6.790 Homework 2

Revision: 10/9/24

Solutions

Questions 1 and 2 are relatively stand-alone warm-ups. Questions 3, 4, and 5 are more extended practice and illustrations of the ideas of this material. Question 6 requires running some code and question 7 requires a small amount of implementation and ask you to answer some questions and include some plots in your submission. *Do not submit your code!*

The computational problems are based on this Google Colab notebook.

There are some rhetorical questions in blue boxes. You don't need to answer them—they're just for thinking about.

Please hand in your work via Gradescope via the link at https://gradml.mit.edu/info/homeworks/. If you were not added to the course automatically, please use Entry Code R7RGGX to add yourself to Gradescope.

1. Latex is not required, but if you are hand-writing your solutions, please write clearly and carefully. You should include enough work to show how you derived your answers, but you don't have to give careful proofs.

2. Homework is due on Tuesday October 1 at 11PM.

3. Lateness and extension policies are described at https://gradml.mit.edu/info/class_policy/.

Contents

Solution: Don't look at the solutions until you have tried your absolute hardest to solve the problems. This is especially true for optional problems that you didn't work on—it's a good idea to come back to them when studying for exams.

1 Bayesian Regression (7 points)

In this problem we will consider the standard Bayesian approach to linear regression, in which we put a Gaussian prior on the weights. Assume $x^{(i)} \in \mathbb{R}^2$, where the first feature of each $x^{(i)}$ is 1. So our data set will have the form $\mathcal{D} = \{((1, x_2^{(i)}), y^{(i)})\}_{i=1}^{n}$. And let

$$\begin{aligned} p(Y \mid X) &= \text{Normal}(W^T X, 1) \\ p(W) &= \text{Normal}(\mathbf{0}, 3\mathbf{I}) \end{aligned}$$

The figure below has some plots of the posterior on the parameters W, $\Pr(W \mid \mathcal{D})$, and of the data likelihood given parameters W, $\Pr(\mathcal{D} \mid W)$, for different values of $\mathcal{D}$. Each plot is in the space of W, indexed by w_1 and w_2, so that the mean of $\Pr(y \mid x_2) = w_1 + w_2 x_2$.

In the densities, the smallest contour contains 10% of the probability mass, and each larger contour is the next decile. In the likelihood plots, the brighter areas have higher density.

For each of the following quantities, indicate which plot corresponds to it, or **None** if none of them do.

(a) $\Pr(W)$ (Prior)

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H √ **I** ○ None

(b) $\Pr(\mathcal{D} = \{((1,1),1)\} \mid W)$ (Likelihood of one data point)

○ A ○ B ○ C ○ D ○ E ○ F ○ G √ **H** ○ I ○ None

(c) $\Pr(\mathcal{D} = \{((1,-1),-1)\} \mid W)$ (Likelihood of one data point)

○ A ○ B ○ C √ **D** ○ E ○ F ○ G ○ H ○ I ○ None

(d) $\Pr(\mathcal{D} = \{((1,0),-1)\} \mid W)$ (Likelihood of one data point)

√ **A** ○ B ○ C ○ D ○ E ○ F ○ G ○ H ○ I ○ None

(e) $\Pr(W \mid \mathcal{D} = \{((1,1),1)\})$ (Posterior after one data point)

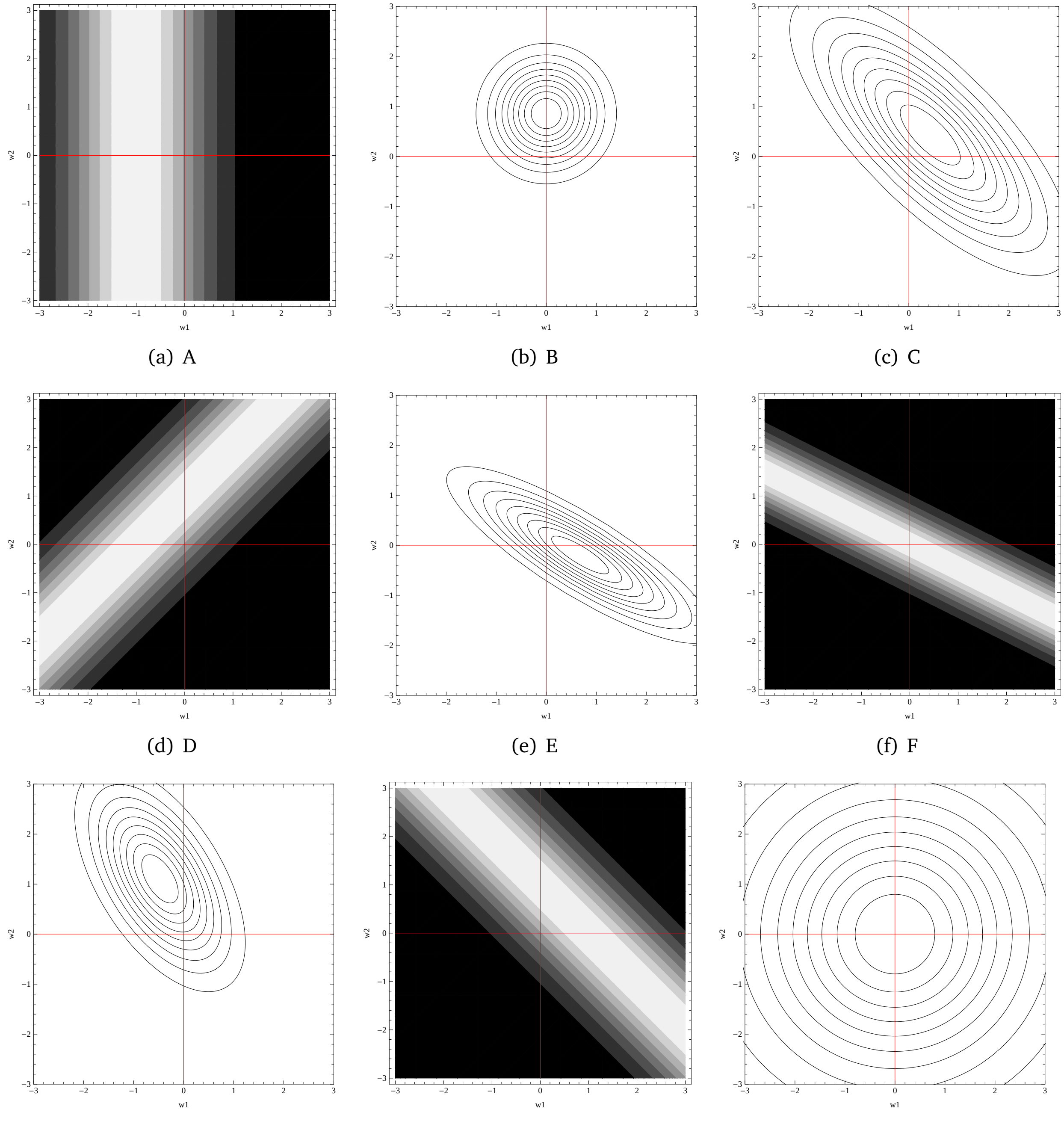
○ A ○ B √ **C** ○ D ○ E ○ F ○ G ○ H ○ I ○ None

(f) $\Pr(W \mid \mathcal{D} = \{((1,1),1),((1,-1),-1)\})$ (Posterior after two data points)

○ A √ **B** ○ C ○ D ○ E ○ F ○ G ○ H ○ I ○ None

(g) $\Pr(W \mid \mathcal{D} = \{((1,1),1),((1,0),-1)\})$ (Posterior after two data points)

○ A ○ B ○ C ○ D ○ E ○ F √ **G** ○ H ○ I ○ None



(a) A

(b) B

(c) C

(d) D

(e) E

(f) F

(g) G

(h) H

(i) I

Figure 1: **Linear Regression Plots**

2 The New Normal (8 points)

You have just discovered a time machine and want to use it to regress back in time to your first birthday. There is a big knob that seems to be freely turnable in both directions; when you turn it, there is a numeric "read-out" on the console of the time machine that varies linearly with the amount the knob is turned. Right now, the numbers on the display read 2024.75, which happens to be the current time, measured in years. You think that the amount the knob is turned correlates with the year the time machine goes to.

You begin to do some experiments. You find that when you arrive at a new time, you can estimate the year, with a standard deviation of about 2 years. Your best guess, initially, is that the the display is in direct correspondence with the date, but you assign a variance of 1 to the parameters of the linear dependence and you don't think the parameters are correlated.

(a) (2 points) You turn the knob to 2000. What is your distribution on what year you will end up in?

(b) (3 points) Once there, you realize that the year is 1015. What is your distribution on the parameters governing the relationship between the knob and the year?

(c) (3 points) You turn the knob to 2010. What is your distribution on the year you will end up in?

Solution:

(a) Let x be the year shown on the machine, and y be the estimated year. Then the relationship between x and y is:

$$p(y \mid \theta) = \mathcal{N}(y \mid \theta_0 + \theta_1 x, 4),$$

The prior on θ is

$$p(\theta) = \mathcal{N}((\theta_0, \theta_1) \mid (0, 1), I) \ .$$

The predicted distribution of the year we will arrive at when we turn the knob to 2000 is the expected output given the posterior on the weights and an input of (1, 2000.0):

$$\begin{aligned} p(y \mid 2000.0) &= \mathcal{N}(y \mid (0,1)^T(1, 2000.0), (1, 2000.0)^T I(1, 2000.0) + 4) \\ p(y \mid 2000.0) &= \mathcal{N}(y \mid 2000.0, 4000005.0) \end{aligned}$$

Whoa! That's a big variance. Why? Uncertainty about the slope has a big lever when we're predicting a point so far away.

(b) Now, we get one observation $(x^{(1)}, y^{(1)}) = ((1, 2000.0), 1015)$. Looking at the previous part, we find

$$\begin{aligned} \mu_n &= \Sigma_n \left(I(0,1) + \frac{1}{4}(1, 2000) \times 1015 \right) \\ \Sigma_n &= \left(I + \frac{1}{4} \begin{pmatrix} 1 & 2000 \\ 2000 & 2000^2 \end{pmatrix} \right)^{-1} \end{aligned}$$

Working this out we get:

$$\mu_n = (-0.00024, 0.507)$$
$$\Sigma_n = \frac{1}{4000005}\begin{pmatrix} 4000004 & -2000 \\ -2000 & 5 \end{pmatrix}$$

(c) Now, our distribution on what year we will go to when we turn the knob to 2010 is

$$\begin{aligned} y^{(n+1)} &\sim N(y|\mu_n^T x^{(n+1)}, x^{(n+1)\top}\Sigma_n x^{(n+1)} + \sigma^2) \\ &= N(y|1020.075, 8.04) \end{aligned}$$

(d) Let $x = \{1, 2010\}$, then:
Variance of prior prediction: $x^T\Sigma_0 x + \sigma^2 = 4040105$.
Variance of posterior prediction: $x^T\Sigma_n x + \sigma^2 = 8.04$.
Prior variance on mean of weight distribution: $\Sigma_0 = I$
Variance of measurements: $\sigma^2 = 4$
Posterior variance on mean of weight distribution: Σ_n (see part b).

3 One parameter, two estimators (18 points)

In this problem, we're going to explore the bias-variance trade-off in a very simple setting. We have a set of unidimensional data, $x^{(1)}, \ldots, x^{(n)}$, drawn from the positive reals. Consider a simple model for its distribution (in a later problem we will consider a slightly different model):

- **Model 1:** The data are drawn from a uniform distribution on the interval $[0, b]$. This model has a single positive real parameter b, such that $0 < b$.

We are interested in estimates of the mean of the distribution.

(a) (1 point) What's the mean of the Model 1 distribution?

Solution: The model density is $\frac{1}{b}$ (over $[0, b]$) giving a mean $\frac{b}{2}$.

(b) (1 point) Let's start by considering the situation in which the data were, in fact, drawn from an instance of the model under consideration: a uniform distribution on $[0, b]$ (for model 1),

In model 1, the ML estimator for b is $b_{\mathbf{ml}} = \max_i x^{(i)}$. The likelihood of the data is:

$$L(b_{\mathbf{ml}}) = \prod_{i=1}^{n} \begin{cases} b_{\mathbf{ml}}^{-1} & \text{if } x^{(i)} \leqslant b_{\mathbf{ml}} \\ 0 & \text{otherwise} \end{cases}$$

We can see that if $b_{\mathbf{ml}} < x^{(i)}$, for any $x^{(i)}$, then the likelihood of the whole data set must be 0. So, we should pick $b_{\mathbf{ml}}$ to be as small as possible subject to the constraint that $b_{\mathbf{ml}} \geqslant x^{(i)}$, which means $b_{\mathbf{ml}} = \max_i x^{(i)}$.

To understand the properties of this estimator we have to start by deriving their PDFs. The minimum and maximum of a data set are also known as their first and n-th *order statistics*, and sometimes written $x^{[1]}$ and $x^{[n]}$ (we're using square brackets to distinguish these from our notation for samples in a data set).

In model 1, we just need to consider the distribution of $b_{\mathbf{ml}}$. Generally speaking, the pdf of the maximum of a set of data drawn from pdf f, with cdf F, is:

$$f_{b_{\mathbf{ml}}}(x) = nF(x)^{n-1}f(x) \tag{1}$$

The idea is that, if x is the maximum, then $n-1$ of the other data values will have to be less than x, and the probability of that is $F(x)^{n-1}$, and then one value will have to equal x, the probability of which is $f(x)$. We multiply by n because there are n different ways to choose the data value that could be the maximum.

What is the maximum likelihood estimate of the mean, $\mu_{\mathbf{ml}}$, of the distribution?

Solution: Given the MLE $b_{\mathbf{ml}}$ of b, which is $x^{[n]}$ the maximum of the data set, the MLE of the mean is $\frac{b_{\mathrm{ml}}}{2} = \frac{x^{[n]}}{2}$ (from our expression of the mean in part a).

(c) (2 points) What is $f_{b_{\mathbf{ml}}}$ for this particular case where the data are drawn uniformly from 0 to b?

Solution: $f(x) = \frac{1}{b}$, $F(x) = \frac{x}{b}$, hence $f_{b_{\mathbf{ml}}}(x) = n\frac{x^{n-1}}{b^n}$ over $[0, b]$, and is zero otherwise.

(d) (2 points) Let's look at the expected value of $\mu_{\mathbf{ml}}$.

The pdf of the max of n data points was given in Equation 1 above. Given that the max value is x, the mean is $\frac{x}{2}$ from Q1. Hence:

$$E[\mu_{\mathbf{ml}}] = \int_0^b \frac{x}{2} f_{b_{\mathbf{ml}}}(x)\,dx = \int_0^b \frac{x}{2} n\frac{x^{n-1}}{b^n}\,dx = \frac{b}{2}\frac{n}{n+1}$$

Now we can answer the question: what is the squared bias of $\mu_{\mathbf{ml}}$? Is this estimator unbiased? Is it asymptotically unbiased?

Solution: Using the given equations,

$$\mathrm{bias}^2(\mu_{\mathbf{ml}}) = (\mathbb{E}[\mu_{\mathbf{ml}}] - \mu)^2 = \left(\frac{b}{2}\frac{n}{n+1} - \frac{b}{2}\right)^2 = \frac{b^2}{4(n+1)^2}$$

Because bias^2 is not zero, it is biased. However, since $\mathrm{bias}^2 \to 0$ as $n \to \infty$, it is asymptotically unbiased.

(e) (1 point) Now, let's look at the variance of $\mu_{\mathbf{ml}}$.

$$\begin{aligned} V[\mu_{\mathbf{ml}}] &= \mathbb{E}\left[\mu_{\mathbf{ml}}^2\right] - [\mathbb{E}[\mu_{\mathbf{ml}}]]^2 \\ &= \int_0^b \left(\frac{x}{2}\right)^2 f_{b_{\mathbf{ml}}}(x)\,dx - [\mathbb{E}[\mu_{\mathbf{ml}}]]^2 \\ &= \int_0^b \frac{x^2}{4} n \frac{x^{n-1}}{b^n}\,dx - \left[\frac{b}{2}\frac{n}{n+1}\right]^2 \\ &= \frac{b^2}{4}\frac{n}{(n+1)^2(n+2)} \end{aligned}$$

What is the mean squared error of $\mu_{\mathbf{ml}}$?

Solution:

$$\text{MSE}(\mu_{\mathbf{ml}}) = \text{bias}^2(\mu_{\mathbf{ml}}) + \text{var}(\mu_{\mathbf{ml}}) = \frac{b^2}{4(n+1)^2} + \frac{b^2}{4}\frac{n}{(n+1)^2(n+2)} = \frac{b^2}{2(n+1)(n+2)}$$

(f) (1 point) So far, we have been considering the error of the *estimator*, comparing the estimated value of the mean with its actual value. We will often want to use the estimator to make predictions, and so we might be interested in the expected error of a prediction.

Assume the loss function for your predictions is $L(g, a) = (g - a)^2$. Given an estimate $\hat{\mu}$ of the mean of the distribution, what value should you predict?

Solution: This is an interesting question! If we knew the actual mean, then because the loss function is symmetric, we should just predict the mean.

But we don't know the actual mean (sad face). We just have an estimate that we have shown to be biased. We could use the ML estimate of the mean as the prediction. Would we expect it to be too high, in general? Or too low? Should we try to adjust it?

In the next section, we'll discover a better strategy.

(g) (3 points) What is the expected loss (risk) of this prediction? Take into account both the error due to inaccuracies in estimating the mean as well as the error due to noise in the generation of the actual value. Just write out the expression with integrals in it, where the only "free" variables (not being integrated over) are n and μ.

Solution: Assume we predict the ML estimate of the mean, the loss is given by:

$$
\begin{aligned}
&E_{\mu_{\mathbf{ml}}}\left[\mathbb{E}_a[(g-a)^2]\right] && (2)\\
=&E_{\mu_{\mathbf{ml}}}\left[\int_0^{2\mu}(g-a)^2 P(a|\mu)\,da\right] && (3)\\
=&E_{\mu_{\mathbf{ml}}}\left[\int_0^{2\mu}(\mu_{\mathbf{ml}}-a)^2\frac{1}{2\mu}da\right] && (4)\\
=&E_{\mu_{\mathbf{ml}}}\left[\frac{(2\mu-\mu_{\mathbf{ml}})^3+{\mu_{\mathbf{ml}}}^3}{6\mu}\right] && (5)\\
=&\int_0^{\mu}P(\mu_{\mathbf{ml}}|\mu)\frac{(2\mu-\mu_{\mathbf{ml}})^3+{\mu_{\mathbf{ml}}}^3}{6\mu}d\mu_{\mathbf{ml}} && (6)\\
=&\frac{n}{6\mu^n}\int_0^{\mu}(8\mu^2{\mu_{\mathbf{ml}}}^{n-1}-12\mu{\mu_{\mathbf{ml}}}^n+6{\mu_{\mathbf{ml}}}^{n+1})d\mu_{\mathbf{ml}} && (7)\\
=&\frac{n}{6\mu^n}\left(\frac{8\mu^{n+2}}{n}-\frac{12\mu^{n+2}}{n+1}+\frac{6\mu^{n+2}}{n+2}\right)=\frac{n^2+3n+8}{3(n+1)(n+2)}\mu^2 && (8)
\end{aligned}
$$

One thing to notice is that, when $n\to\infty$, the risk converges $\frac{1}{3}\mu^2$, which is the variance of uniform distribution.

Another estimator We might consider something other than the MLE for Model 1 (labeled o for other). Consider the estimator

$$\mu_o=\frac{x^{[n]}(n+1)}{2n}\quad .$$

where $x^{[n]}$ is the maximum of the data set.

(h) (3 points) Write an expression for the expected value of this version of μ_o as an integral where the only free variables are b and n. It integrates to $b/2$.

Solution:

$$\mathbb{E}[\mu_o]=\int_0^b\frac{x(n+1)}{2n}f_{b_o}\,dx=\int_0^b\frac{x(n+1)}{2n}n\frac{x^{n-1}}{b^n}\,dx=\frac{b}{2}$$

Where f_{b_o} is the analogous of $f_{b_{ml}}$ and also assumes the data is drawn uniformly from 0 to b.

(i) (1 point) What is the squared bias of this estimator for μ_o? Is this estimator unbiased? Is it asymptotically unbiased?

Solution:

$$\text{bias}^2(\mu_o)=(\mathbb{E}[\mu_o]-\mu)^2=\left(\frac{b}{2}-\frac{b}{2}\right)^2=0$$

The estimator is unbiased (and asymptotically unbiased).

The variance of μ_o is

$$\begin{aligned} V[\mu_o] &= \int_0^b \left(\frac{x(n+1)}{2n}\right)^2 f_{b_o}\, dx - [\mathbb{E}[\mu_o]]^2 \\ &= \int_0^b \frac{x^2(n+1)^2}{4n^2} n \frac{x^{n-1}}{b^n}\, dx - \frac{b^2}{4} \\ &= \frac{b^2}{4n(n+2)} \end{aligned}$$

(j) (1 point) What is the mean squared error of this version of μ_o?

Solution: Since the bias is zero, the MSE is the same as the variance $V[\mu_o]$.

(k) (2 points) What are the relative advantages and disadvantages of the estimators $\mu_{\mathbf{ml}}$ and μ_o?

Solution: The unbiased estimator is strictly better. It always has smaller MSE, even if the variance is higher.

4 One problem, two models (22 points)

In this problem, we're going to continue exploring the bias-variance trade-off in a very simple setting. We have a set of unidimensional data, $x^{(1)}, \ldots, x^{(n)}$, drawn from the positive reals. We will consider two different models for its distribution:

- **Model 1:** The data are drawn from a uniform distribution on the interval $[0, b]$. This model has a single positive real parameter b, such that $0 < b$.
- **Model 2:** The data are drawn from a uniform distribution on the interval $[a, b]$. This model has two positive real parameters, a and b, such that $0 < a < b$.

We are interested in comparing estimates of the mean of the distribution, derived from each of these two models.

4.1 Using Model 2

(a) (1 point) What's the mean of the Model 2 distribution?

Solution: The model density is $\frac{1}{b-a}$ (over $[a, b]$) giving a mean $\frac{a+b}{2}$.

(b) (3 points) Let's consider the situation in which the data were, in fact, drawn from an instance of the model under consideration: either a uniform distribution on $[0, b]$ (for model 1) or a uniform distribution on $[a, b]$ (for model 2).

We saw that, in model 1, the ML estimator for b is $b_{\mathbf{ml}} = \max_i x^{(i)}$.

By a similar argument in model 2, the ML estimator for b remains the same and the ML estimator for a is $a_{\mathbf{ml}} = \min_i x^{(i)}$.

We started our analysis of Model 1 in question 3. Now, let's do the same thing, but for the MLE for model 2. We have to start by thinking about the joint distribution of MLE's $a_{\mathbf{ml}}$ and $b_{\mathbf{ml}}$. Generally speaking, the joint pdf of the minimum and the maximum of a set of data drawn from pdf f, with cdf F, is

$$f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y) = n(n-1)(F(y)-F(x))^{n-2}f(x)f(y) \quad .$$

Explain in words why this makes sense.

Solution: The argument for $f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y)$ is similar to the one for $f_{b_{\mathbf{ml}}}(x)$. However, we have to choose a minimum value x in addition to the maximum value y and ensure all other values fall between x and y. First, we factor in the probability (density) of x and y, giving the final $f(x)f(y)$ terms. The other $(n-2)$ data points must all be between x and y, which is true with probability $(F(y)-F(x))^{(n-2)}$. Finally, there are $n(n-1)$ different ways of choosing the maximum and minimum points. (Note that the ordering of these two points matters, so the multiplicative factor is *not* $\binom{n}{2} = \frac{n(n-1)}{2}$).

(c) (2 points) What is $f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}$ in the particular case where the data are drawn uniformly from a to b?

Solution: $f(x) = \frac{1}{b-a}$, $F(x) = \frac{x-a}{b-a}$, hence $f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y) = n(n-1)\frac{(y-x)^{n-2}}{(b-a)^n}$ for $a \leqslant x \leqslant y \leqslant b$, and is zero otherwise.

(d) (2 point) Let's look at expected value of $\mu_{\mathbf{ml}}$:

Given that x and y are the min and max values for Model 2, the MLE is now $\frac{x+y}{2}$. Hence:

$$E[\mu_{\mathbf{ml}}] = \int\int \frac{x+y}{2} f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y)\,dxdy = \int_a^b \int_a^y \frac{x+y}{2} n(n-1)\frac{(y-x)^{n-2}}{(b-a)^n}\,dxdy = \frac{a+b}{2}$$

What is the squared bias of $\mu_{\mathbf{ml}}$? Is this estimator unbiased? Is it asymptotically unbiased?

Solution: $\text{bias}^2(\mu_{\mathbf{ml}}) = (E[\mu_{\mathbf{ml}}] - \mu)^2 = \left(\frac{a+b}{2} - \frac{a+b}{2}\right)^2 = 0$. The estimator is unbiased (and asymptotically unbiased).

(e) (2 point) And now the variance of $\mu_{\mathbf{ml}}$:

$$\begin{aligned} V[\mu_{\mathbf{ml}}] &= \int\int \left(\frac{x+y}{2}\right)^2 f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y)\,dxdy - [E[\mu_{\mathbf{ml}}]]^2 \\ &= \int_a^b \int_a^y \frac{(x+y)^2}{4} n(n-1)\frac{(y-x)^{n-2}}{(b-a)^n}\,dxdy - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{2(n+1)(n+2)} \end{aligned}$$

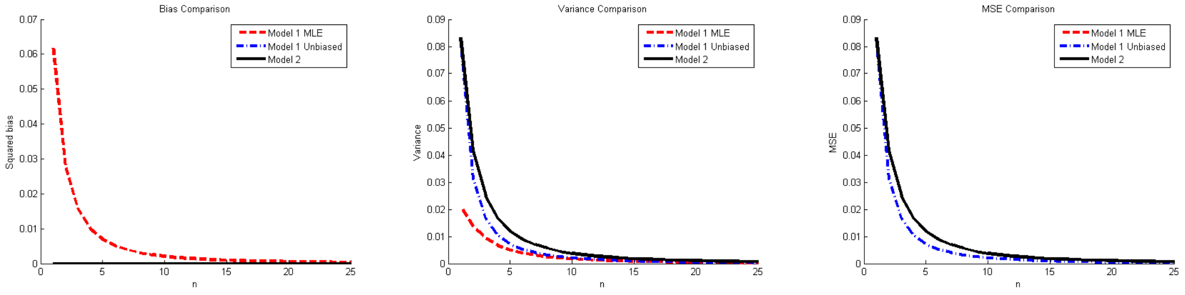
What is the mean squared error of $\mu_{\mathbf{ml}}$?

Solution: Since the bias is zero, the MSE is the same as the variance $V[\mu_{\mathbf{ml}}]$.

4.2 Comparing Models

What if we have data that is actually drawn from the interval $[0, 1]$? Both models seem like reasonable choices.

(a) (3 points) Figure 2 has plots that compare the bias, variance, and MSE of each of the estimators we've considered on that data, as a function of n. Write a paragraph in English explaining your results. What estimator would you use?



(a) Squared bias (black overlaps blue) (b) Variance (c) MSE (black overlaps red)

Figure 2: Red = model 1 MLE, blue = model 1 unbiased, black = model 2.

Solution: Note that both Model 1 unbiased (blue) and Model 2 (black) have zero bias in Figure 2(a). Also, for $a = 0$, the MSE for Model 1 MLE (red) and Model 2 are the same, so they overlap in Figure 2(c) (red under black). We already know that the Model 1 unbiased estimator (blue) has lower error than the Model 1 MLE (red). Since the MSE for Model 2 and Model 1 MLE are the same, we conclude that the Model 1 unbiased estimator is superior for data from $[0, 1]$ due to its lower variance.

(b) (2 points) Now, what if we have data that is actually drawn from the interval $[.1, 1]$? It seems like model 2 is the only reasonable choice. But is it?

We already know the bias, variance, and MSE for model 2 in this case. But what about the MLE and unbiased estimators for model 1? Let's characterize the general behavior when we use the estimator $\mu_o = x^{[n]}(n + 1)/(2n)$ on data drawn from an interval $[a, b]$.

Here is the expected value of μ_o:

$$E[\mu_o] = \int\int \frac{y(n+1)}{2n} f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x, y)\, dx dy = \int_a^b \int_a^y \frac{y(n+1)}{2n} n(n-1) \frac{(y-x)^{n-2}}{(b-a)^n}\, dx dy = \frac{a + bn}{2n}$$

Explain in English why this answer makes sense.

Solution: For small n (and in particular for $n = 1$), since the maximum value in fact cannot be less than a, a high value of a means that initial maximum values will be higher, and hence the estimated mean is higher. Ultimately, the estimator only depends on the maximum of the data b, and as we saw earlier the expected value is $\frac{b}{2}$. The expression above tends to this as $n \to \infty$, since with many data points, it is likely that their maximum is close to b.

(c) (3 points) What is the squared bias of this μ_o? Explain in English why your answer makes sense. Consider how it behaves as a increases, and how it behaves as n increases.

Solution: $\text{bias}^2(\mu_o) = (E[\mu_o] - \mu)^2 = \left(\frac{a+bn}{2n} - \frac{a+b}{2}\right)^2 = \frac{a^2(n-1)^2}{4n^2}$. We already know it's unbiased if $a = 0$; as a increases, this is an increasingly bad (inaccurate) model. Furthermore, for fixed a, the bias increases as a function of n, because the expected answer gets closer to $\frac{b}{2}$ (and farther from the true $\frac{a+b}{2}$).

(d) (2 points) The variance of this μ_o is:

$$
\begin{aligned}
V[\mu_o] &= \int\int \left(\frac{y(n+1)}{2n}\right)^2 f_{a_{\mathbf{ml}}, b_{\mathbf{ml}}}(x, y)\, dx dy - [E[\mu_o]]^2 \\
&= \int_a^b \int_a^y \frac{y^2(n+1)^2}{4n^2} n(n-1) \frac{(y-x)^{n-2}}{(b-a)^n}\, dx dy - \frac{(a+bn)^2}{4n^2} = \frac{(b-a)^2}{4n(n+2)}
\end{aligned}
$$

To save you some tedious algebra, we'll tell you that the mean squared error of this $\mu_{\mathbf{ml}}$ is (apologies for the ugliness; let us know if you find a beautiful rewrite)

$$
\frac{b^2 n - 2abn + a^2(2 - 2n + n^3)}{4n^2(n+2)} .
$$

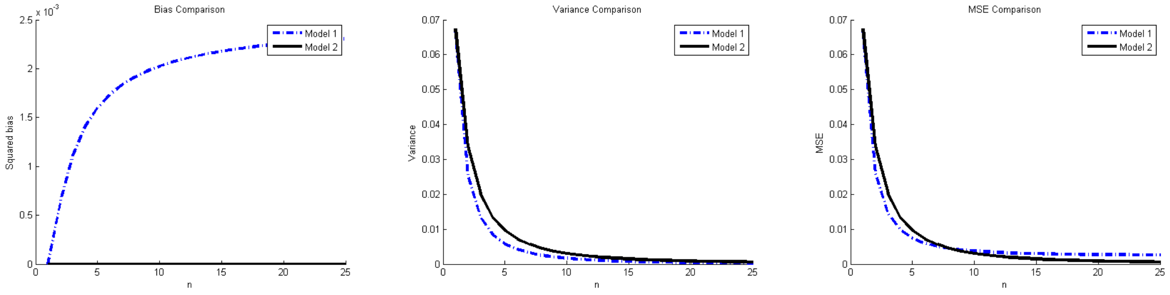
Figure 3 has plots that compare the bias, variance, and MSE of this estimator with the regular model 2 estimator on data drawn from $[0.1, 1]$, as a function of n.

Are there circumstances in which it would be better to use this estimator? If so, what are they and why? If not, why not?

Solution: The MSE plots cross over at around 8. That is, for $n < 8$, using the model 1 estimator is better, and beyond that the model 2 estimator should be used. Although model 1 has lower variance than model 2, the bias in using model 1 takes over for larger n.

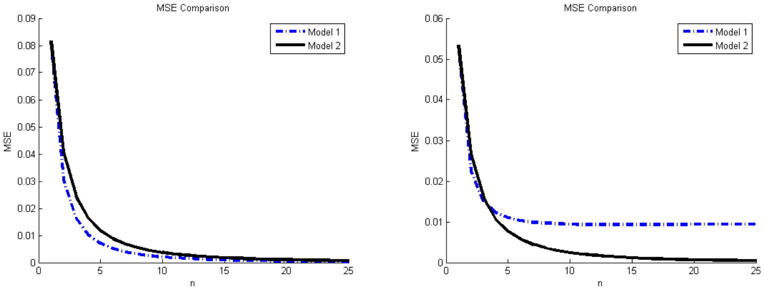
(e) (2 points) Figure 4 has plots of MSE of both estimators, as a function of n on data drawn from $[.01, 1]$ and on data drawn from $[.2, 1]$.

How do things change? Explain why this makes sense.



(a) Squared bias (b) Variance (c) MSE

Figure 3: MSE Plots: Blue = model 1, black = model 2. Data from [.1,1].



(a) [.01,1] (b) [.2,1]

Figure 4: Blue = model 1, black = model 2.

Solution: For data from $[.01, 1]$, model 1 is a very good approximation. Although the model 1 estimator is still biased, because a is very small, the effect of the bias is much smaller, and model 1 is superior for a larger range of n due to its lower variance. In contrast, for data from $[.2, 1]$, model 1 is less accurate compared to its application in Figure 3. The model 1 estimator is more biased and is inferior for $n > 3$.

5 Ridge Regression (25 points)

The goal of this question is to understand the various interpretations and properties of regularized regression. Suppose we have access to n data points: $(x^{(i)}, y^{(i)})$, $i = 1, \dots, n$. First, recall that the ridge regression algorithm finds an appropriate model by solving the following optimization problem:

$$\min_{w} \sum_{i=1}^{n} \left(y^{(i)} - (w^\top x^{(i)} + w_0)\right)^2 + \lambda \|w\|^2.$$

Note that if the hyperparameter λ is set to 0 (i.e., no regularization), the problem is identical to the Ordinary Least Squares (OLS) problem, which has closed form solution without intercept $\hat{w}^{\mathrm{OLS}} = (X^\top X)^{-1} X^\top y$, where $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^{n \times 1}$.

Assume for this problem that all data is centered (i.e., both X and y have mean 0) and thus we don't need a bias term in the ridge regression. In this case, the ridge regression objective, as we have seen in class, has a similar solution which is modulated by regularization parameter λ: $\hat{w}^{Ridge}(\lambda) = (X^\top X + \lambda I)^{-1} X^\top y$. Note that when the data is not centered, $\hat{w}^{Ridge}$ cannot be written as the above simple form because w_0 is not regularized. You are encouraged to work out the general case when the data is not centered, but it is not required for this problem.

1. (Optional, but highly educational!) Suppose that data is truly generated by a linear model: $y^{(i)} = w_*^\top x^{(i)} + z^{(i)}$, where $z^{(i)}$ are zero-mean and variance σ^2 iid noise variables and w_* is the true value of the weight vector. Let $f(\lambda) = \mathbb{E}[\|\hat{w}^{Ridge}(\lambda) - w_*\|^2]$ be the average error of the ridge estimator. (Here, the expectation is only over y_i, with x_i considered fixed.) Compute the sign of the derivative $f'(0) = \lim_{\lambda \to 0+} \frac{f(\lambda) - f(0)}{\lambda}$. What conclusion can you draw regarding using $\lambda = 0$ (the ordinary unregularized least squares)?

   You may assume that $\text{rank}(X) = d$, where X is the $n \times d$ design matrix whose rows are $x^{(i)}$.

   Hint: Express $f(\lambda)$ as the sum of two parts, one corresponding to the bias of the ridge regression and the other corresponding to the variance, and find their derivatives separately. What happens to each as λ is increased?

Solution: $f'(0) < 0$, and therefore a sufficiently small regularization constant λ will always be able to produce a smaller average error.

Remark: The solution still works if X is not full-column rank, but then $f'(0)$ has to be replaced by $\lim_{\lambda \to 0} f'(\lambda)$.

Key points: The following key points should be in your solution:

1. Average error = $\text{Bias}^2(\lambda) + \text{Variance}(\lambda)$
2. $\frac{d}{d\lambda}\text{Bias}^2(\lambda) = 0$ at $\lambda = 0$
3. $\frac{d}{d\lambda}\text{Variance}(\lambda) = -2\sigma^2 \sum_j \frac{1}{d_j^4} < 0$ where d_j are the singular values of X.

Equivalent in (3) is $-2\sigma^2 \sum_j \frac{1}{\nu_j^2}$ where ν_j are the eigenvalues of $X^\top X$, or $-2\sigma^2 \text{Tr}(X^\top X)$ (trace). The solution below is intentionally overly-thorough.

For ease of reading, let the ridge regression operator be denoted $\widehat{w}(\lambda)$, and let 0_d denote a d-dimensional (column) vector of all zeros.

First, $f(\lambda)$ can be broken down into bias-squared and variance terms as follows: if $\epsilon \in \mathbb{R}^n$ is the random noise, then $y = Xw_* + \epsilon$ and thus

$$
\begin{aligned}
f(\lambda) &= \mathbb{E}\left[\|\widehat{w}(\lambda) - w_*\|^2\right] = \mathbb{E}\left[\|(X^\top X + \lambda I)^{-1} X^\top (Xw_* + \epsilon) - w_*\|^2\right] && (9) \\
&= \|(X^\top X + \lambda I)^{-1} X^\top X w_* - w_*\|^2 + \mathbb{E}\left[\|(X^\top X + \lambda I)^{-1} X^\top \epsilon\|^2\right] && (10)
\end{aligned}
$$

Equation (10) is justified because $\mathbb{E}[\epsilon] = 0$ so cross-terms can be removed, and the left term is not random so we remove the expectation. We can then take the derivative w.r.t. λ at $\lambda = 0$ on each term separately. We use the Singular Value Decomposition $X = UDV^\top$.

For the bias-squared term,

$$(X^\top X + \lambda I)^{-1} X^\top X w_* - w_* = ((X^\top X + \lambda I)^{-1} X^\top X - I) w^* \tag{11}$$

Looking at $(X^\top X + \lambda I)^{-1} X^\top X - I$, with the SVD we write

$$(X^\top X + \lambda I)^{-1} X^\top X - I = V((D^\top D + \lambda I)^{-1} D^\top D - I) V^\top \tag{12}$$

Thus, taking the squared norm (and canceling $V^\top V = I$), we get

$$\|(X^\top X + \lambda I)^{-1} X^\top X w_* - w_*\|^2 = (w^*)^\top V((D^\top D + \lambda I)^{-1} D^\top D - I)^2 V^\top w^* \tag{13}$$

The terms in the diagonal matrix $((D^\top D + \lambda I)^{-1} D^\top D - I)^2$ are

$$\left(\frac{d_j^2}{d_j^2 + \lambda} - 1\right)^2 = \frac{\lambda^2}{(d_j^2 + \lambda)^2} \tag{14}$$

and (no matter what vector $V^\top w^*$ is), the derivative of equation (13) w.r.t. λ at $\lambda = 0$ is 0 because all terms have the quadratic λ^2 in the numerator.

For the variance term, because the noise ϵ is IID with variance σ^2, we know that $\mathbb{E}[\|(X^\top X + \lambda I)^{-1} X^\top \epsilon\|^2]$ is σ^2 times the sum of the squares of the singular values of $(X^\top X + \lambda I)^{-1} X^\top$. Using $X = UDV^\top$, we get

$$\begin{aligned} (X^\top X + \lambda I)^{-1} X^\top &= (V(D^\top D + \lambda I)^{-1} V^\top) V D^\top U^\top && (15) \\ &= V(D^\top D + \lambda I)^{-1} D^\top U^\top && (16) \end{aligned}$$

Therefore, its singular values are $\frac{d_j}{d_j^2 + \lambda}$ and so the sum of the squares of the singular values, times σ^2, is

$$\mathbb{E}[\|(X^\top X + \lambda I)^{-1} X^\top \epsilon\|^2] = \sigma^2 \sum_{j=1}^{d} \frac{d_j^2}{(d_j^2 + \lambda)^2} \tag{17}$$

Taking the derivative at $\lambda = 0$ yields

$$-2\sigma^2 \sum_{j=1}^{d} \frac{d_j^2}{(d_j^2 + \lambda)^3} = -2\sigma^2 \sum_{j=1}^{d} \frac{1}{d_j^4} < 0 \tag{18}$$

(since $\lambda = 0$)

Thus, the bias-squared term has derivative 0 at $\lambda = 0$, and the variance term has derivative < 0 at $\lambda = 0$, and so $f'(0) < 0$.

2. (8 points) Show that the closed form solution of ridge regression can be obtained by solving the ordinary least squares problem using the following augmented data set.

   To form our augmented dataset, we define the augmented data matrix C to be X with d additional rows containing $\sqrt{\lambda}I_d$ (where I_d is the $d \times d$ identity matrix), and form our augmented target z to be y with d additional zeros.

   Under this interpretation, by introducing artificial data with response value zero, the fitting procedure is forced to shrink the coefficients toward zero. This is related to the idea of using a regularization parameter to penalize the magnitude of the weight vector to prevent overfitting.

Solution:

Key points: Note that the problem requires the derivation of the *closed-form* ridge regression solution from OLS.

- Any algebra equivalent to what is below is correct; it's just clearer when written in matrix block notation.
- Partial credit for showing that the augmented OLS and the ridge regression problems yield the same w (e.g. by writing the augmented OLS and ridge regression as minimization problems and showing the problems are equivalent); this shows the same w minimizes both, but needs a little extra algebra to derive the closed-form expression of the ridge regression minimizer (as required in the problem).

We claim that the closed form of the ridge regression solution can be obtained by applying OLS to an augmented data set. Let C be X with d extra rows containing $\sqrt{\lambda}I_d$ and z be y with d additional zeros, i.e.

$$C = \begin{bmatrix} X \\ \sqrt{\lambda}I_d \end{bmatrix} \quad \text{and} \quad z = \begin{bmatrix} y \\ 0_d \end{bmatrix} \tag{19}$$

Therefore the claim is that

$$(C^\top C)^{-1}C^\top z = (X^\top X + \lambda I_d)^{-1}X^\top y \tag{20}$$

We break this into two parts:

$$C^\top C = \begin{bmatrix} X^\top & \sqrt{\lambda}I_d \end{bmatrix} \begin{bmatrix} X \\ \sqrt{\lambda}I_d \end{bmatrix} = X^\top X + \lambda I_d \tag{21}$$

and

$$C^\top z = \begin{bmatrix} X^\top & \sqrt{\lambda}I_d \end{bmatrix} \begin{bmatrix} y \\ 0_d \end{bmatrix} = X^\top y \tag{22}$$

Thus, we have our claim (20) by substitution.

3. (9 points) In the Bayesian regression setup one introduces a prior distribution on the weight parameter vector $w \sim \mathbb{P}[w]$ and then computes the posterior distribution given the data as $\mathbb{P}[w|D] \propto \mathbb{P}[w]\mathbb{P}[D|w]$ where $D = \{x^{(i)}, y^{(i)}\}_{i=1\ldots n}$. Let us set a Gaussian prior on $w \sim \mathcal{N}(0_d, \tau^2 I_d)$ and use the standard Gaussian generative assumption $y \sim \mathcal{N}(Xw, \sigma^2 I_n)$.

   Show that the closed form solution of ridge regression is the mean (and mode) of the above posterior distribution. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ^2 and σ^2 in the Gaussian formulation. Again, assume that the data is centered, and thus we don't need a bias term.

   Under this interpretation, the regularized least squares objective can be viewed as a Maximum A Posteriori (MAP) estimation under an assumption of normally distributed residuals. In this framework, the regularization terms of OLS can be understood as encoding priors on w.

Solution: $\lambda = \sigma^2/\tau^2$.

Key points: Writing out $\mathbb{P}[w|D] \propto \mathbb{P}[w]\mathbb{P}[D|w]$ and noting that it is proportional to a Gaussian pdf, then solving for the mean (which is also the mode). Solving for the mean can be done directly or (with some algebra) by the standard technique of taking the gradient at w and setting it to 0.

Suppose that the value w is generated from a Gaussian prior $\mathcal{N}(0_d, \tau^2 I_d)$, and that $y \sim \mathcal{N}(Xw, \sigma^2 I_n)$ as usual; then given data $D = (X, y)$, we have a posterior distribution for w given by

$$\mathbb{P}[w|D] \propto \mathbb{P}[w]\,\mathbb{P}[D|w] \tag{23}$$

We treat X as fixed and centered. We now claim that the mean and mode of the posterior distribution is given by ridge regression with an appropriate λ.

We can simply compute the density given w, X, y:

$$\mathbb{P}[w]\,\mathbb{P}[D|w] \propto \frac{\exp(-\frac{1}{2\tau^2}w^\top w)}{(2\pi\tau^2)^{d/2}} \frac{\exp(-\frac{1}{2\sigma^2}(y - Xw)^\top(y - Xw))}{(2\pi\sigma^2)^{n/2}} \tag{24}$$

We then have an exponential of

$$\begin{aligned}
&-\frac{1}{2\tau^2}w^\top w - \frac{1}{2\sigma^2}(y - Xw)^\top(y - Xw) && (25)\\
&= -\frac{1}{2}\left(w^\top \frac{1}{\tau^2} I w + \left(\frac{1}{\sigma^2}y^\top y - 2y^\top \frac{1}{\sigma^2}Xw + w^\top \frac{1}{\sigma^2}X^\top X w\right)\right) && (26)\\
&= -\frac{1}{2}\left(\frac{1}{\sigma^2}\left(y^\top y - 2y^\top Xw + w^\top\left(X^\top X + \frac{\sigma^2}{\tau^2} I\right)w\right)\right) && (27)
\end{aligned}$$

This is a quadratic form on w as the (negative of the) exponent, hence it is also Gaussian, and we are after its mean (which is also the mode). We therefore set $\lambda = \sigma^2/\tau^2$ and show that $\widehat{w}(\lambda) = (X^\top X + \lambda I)X^\top y$ is the mean by matching the following with the Gaussian exponent:

$$(w - \widehat{w}(\lambda))^\top (X^\top X + \lambda I)(w - \widehat{w}(\lambda)) \tag{28}$$

$$= w^\top (X^\top X + \lambda I) w - 2w^\top (X^\top X + \lambda I)\widehat{w}(\lambda) + \widehat{w}(\lambda)^\top (X^\top X + \lambda I)\widehat{w}(\lambda) \tag{29}$$

Note that the first term is already what we want. The second term is

$$-2w^\top (X^\top X + \lambda I)\widehat{w}(\lambda) = -2w^\top (X^\top X + \lambda I)(X^\top X + \lambda I)^{-1} X^\top y = -2y^\top X w \tag{30}$$

The last term (which contains $\widehat{w}(\lambda)$ but no w) is a constant term, which becomes multiplicative (i.e. if it doesn't exactly match $y^\top y$ this is due to the multiplicative normalization terms). Hence, our Gaussian exponent is really

$$-\frac{1}{2\sigma^2}(w - \widehat{w}(\lambda))^\top (X^\top X + \lambda I)(w - \widehat{w}(\lambda)) \tag{31}$$

yielding a covariance matrix of $\sigma^2(X^\top X + \lambda I)^{-1}$ and a mean of $\widehat{w}(\lambda)$ where $\lambda = \sigma^2/\tau^2$.

4. (8 points) Consider a linear prediction model of the form

$$\hat{y}(x) = w_0 + \sum_{j=1}^{d} w_j x_j$$

and recall that the OLS finds w by minimizing the empirical risk

$$\mathrm{Err}_D(w) = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}(x^{(i)}) - y^{(i)})^2$$

Now given the dataset $D = \{x^{(i)}, y^{(i)}\}_{i=1\ldots n}$ let us introduce a new random dataset $D' = \{x^{(i)} + \epsilon^{(i)}, y^{(i)}\}_{i=1\ldots n}$, where $\epsilon^{(i)} \sim \mathcal{N}(0, \tau^2 I_d)$. Show that minimizer of the following problem is $\hat{w}^{(\mathrm{Ridge})}$

$$\min_w \mathbb{E}_{\text{ffl}}[\mathrm{Err}_{D'}(w)],$$

where $\mathbb{E}_{\text{ffl}}$ denotes the expectation over $\epsilon^{(i)}$'s only. Derive dependence between τ^2 and λ in the ridge setup.

Solution: $\lambda = n\tau^2$; however $\lambda = \tau^2$ will also be accepted (see Note below).

The way that Ridge Regression is defined in the lecture notes (without any normalization constant), which is the standard definition, yields $\lambda = n\tau^2$ since at the end we must multiply through by n to get the Ridge Regression minimization problem. However, as stated above, for this problem not multiplying through by n will also be acceptable.

Key points: Expanding the error $\text{Err}_{D'}(w_0, w)$ and taking the expectation to show it produces the ridge regression minimization problem (using the fact that the noise is $\mathcal{N}(0, \tau^2 I_d)$ to cancel linear terms of the noise and compute quadratic terms of the noise).

We let $\widetilde{y}^{(i)}(x) = w_0 + w^\top(x + \varepsilon^{(i)})$. Then, letting $E^{(i)} = \varepsilon^{(i)}(\varepsilon^{(i)})^\top$:

$$\text{Err}_{D'}(w_0, w) = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{y}^{(i)}(x^{(i)}) - y^{(i)})^2 \tag{32}$$

$$= \frac{1}{n}\sum_{i=1}^{n}((\widehat{y}(x^{(i)}) - y^{(i)} + w^\top\varepsilon^{(i)})^2 \tag{33}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}(x^{(i)}) - y^{(i)})^2 + 2(\widehat{y}(x^{(i)}) - y^{(i)})w^\top\varepsilon^{(i)} + w^\top E^{(i)} w \tag{34}$$

$$\implies \mathbb{E}_\varepsilon[\text{Err}_{D'}(w_0, w)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\varepsilon^{(i)}}[(\widehat{y}(x^{(i)}) - y^{(i)})^2 + 2(\widehat{y}(x^{(i)}) - y^{(i)})w^\top\varepsilon^{(i)} + w^\top E^{(i)} w] \tag{35}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left((\widehat{y}(x^{(i)}) - y^{(i)})^2 + \tau^2\|w\|^2\right) \tag{36}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n}(\widehat{y}(x^{(i)}) - y^{(i)})^2\right) + \tau^2\|w\|^2 \tag{37}$$

which is the minimization problem solved by ridge regression with $\lambda = \tau^2$. Equation (36) is because $\varepsilon^{(i)}$ has mean 0_d, has independent components, and variance τ^2, so $\mathbb{E}[\varepsilon^{(i)}] = 0_d$ and $\mathbb{E}[E^{(i)}] = \tau^2 I_d$.

Multiplying through by n (to get the usual ridge regression expression) yields $\lambda = n\tau^2$.

6 Computational Problem 1: Regression Model Classes (10 points)

In this problem, we compare the performance of different regression models for a particular dataset. Please load the dataset on Canvas, and follow the notebook to generate graphs for the following regression models:

1. Linear regression.
2. Nearest-neighbor regression.
3. Neural network.
4. Linear regression in polynomial space.
5. Linear regression in Fourier feature space.

Using the results, answer the following questions.

(a) (2 points) After understanding the provided code for polynomial features, in 1 sentence, explain how X was transformed before passing into the linear regression model.

> **Solution:** Each x_i is transformed to $x_i, x_i^2, \cdots, x_i^k$ before passing into the regression model.

(b) (2 points) Which model gives the lowest training error?

> **Solution:** Answers vary based on hyperparameters chosen. Acceptable answers include Nearest-Neighbor Regression (if n_neighbor is low), neural net, or linear regression with Fourier features.

(c) (2 points) Approximately what value would each model predict for $x = 20$?

> **Solution:** Estimate for linear regression should be around 0; estimate for Nearest-Neighbor regression should be between -0.5 and 0.5 (varies based on the hyperparameter); estimate for linear regression with Fourier features should be between -1 and 1 (varies based on the hyperparameter).

(d) (4 points) Which model generalizes the best? Why? Please include the plot of this model from your notebook.[1]

> **Solution:** Regression in Fourier feature space generalizes the best. This is because the periodic form of Fourier features work particularly well for sinusoidal data.

7 Computational Problem 2: Bayesian Regression (10 points)

This problem explores Bayesian regression.

(a) (4 points) Complete the function for Bayesian regression in the notebook, and run the model for linear and quadratic data. Include the generated plots in your submission.

> **Solution:**

```python
def bayes_lin_reg(X_in, y, mu_w, sigma_w, sigma_y):
    X = np.hstack([np.ones((X_in.shape[0], 1)), X_in])

    sigma_w_post = np.linalg.inv(np.linalg.inv(sigma_w) + X.T @ X / sigma_y**2)
    mu_w_post = sigma_w_post @ (np.linalg.inv(sigma_w) @ mu_w + X.T @ y / sigma_y**2)

    def pred(X_in, return_std=False):
```

[1] Don't panic! We aren't going to check to see if you tuned the parameters super-carefully. This is all the general ideas.

```

        X = np.hstack([np.ones((X_in.shape[0], 1)), X_in])

        if return_std:
            return np.einsum('xd, da -> x', X, mu_w_post), \
                    np.sqrt(sigma_y**2 + np.einsum('xd,de,xe -> x', X, sigma_w_post
    , X))
        else:
            return X @ mu_w_post
    return RegressionResult((mu_w_post, sigma_w_post), pred)
```

Listing 1: Bayesian regression.

(b) (3 points) What factors contribute to the variance in posterior predictive distribution? How does that explain the standard deviations in your plots?

Solution: There are two factors contributing to the variance in the posterior predictive distribution:

1. Uncertainty in the parameter, and
2. Noise in the regression model itself.

Thus, note that even when we have a lot of data, there would still be variance in the posterior predictive distribution due to the second factor we identified above.

(c) (3 points) Run the model for sinusoidal dataset (with different amount of data). What property of the posterior model do you see in this case?

Solution: We observe there is higher uncertainty near places where there are more data, and the farther away we get from the data we observe, the larger the uncertainty is.