGD learner. For BGD, you choose a step size such that you are sure the gradient steps do not diverge. You run BGD for: 10, 50, and 150 epochs, obtain a predictor for each, and plot them along with the closed-form predictor. Which of the following statements are true?



(a) Fig 1 (b) Fig 2 (c) Fig 3

A. The BGD predictor after 10, 50, and 150 epochs is shown in plots Fig 1, Fig 3, and Fig 2, respectively.

B. The BGD predictor after 10, 50, and 150 epochs is shown in plots Fig 2, Fig 3, and Fig 1, respectively.

C. The BGD predictor after 10, 50, and 150 epochs is shown in plots Fig 1, Fig 2, and Fig 3, respectively.

D. The estimated loss for the closed-form predictor is always less than or equal to the estimated loss for the BGD predictor for any number of epochs.

Solution 4. Correct answer(s): B, D

Explanation:

A. **False.**

Statement A claims that the BGD predictors after 10, 50, and 150 epochs correspond to Fig 1, Fig 3, and Fig 2, respectively. However, Fig 2 shows the BGD predictor after 10 epochs (farthest from the closed-form predictor), Fig 3 after 50 epochs (closer), and Fig 1 after 150 epochs (closest). Therefore, the mapping in Statement A is incorrect.

B. **True.**

Statement B correctly identifies the association between the number of epochs and the corresponding figures:

- After 10 epochs: Fig 2 (farthest from the closed-form predictor)
- After 50 epochs: Fig 3 (closer to the closed-form predictor)

- After 150 epochs: Fig 1 (closest to the closed-form predictor)

This alignment accurately reflects the progression of the BGD predictor approaching the closed-form solution as the number of epochs increases. Thus, Statement B is true.

C. **False.**

Statement C incorrectly maps the epochs to the figures by stating that after 10, 50, and 150 epochs correspond to Fig 1, Fig 2, and Fig 3, respectively. As established, Fig 1 corresponds to 150 epochs, Fig 2 to 10 epochs, and Fig 3 to 50 epochs. Therefore, the mapping in Statement C is incorrect.

D. **True.**

Statement D asserts that the estimated loss for the closed-form predictor is always less than or equal to the estimated loss for the BGD predictor for any number of epochs. This is true because the closed-form predictor provides the global minimum of the loss function. In contrast, the BGD predictor approaches this minimum asymptotically as the number of epochs increases. For any finite number of epochs, the BGD predictor may not yet have fully converged to the minimum loss achieved by the closed-form predictor. Therefore, the estimated loss for the BGD predictor is always at least as large as the estimated loss for the closed-form predictor.

Question 5. [1 MARK]

The binomial distribution is a discrete probability distribution that models the number of heads in n independent flips of a coin with probability θ of landing heads. The pmf of the binomial distribution is given by:

$$p(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k},$$

where k is the number of heads, n is the number of flips, θ is the probability of heads, $\binom{n}{k} = \frac{n!}{(n-k)!k!}$, and $a! = a \times (a-1) \times \ldots \times 1$ is the factorial function.
Now suppose we have data $D = (X_1, X_2, X_3) = (2, 2, 1)$ where each X_i is independently drawn from the same binomial distribution with $n = 4$. We want to estimate the probability of heads θ using the maximum likelihood estimation (MLE) method.
Which of the following statements are true?

A. The likelihood function is given by $6^2 \cdot 4 \cdot \theta^5(1-\theta)^7$.

B. The likelihood function is given by $6 \cdot 4 \cdot \theta^5(1-\theta)^7$.

C. The likelihood function is given by $6^2 \cdot 4 \cdot \theta^4(1-\theta)^7$.

D. The likelihood function is given by $6^2 \cdot 4 \cdot \theta^5(1-\theta)^6$.

Solution 5. **Correct answer(s): A**

Explanation:

A. **True.** The likelihood function is the product of the pmf evaluated at each data point. For the data $\mathcal{D} = (2, 2, 1)$:

$$p(2) \cdot p(2) \cdot p(1)$$

Substituting the pmf:

$$\left[\binom{4}{2}\theta^2(1-\theta)^2\right]^2 \cdot \left[\binom{4}{1}\theta(1-\theta)^3\right]$$

Calculating the binomial coefficients:

$$\binom{4}{2} = 6 \quad \text{and} \quad \binom{4}{1} = 4$$

Therefore likelihood function is:

$$(6\theta^2(1-\theta)^2)^2 \cdot (4\theta(1-\theta)^3) = 6^2 \cdot 4 \cdot \theta^5(1-\theta)^7$$

This matches statement A.

B. **False.** Statement B incorrectly omits the exponent on the first binomial coefficient. The correct likelihood should include 6^2, not just 6.

C. **False.** Statement C incorrectly lowers the exponent of θ from 5 to 4. The correct exponent for θ is 5, as derived above.

D. **False.** Statement D incorrectly lowers the exponent of $(1-\theta)$ from 7 to 6. The correct exponent for $(1-\theta)$ is 7, as derived above.

Question 6. [1 MARK]

Let everything be defined as in the previous question. Recall the logarithm property that $\log(x^a) = a\log(x)$. Which of the following statements are true?

A. The maximum likelihood estimate of θ is $\theta_{\text{MLE}} = \frac{5}{6}$.

B. The maximum likelihood estimate of θ is $\theta_{\text{MLE}} = \frac{5}{12}$.

C. The maximum likelihood estimate of θ is $\theta_{\text{MLE}} = \frac{7}{12}$.

D. The maximum likelihood estimate of θ is $\theta_{\text{MLE}} = \frac{7}{6}$.

Solution 6. Correct answer(s): B

Explanation:

A. **False.** Incorrectly states the MLE of θ as $\frac{5}{6}$. Based on the calculation below, the correct MLE is $\frac{5}{12}$, not $\frac{5}{6}$.

B. **True.** To find the maximum likelihood estimate (MLE) of θ, we use the likelihood function from the previous question:

$$g(\theta) = 6^2 \cdot 4 \cdot \theta^5(1-\theta)^7$$

To find the MLE, we take the natural logarithm of the likelihood function (log-likelihood):

$$\log(g(\theta)) = 2\log(6) + \log(4) + 5\log(\theta) + 7\log(1-\theta)$$

Taking the derivative with respect to θ and setting it to zero:

$$\frac{d}{d\theta}\log g(\theta) = \frac{5}{\theta} - \frac{7}{1-\theta} = 0$$

Solving for θ:

$$\frac{5}{\theta} = \frac{7}{1-\theta} \quad \Rightarrow \quad 5(1-\theta) = 7\theta \quad \Rightarrow \quad 5 - 5\theta = 7\theta \quad \Rightarrow \quad 5 = 12\theta \quad \Rightarrow \quad \theta = \frac{5}{12}$$

Therefore, the maximum likelihood estimate of θ is $\frac{5}{12}$.

C. **False.** Incorrectly suggests the MLE of θ is $\frac{7}{12}$. This does not align with the derived value of $\frac{5}{12}$.

D. **False.** Incorrectly states the MLE of θ as $\frac{7}{6}$, which is greater than 1 and thus not a valid probability.

Question 7. [1 MARK]

Suppose we have a coin with an unknown probability of landing heads, denoted by θ. We place a Beta prior distribution on $\theta \in (0,1)$, such that $\theta \sim \text{Beta}(\alpha, \beta)$, where the pdf of the Beta distribution is given by:

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

where $B(\alpha,\beta)$ is the Beta function and is a constant. After observing n coin flips, we see k heads and $n-k$ tails. The likelihood function is given by $\theta^k(1-\theta)^{n-k}$. What is the posterior distribution of θ given this data? Here $\propto$ means proportional to, that is, excluding the constants.

A. $p(\theta \mid \mathcal{D}) \propto \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}$

B. $p(\theta \mid \mathcal{D}) \propto \theta^{\alpha+n-k-1}(1-\theta)^{\beta+k-1}$

C. $p(\theta \mid \mathcal{D}) \propto \theta^{\alpha+k}(1-\theta)^{\beta+n-k}$

D. $p(\theta \mid \mathcal{D}) \propto \theta^{\alpha+n-1}(1-\theta)^{\beta+k-1}$

Solution 7. Correct answer(s): A

Explanation: The posterior distribution is proportional to the product of the likelihood function and the prior distribution. Therefore, the posterior distribution is given by:

$$p(\theta \mid X) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \theta^k(1-\theta)^{n-k} = \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}$$

Question 8. [1 MARK]

Let everything be defined as in the previous question. Which of the following statements are true?

A. If α and β are both small compared to n and k, then the posterior distribution is almost the same as the prior distribution.

B. If α and β are both large compared to n and k, then the posterior distribution is almost the same as the prior distribution.

C. If $\alpha = 1$ and $\beta = 1$, then the posterior distribution is proportional to the likelihood function.

D. If $\alpha = 1$ and $\beta = 1$, then the posterior distribution is not proportional to the likelihood function.

Solution 8. Correct answer(s): B, C

Explanation:

A. **False.** This is false since the opposite is true, as explained below.

B. **True.** When both α and β are large relative to k and n, then the terms $\alpha + k$ and $\beta + n - k$ in the posterior distribution are approximately equal to α and β, respectively, which gives the prior.

C. **True.** Setting $\alpha = 1$ and $\beta = 1$ defines a uniform prior distribution (Beta$(1, 1)$), which assigns equal probability to all possible values of θ in the interval $[0, 1]$. In this case, the posterior distribution becomes:

$$\theta \mid D \sim \text{Beta}(1 + k, 1 + n - k) = \text{Beta}(k + 1, n - k + 1)$$

Since the prior is uniform, the posterior distribution is directly proportional to the likelihood function:

$$p(\theta \mid D) \propto p(D \mid \theta) \times p(\theta) = p(D \mid \theta) \times 1 = p(D \mid \theta)$$

Therefore, the posterior distribution is indeed proportional to the likelihood function when $\alpha = 1$ and $\beta = 1$.

D. **False.** This statement directly contradicts statement 3. As established, when $\alpha = 1$ and $\beta = 1$, the posterior distribution is proportional to the likelihood function. Therefore, statement 4 is false.

Question 9. [1 MARK]

In Lasso regression, the data is assumed to be generated from a Gaussian distribution with mean $\mathbf{x}^T\mathbf{w}$ and variance 1. The prior distribution on the weights w_j is a Laplace distribution, with pdf given by $p(w_j) = \frac{\lambda}{2}\exp(-\lambda|w_j|)$, for $j = 1, \ldots, d$ where $\lambda \geq 0$ is the regularization parameter. The bias term w_0 is assumed to be generated from a uniform distribution, with pdf given by $p(w_0) = \frac{1}{2a}$ with a very large a. All the weights $w_0, w_1, \ldots, w_d$ are independent. The MAP estimate of the weights is given by

$$\mathbf{w}_{\text{MAP}} = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\mathbf{w})^2 - \log(p(\mathbf{w})) \right\}.$$

What is the $\log(p(\mathbf{w}))$ term above equal to?

A. $\sum_{j=1}^{d}\left(\log\left(\frac{\lambda}{2}\right) - \lambda|w_j|\right) + \log\left(\frac{1}{2a}\right)$

B. $\sum_{j=1}^{d} -\lambda|w_j| + \log\left(\frac{1}{2a}\right)$

C. $\sum_{j=1}^{d} -\lambda|w_j|$

D. $\sum_{j=1}^{d} -\lambda|w_j| + d\log\left(\frac{\lambda}{2}\right) + \log\left(\frac{1}{2a}\right)$

Solution 9. Correct answer(s): A, D

Explanation:
To determine $\log(p(\mathbf{w}))$, we consider the prior distributions of all weights $w_0, w_1, \ldots, w_d$:
Bias Term w_0:

$$p(w_0) = \frac{1}{2a}$$

Taking the logarithm:

$$\log(p(w_0)) = \log\left(\frac{1}{2a}\right)$$

Weights w_j for $j = 1, \ldots, d$:

$$p(w_j) = \frac{\lambda}{2}\exp(-\lambda|w_j|)$$

Taking the logarithm:

$$\log(p(w_j)) = \log\left(\frac{\lambda}{2}\right) - \lambda|w_j|$$

Combined Log-Prior for All Weights: Since all weights are independent, the log-prior of the entire weight vector $\mathbf{w}$ is the sum of the log-priors of each individual weight:

$$\log(p(\mathbf{w})) = \log(p(w_0)) + \sum_{j=1}^{d}\log(p(w_j))$$

Substituting the expressions from above and simplifying:

$$\begin{aligned}\log(p(\mathbf{w})) &= \sum_{j=1}^{d}\left(\log\left(\frac{\lambda}{2}\right) - \lambda|w_j|\right) + \log\left(\frac{1}{2a}\right) \\ &= \sum_{j=1}^{d} -\lambda|w_j| + d\log\left(\frac{\lambda}{2}\right) + \log\left(\frac{1}{2a}\right)\end{aligned}$$

Question 10. [1 MARK]

Let the dataset be $\mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n))$, the mini-batch size $b \in \mathbb{N}$, and $M = \text{floor}(n/b)$. In class we learned about mini-batch gradient descent. However, if the size of the dataset n was not divisible by the mini-batch size b, then we discarded the last batch of data. In this question we are interested in developing a mini-batch gradient descent algorithm that uses all the data points. Which of the following statements are true?

A. There are always M mini-batches.

B. If n is divisible by b then there are M mini-batches.

C. There are always $M + 1$ mini-batches.

D. If n is not divisible by b then the size of the last mini-batch is $n - Mb$.

Solution 10. **Correct answer(s): B, D**

Explanation:

A. **False.**

If the dataset size n is not divisible by the mini-batch size b, then using only $M = \lfloor n/b \rfloor$ mini-batches would leave out the remaining $n - Mb$ data points. Therefore, there cannot always be exactly M mini-batches when n is not divisible by b.

B. **True.**

When n is divisible by b, i.e., $n = Mb$, the dataset can be perfectly divided into M mini-batches with each mini-batch containing exactly b data points. Hence, there are indeed M mini-batches in this case.

C. **False.**

The statement claims that there are always $M+1$ mini-batches. However, when n is divisible by b, there are exactly M mini-batches, not $M + 1$. The additional mini-batch only appears when n is not divisible by b.

D. **True.**

If n is not divisible by b, the first M mini-batches will each contain b data points, accounting for Mb data points in total. The remaining $n - Mb$ data points will form the last mini-batch, ensuring that all data points are utilized in the mini-batch gradient descent algorithm.

Question 11. [1 MARK]

Let everything be defined as in the previous question. Which of the following statements are true?

A. If n is not divisible by b then the estimated loss based on the last mini-batch is

$$\frac{1}{b}\sum\nolimits_{i=(M-1)b+1}^{Mb}\ell(f(\mathbf{x}_i),y_i).$$

B. If n is not divisible by b then the estimated loss based on the last mini-batch is

$$\frac{1}{n-Mb}\sum\nolimits_{i=Mb+1}^{n}\ell(f(\mathbf{x}_i),y_i).$$

C. If n is divisible by b then the estimated loss based on the last mini-batch is

$$\frac{1}{b}\sum\nolimits_{i=(M-1)b+1}^{Mb}\ell(f(\mathbf{x}_i),y_i).$$

D. If n is not divisible by b then the variance of the estimated loss based on the last mini-batch is larger than the variance of the estimated loss based on any of the other mini-batches.

Solution 11. Correct answer(s): B, C, D

Explanation:

A. **False.**

If n is not divisible by b, the last mini-batch will contain fewer than b data points, specifically $n-Mb$ data points. Therefore, the correct scaling factor for the estimated loss based on the last mini-batch should be $\frac{1}{n-Mb}$, not $\frac{1}{b}$.

B. **True.**

When n is not divisible by b, the last mini-batch consists of $n-Mb$ data points. The estimated loss for this mini-batch is appropriately given by:

$$\frac{1}{n-Mb}\sum_{i=Mb+1}^{n}\ell(f(\mathbf{x}_i),y_i).$$

C. **True.**

When n is divisible by b, the last mini-batch contains exactly b data points. Thus, the estimated loss based on this mini-batch is calculated as:

$$\frac{1}{b}\sum_{i=(M-1)b+1}^{Mb}\ell(f(\mathbf{x}_i),y_i).$$

D. **True.**

When n is not divisible by b, the last mini-batch has fewer data points ($n-Mb$ rather than b), resulting in a higher variance for the estimated loss based on this mini-batch compared to other mini-batches, which all have b data points.

Question 12. [1 MARK]

Let everything be defined as in the previous two questions. Your friend is trying to implement the version of mini-batch gradient descent discussed in the previous two questions with a constant step size. They have written the following pseudocode and asked you to review it. Which of the following statements are true?

Algorithm 1: MBGD Linear Regression Learner (with a constant step size and last mini-batch)

input: $\mathcal{D} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$, step size η, number of epochs T, mini-batch size b
$\mathbf{w} \leftarrow$ random vector in $\mathbb{R}^{d+1}$
$M \leftarrow \text{floor}\left(\frac{n}{b}\right)$
for $t = 1, \ldots, T$ **do**
  randomly shuffle $\mathcal{D}$
  for $m = 1, \ldots, M$ **do**
    $\nabla \hat{L}(\mathbf{w}) \leftarrow \frac{2}{b} \sum_{i=(m-1)b+1}^{mb} (\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_i$
    $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \hat{L}(\mathbf{w})$
  if $n > Mb$ **then**
    $\nabla \hat{L}(\mathbf{w}) \leftarrow \frac{2}{n-Mb} \sum_{i=Mb+1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_i$
    $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \hat{L}(\mathbf{w})$
return $\hat{f}(\mathbf{x}) = \mathbf{x}^\top \hat{\mathbf{w}}$

A. The pseudocode is correct.

B. The pseudocode is incorrect because the step size should be updated at each epoch.

C. The pseudocode is incorrect because the gradient calculation for the last mini-batch is incorrect.

D. The pseudocode is incorrect because the if statement should be inside the for loop over mini-batches.

Solution 12. Correct answer(s): A

Explanation:

A. **True.**

The pseudocode correctly implements mini-batch gradient descent with a constant step size. It processes all mini-batches, including the last one even when the dataset size n is not perfectly divisible by the mini-batch size b. The gradient is appropriately scaled by the mini-batch size, and the step updates are correctly applied both within the loop for full mini-batches and outside the loop for the last mini-batch if it contains fewer than b data points.

B. **False.** A constant step size is used, and it is not necessary to update the step size at each epoch. The step size is fixed throughout the training process, and updating it at each epoch is not a requirement for the correctness of the algorithm.

C. **False.**

The gradient calculation for the last mini-batch is correctly implemented. When n is not divisible by b, the last mini-batch contains $n - Mb$ data points. The gradient is computed by averaging over these $n - Mb$ data points, as shown in the pseudocode:

$$\nabla \hat{L}(\mathbf{w}) \leftarrow \frac{2}{n - Mb} \sum_{i=Mb+1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_i$$

This ensures that all data points are utilized and the gradient is properly scaled.

D. **False.**

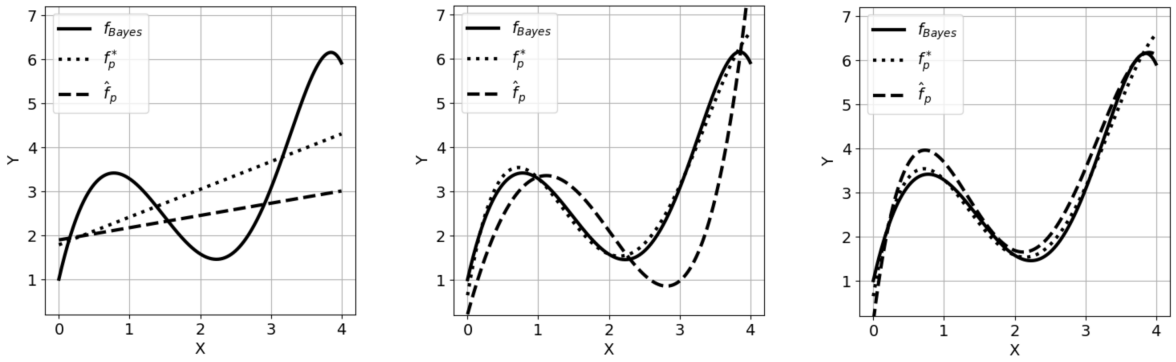
Placing the if statement outside the for loop over mini-batches is appropriate. The for loop handles all full mini-batches of size b, and the if statement subsequently handles the potential last mini-batch that may contain fewer than b data points. If the if statement were inside the for loop, it would incorrectly attempt to process the last mini-batch multiple times or interfere with the processing of full mini-batches. Therefore, the current placement of the if statement ensures correct and efficient processing of all data points.

Question 13. [1 MARK]

Let ϕ_p be the polynomial feature map of degree p, and $\mathcal{F}_p$ the function class containing all polynomials of degree p or less. Recall that

$$f_{\text{Bayes}} = \underset{f \in \{f | f:\mathbb{R}^{d+1} \to \mathbb{R}\}}{\arg\min} L(f), \quad f_p^* = \underset{f \in \mathcal{F}_p}{\arg\min} L(f), \quad \hat{f}_p = \underset{f \in \mathcal{F}_p}{\arg\min} \hat{L}(f).$$

Below are plots for different values of p and dataset size n. Which of the following statements are true?



(a) Fig 1 (b) Fig 2 (c) Fig 3

A. The polynomial degree p does not affect f_p^*.

B. The polynomial degree p does not affect f_{Bayes}.

C. The value of p used in Fig 1 is less than the value of p used in Fig 2, since f_p^* is closer to f_{Bayes} in Fig 2.

D. In Fig 1, p is likely to be equal to 1, since both f_p^* and $\hat{f}_p$ are lines, while f_{Bayes} is a curve.

Solution 13. Correct answer(s): B, C, D

Explanation:

A. **False.**

The function f_p^* is the best approximation within the function class $\mathcal{F}_p$ that minimizes the loss $L(f)$. The polynomial degree p directly influences the complexity and flexibility of the function class $\mathcal{F}_p$. A higher degree p allows for more complex models that can better fit the data. Therefore, changing p will generally affect f_p^*. Hence, statement A is false.

B. **True.**

The function f_{Bayes} is defined as the minimizer of the loss over all possible functions from $\mathbb{R}^{d+1}$ to $\mathbb{R}$. It represents the optimal predictor irrespective of any constraints on the function class. Therefore, the polynomial degree p of the function class $\mathcal{F}_p$ does not influence f_{Bayes}. Statement B is true.

C. **True.**

In Fig 2, f_p^* is closer to f_{Bayes} compared to Fig 1, indicating that a higher polynomial degree p was used in Fig 2 to achieve a better approximation of the true underlying function. Conversely, Fig 1 shows f_p^* further from f_{Bayes}, suggesting a lower degree p. Therefore, the value of p in Fig 1 is less than that in Fig 2. Statement C is true.

D. **True.**

In Fig 1, both f_p^* and $\hat{f}_p$ are represented as straight lines, which is characteristic of linear models (polynomials of degree $p = 1$). Meanwhile, f_{Bayes} is depicted as a curve, indicating a more complex underlying relationship that cannot be captured by a linear model. This suggests that in Fig 1, the polynomial degree p is likely set to 1. Therefore, statement D is true.

Question 14. [1 MARK]

Let everything be defined as in the previous question. Which of the following statements are true?

A. The value of p in Fig 2 and Fig 3 is the same since f_p^* is the same in both figures.

B. The value of n in Fig 3 is likely larger than in Fig 2 since $\hat{f}_p$ is much closer to f_p^* in Fig 3.

C. The value of n in Fig 2 is likely larger than in Fig 1 since f_p^* is much closer to f_{Bayes} in Fig 2.

D. The approximation error in Fig 1 is likely smaller than in Fig 2 as $\hat{f}_p$ is closer to f_p^* in Fig 1.

Solution 14. Correct answer(s): B

Explanation:

A. **False.**

The parameter p in Fig 2 and Fig 3 is not necessarily the same value, even if the functions f_p^* are identical in both figures.

Consider the case where $f_5^* = \arg\min_{f \in \mathcal{F}_5} L(f)$ satisfies $f_5^* \in \mathcal{F}_4$. This implies that the optimal function for $\mathcal{F}_4$ is also

$$f_4^* = \arg\min_{f \in \mathcal{F}_4} L(f) = f_5^*.$$

Therefore, if $p = 4$ in Figure 2 and $p = 5$ in Figure 3, the curves representing f_p^* will be identical in both figures, despite the different values of p.

In other words, although increasing p is likely to make f_p^* closer to f_{Bayes}, it is not guaranteed.

B. **True.**

Since the estimation error goes down as n increases, if $\hat{f}_p$ is much closer to f_p^* in Fig 3 compared to Fig 2, it suggests that the dataset size n in Fig 3 is likely larger than in Fig 2. This is because a larger dataset size allows for a more accurate estimation of the optimal predictor f_p^*, resulting in a closer approximation by $\hat{f}_p$. Therefore, Statement B is true.

C. **False.**

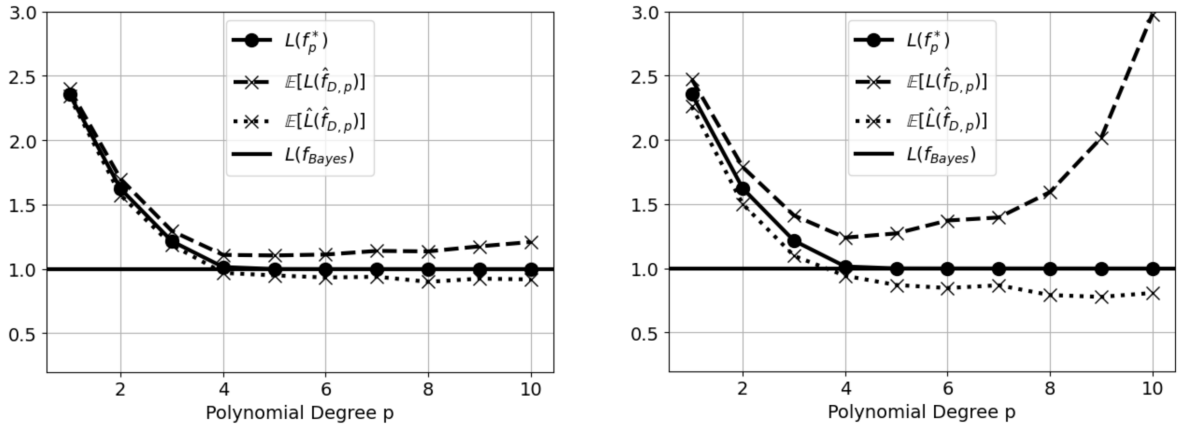
The closeness of f_p^* to f_{Bayes} is only influenced by the polynomial degree p rather than the dataset size n.

D. **False.**

The approximation error refers to how well f_p^* approximates f_{Bayes}, which is determined by the polynomial degree p. The proximity of $\hat{f}_p$ to f_p^* pertains to the estimation error, not the approximation error. Additionally, a closer $\hat{f}_p$ to f_p^* typically indicates a smaller estimation error, not necessarily a smaller approximation error. Therefore, Statement D is false.

Question 15. [1 MARK]

You have access to the true feature-label distribution $\mathbb{P}_{\mathbf{X},Y}$. You are interested in studying the estimation, approximation, and irreducible errors as a function of polynomial degree p and dataset size n. To do this, you plot the following figures. Note that $L(f_p^*), L(f_{\text{Bayes}})$ are identical in both Fig 1 and Fig 2. Which of the following statements are true?



(a) Fig 1

(b) Fig 2

A. The dataset size n used in Fig 1 is likely larger than in Fig 2 since the estimation error is smaller in Fig 1 for all values of p.

B. In Fig 1 the irreducible error is smaller for $p = 2$ than for $p = 8$.

C. In Fig 2 the estimation error is smaller for $p = 2$ than for $p = 8$.

D. In Fig 2 the approximation error is smaller than in Fig 1 for all values of p.

Solution 15. Correct answer(s): A, C

Explanation:

A. **True.** A larger dataset size n reduces estimation error, so Fig 1 likely has a larger n than Fig 2.

B. **False.** The irreducible error is determined by the true underlying distribution and is independent of the polynomial degree p.

C. **True.** Visually, the estimation error $\mathbb{E}[L(\hat{f}_{D,p})] - L(f_p^*)$ is smaller for $p = 2$ than for $p = 8$ in Fig 2.

D. **False.** The approximation error is $L(f_p^*) - L(f_{\text{Bayes}})$ and is the exact same in both figures.

Question 16. [1 MARK]

Let everything be defined as in the previous question. Which of the following statements are true?

A. In Fig 2 the predictor $\hat{f}_{D,p}$ is overfitting for $p = 1$ and underfitting for $p = 10$.

B. The best choice of polynomial degree to use for a learner, based on Fig 2, is $p = 4$ since the expected loss $\mathbb{E}[L(\hat{f}_{D,p})]$ is smallest for $p = 4$.

C. It is impossible to make irreducible error smaller by changing the dataset size n or the polynomial degree p.

D. If you gather new data that includes more features that are relevant to the prediction task, the irreducible error will likely decrease.

Solution 16. Correct answer(s): B, C, D

Explanation:

A. **False.** For $p = 1$, the predictor is too simple and underfits, while for $p = 10$, it is too complex and overfits.

B. **True.** Based on Fig 2, $p = 4$ achieves the smallest expected loss $\mathbb{E}[L(\hat{f}_{D,p})]$, making it the optimal choice.

C. **True.** Irreducible error is inherent to the data distribution and cannot be reduced by altering n or p.

D. **True.** Adding relevant features can capture more information, thereby reducing irreducible error.

Question 17. [1 MARK]

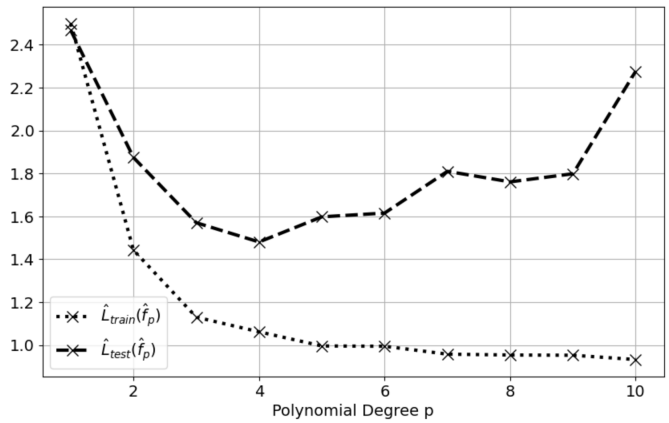
You are trying to decide which polynomial degree p to use for the function class $\mathcal{F}_p$ for a closed-form polynomial regression learner. You have a dataset of size n which you split into a training set and a test set as follows:

$$\mathcal{D}_{\text{train}} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-m}, y_{n-m})), \quad \text{and} \quad \mathcal{D}_{\text{test}} = ((\mathbf{x}_{n-m+1}, y_{n-m+1}), \ldots, (\mathbf{x}_n, y_n)).$$

You train a polynomial regression learner for each p on $\mathcal{D}_{\text{train}}$, giving you a predictor $\hat{f}_p$ for each p. The training and test loss are defined as follows:

$$\hat{L}_{\text{train}}(f) = \frac{1}{n-m} \sum_{i=1}^{n-m} \ell(f(\mathbf{x}_i), y_i), \quad \hat{L}_{\text{test}}(f) = \frac{1}{m} \sum_{i=n-m+1}^{n} \ell(f(\mathbf{x}_i), y_i).$$

You plot the training loss $\hat{L}_{\text{train}}(\hat{f}_p)$ and the test loss $\hat{L}_{\text{test}}(\hat{f}_p)$ as a function of p, which is shown below. Which of the following statements are true?



A. The best choice of p based on the plot is $p = 10$ since the train loss is smallest for $p = 10$.

B. The best choice of p based on the plot is $p = 4$ since the test loss is smallest for $p = 4$.

C. The reason that the train loss decreases as p increases is because $\mathcal{F}_p$ becomes a larger function class as p increases.

D. The test loss is usually a better estimate of $\mathbb{E}[L(\hat{f}_p)]$ than the train loss.

Solution 17. Correct answer(s): B, C, D

Explanation:

A. **False.** Choosing $p = 10$ based solely on the smallest training loss can lead to overfitting.

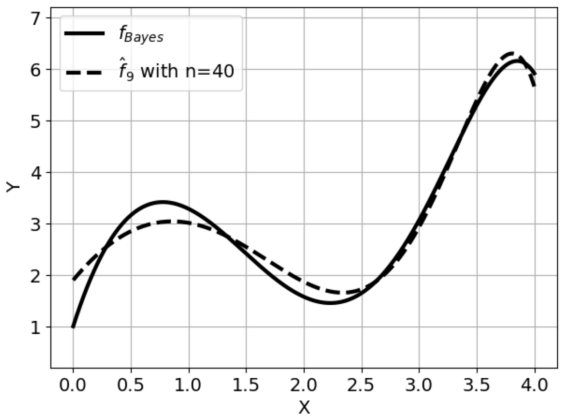
B. **True.** $p = 4$ has the smallest test loss, indicating better generalization.

C. **True.** A higher p increases the function class $\mathcal{F}_p$, allowing lower training loss.

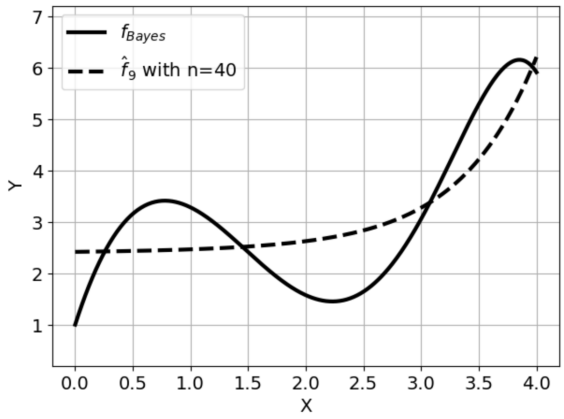
D. **True.** Test loss better reflects the expected loss $\mathbb{E}[L(\hat{f}_p)]$ as it measures performance on unseen data.

Question 18. [1 MARK]

You are working on a ridge regression problem and choose a polynomial feature map of degree $p = 9$. You have a dataset of size $n = 40$ and use a closed-form learner with regularization parameter $\lambda = 0$ and $\lambda = 100$, to obtain two different predictors. You plot both of the predictors below. Which of the following statements are true?



(a) Fig 1



(b) Fig 2

A. The predictor $\hat{f}_9$ in Fig 1 is likely the predictor output by the learner with $\lambda = 0$.

B. The predictor $\hat{f}_9$ in Fig 2 is likely the predictor output by the learner with $\lambda = 0$.

C. The predictor $\hat{f}_9$ in Fig 1 is a better predictor than the predictor $\hat{f}_9$ in Fig 2 since it is closer to f_{Bayes}, indicating it has a smaller expected loss.

D. The predictor $\hat{f}_9$ in Fig 2 is a better predictor than the predictor $\hat{f}_9$ in Fig 1 since it is simpler.

Solution 18. Correct answer(s): A, C

Explanation:

A. **True.** $\lambda = 0$ means no regularization, leading to a complex predictor, like in Fig 1.

B. **False.** Fig 2 corresponds to $\lambda = 100$, not $\lambda = 0$, since the predictor is simpler.

C. **True.** The predictor in Fig 1 is closer to f_{Bayes}, indicating better performance.

D. **False.** While Fig 2's predictor is simpler due to regularization, it is not necessarily better since it is farther from f_{Bayes}.

Question 19. [1 MARK]

Let ϕ_p be the polynomial feature map of degree p. The function class containing all polynomials of degree p or less is

$$\mathcal{F}_p = \{f \mid f : \mathbb{R}^{d+1} \to \mathbb{R}, \text{ and } f(\mathbf{x}) = \phi_p(\mathbf{x})^\top \mathbf{w}, \text{ for some } \mathbf{w} \in \mathbb{R}^{\bar{p}}\}.$$

Which of the following statements are true?

A. $\mathcal{F}_p$ is a subset of $\mathcal{F}_{p+1}$.

B. $\min_{f \in \mathcal{F}_p} \hat{L}(f) \leq \min_{f \in \mathcal{F}_{p+1}} \hat{L}(f)$.

C. If $d = 2$ then $\phi_3(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3) \in \mathbb{R}^{10}$.

D. If $f_4 \in \mathcal{F}_4$, then $f_4 \in \mathcal{F}_3$.

Solution 19. Correct answer(s): A, C

Explanation:

A. **True.**

The function class $\mathcal{F}_p$ contains all polynomials of degree p or less. Since any polynomial of degree p or less is also a polynomial of degree $p + 1$ or less, it follows that $\mathcal{F}_p \subseteq \mathcal{F}_{p+1}$. Therefore, statement A is true.

B. **False.**

Since $\mathcal{F}_p \subseteq \mathcal{F}_{p+1}$, the minimum estimated loss over $\mathcal{F}_{p+1}$ should be less than or equal to the minimum over $\mathcal{F}_p$. That is,

$$\min_{f \in \mathcal{F}_{p+1}} \hat{L}(f) \leq \min_{f \in \mathcal{F}_p} \hat{L}(f).$$

However, the statement claims the opposite inequality, so statement B is false.

C. **True.**

For $d = 2$, the polynomial feature map $\phi_3(\mathbf{x})$ of degree 3 includes all monomials up to degree 3 in variables x_1 and x_2. The total number of such monomials is:

Degree 0: 1 term (constant term)

Degree 1: 2 terms (x_1, x_2)

Degree 2: 3 terms (x_1^2, x_1x_2, x_2^2)

Degree 3: 4 terms (x_1^3, $x_1^2x_2$, $x_1x_2^2$, x_2^3)

Total terms: $1 + 2 + 3 + 4 = 10$ terms.

Therefore, $\phi_3(\mathbf{x}) \in \mathbb{R}^{10}$, and the given expression correctly lists all the terms. Thus, statement C is true.

D. **False.**

The function $f_4 \in \mathcal{F}_4$ is a polynomial of degree up to 4. If f_4 has degree exactly 4, it cannot be represented as a polynomial of degree 3 or less, and hence $f_4 \notin \mathcal{F}_3$. Therefore, statement D is false.

Question 20. [1 MARK]

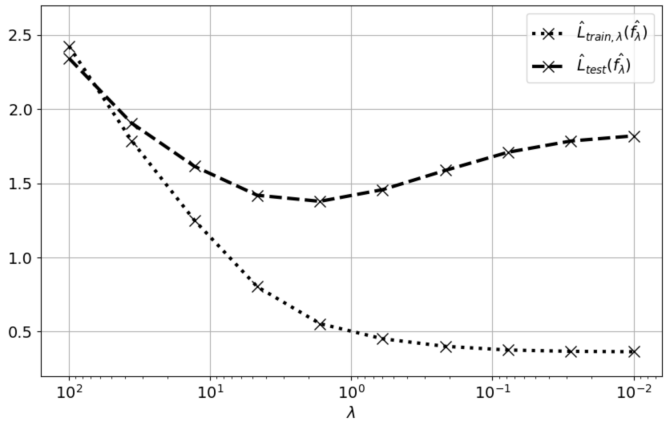
You are trying to decide which regularization parameter value λ to use for a closed-form polynomial regression learner with degree $p = 9$. You have a dataset of size n which you split into a training set and a test set as follows:

$$\mathcal{D}_{\text{train}} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-m}, y_{n-m})), \quad \mathcal{D}_{\text{test}} = ((\mathbf{x}_{n-m+1}, y_{n-m+1}), \ldots, (\mathbf{x}_n, y_n)).$$

You train a polynomial regression learner for 10 different values of λ on $\mathcal{D}_{\text{train}}$, giving you a different predictor $\hat{f}_\lambda$ for each value of λ. The training and test loss are defined as follows:

$$\hat{L}_{\text{train},\lambda}(f) = \frac{1}{n-m} \sum_{i=1}^{n-m} \ell(f(\mathbf{x}_i), y_i) + \frac{\lambda}{n-m} \sum_{j=1}^{\bar{p}-1} w_j^2, \quad \hat{L}_{\text{test}}(f) = \frac{1}{m} \sum_{i=n-m+1}^{n} \ell(f(\mathbf{x}_i), y_i).$$

You plot the training loss $\hat{L}_{\text{train},\lambda}(\hat{f}_\lambda)$ and the test loss $\hat{L}_{\text{test}}(\hat{f}_\lambda)$ as a function of λ, which is shown below. Which of the following statements are true?



A. Based on the plot, for large λ values, such as $\lambda = 100$, the predictor $\hat{f}_\lambda$ is likely underfitting.

B. Based on the plot, for small λ values, such as $\lambda = 0.01$, the predictor $\hat{f}_\lambda$ is likely overfitting.

C. The approximation error is likely higher for $\lambda = 100$ than for $\lambda = 0.01$.

D. Since the training loss is small at $\lambda \approx 2$, it is the best choice of λ.

Solution 20. Correct answer(s): A, B

Explanation:

A. **True.** A large λ imposes strong regularization, simplifying the model and potentially causing underfitting.

B. **True.** A small λ, especially $\lambda = 0$, removes regularization, allowing the model to fit the training data closely and possibly overfitting.

C. **False.** Since the function class is fixed to $\mathcal{F}_9$ the approximation error is the same for all λ. However, the bias changes.

D. **False.** $\lambda \approx 2$ is the best choice since the test loss is the smallest there.