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 11 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 **20 questions**. Each question is worth **1 mark**. The exam is worth a total of **20 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.
The regularized estimated loss is defined as the following function:

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

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

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

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

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

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.

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 $\eta^{(t)} = \frac{1}{\hat{L}''_\lambda(w^{(t)})}$.

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

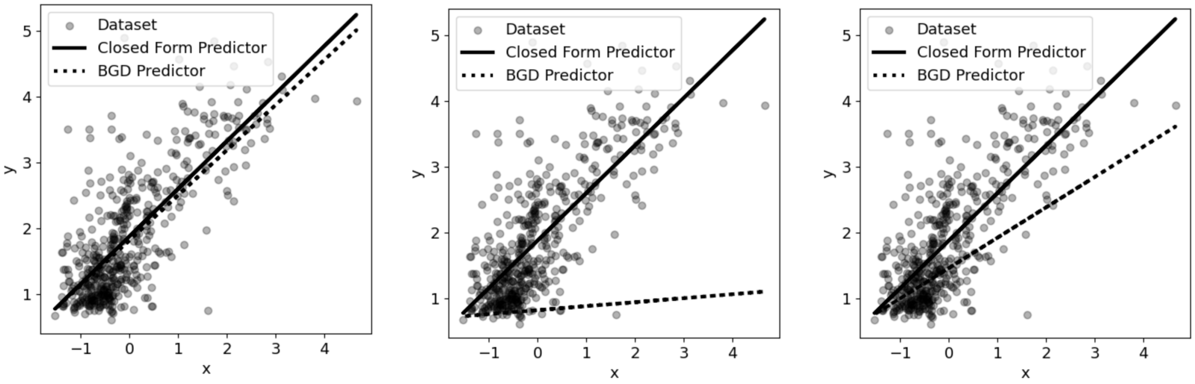
B. We have not learned in class how to check if the estimated loss $\hat{L}$ is convex.

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

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

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.

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 $6^2 \cdot 4 \cdot \theta^5 (1-\theta)^6$.

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 $\theta_{\text{MLE}} = \frac{7}{6}$.

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

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.

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

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 $n - Mb$.

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

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

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

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

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

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

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.

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 $\eta$, 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.

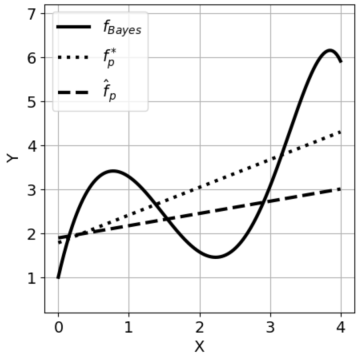
D. The pseudocode is incorrect because the if statement should be inside the for loop over mini-batches.

Question 13. [1 MARK]

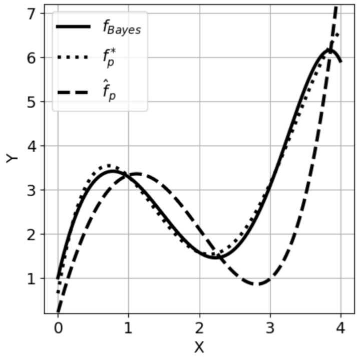
Let ϕ_p be the polynomial feature map of degree p, and $\mathcal{F}_p$ the function class containing all polynomials of degree p or less.
Recall that

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

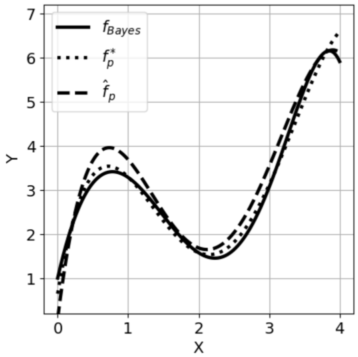
Below are plots for different values of p and dataset size n. Which of the following statements are true?



(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.

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.

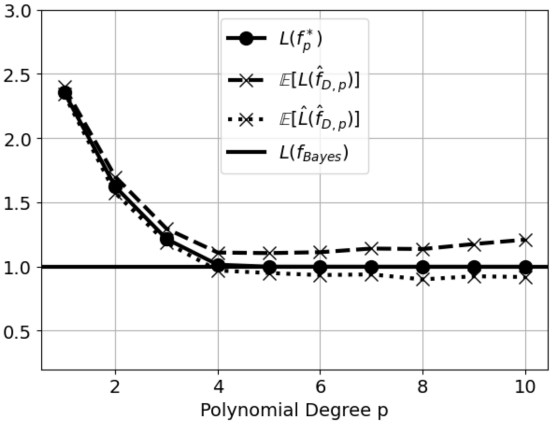
B. The value of n in Fig 3 is likely larger than in Fig 2 since $\hat{f}_p$ is much closer to f_p^* in Fig 3.

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.

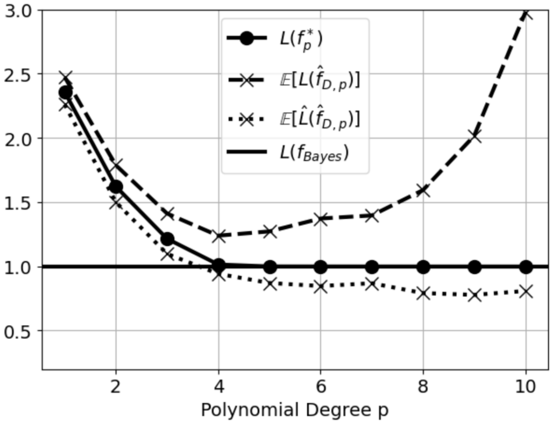
D. The approximation error in Fig 1 is likely smaller than in Fig 2 as $\hat{f}_p$ is closer to f_p^* in Fig 1.

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.

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.

Question 17. [1 MARK]

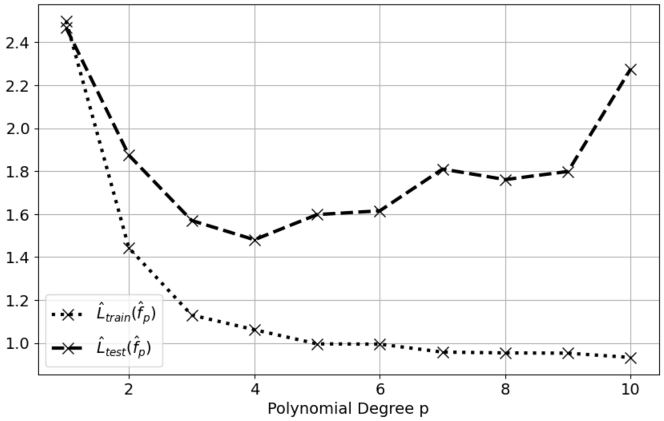
You are trying to decide which polynomial degree p to use for the function class $\mathcal{F}_p$ for a closed-form polynomial regression learner. You have a dataset of size n which you split into a training set and a test set as follows:

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

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 $p = 10$.

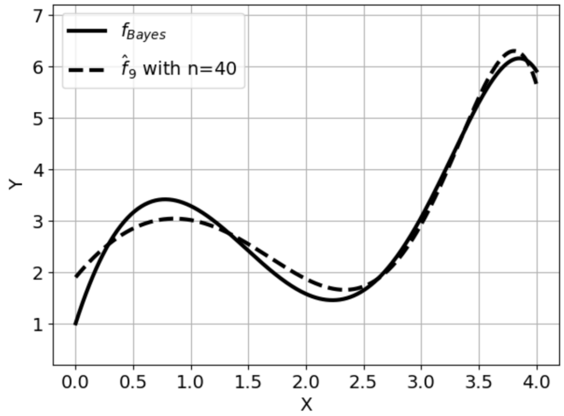
B. The best choice of p based on the plot is $p = 4$ since the test loss is smallest for $p = 4$.

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

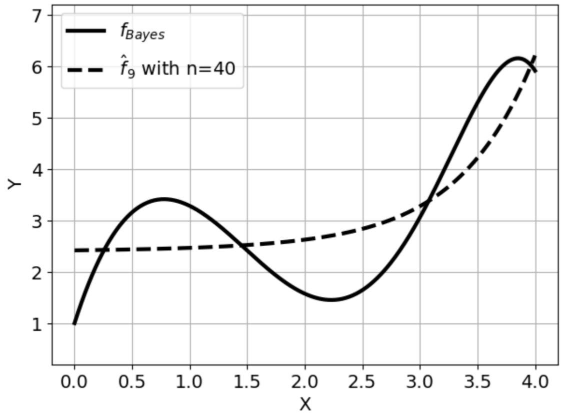
D. The test loss is usually a better estimate of $\mathbb{E}[L(\hat{f}_p)]$ than the train loss.

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.

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

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

Question 20. [1 MARK]

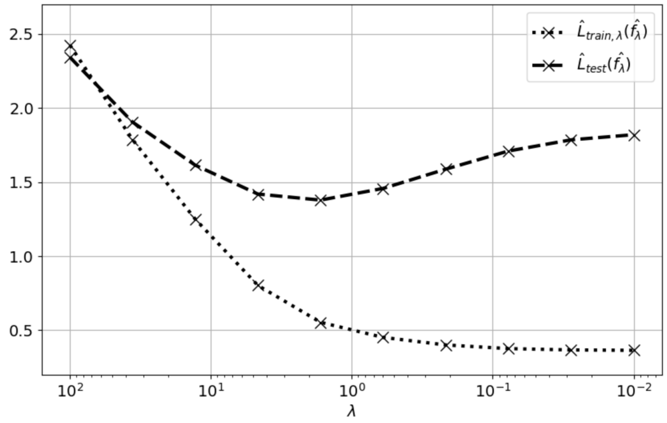
You are trying to decide which regularization parameter value λ to use for a closed-form polynomial regression learner with degree $p = 9$. You have a dataset of size n which you split into a training set and a test set as follows:

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

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.