6.790 Homework 2

Revision: 9/20/24 3:53PM

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

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

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

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

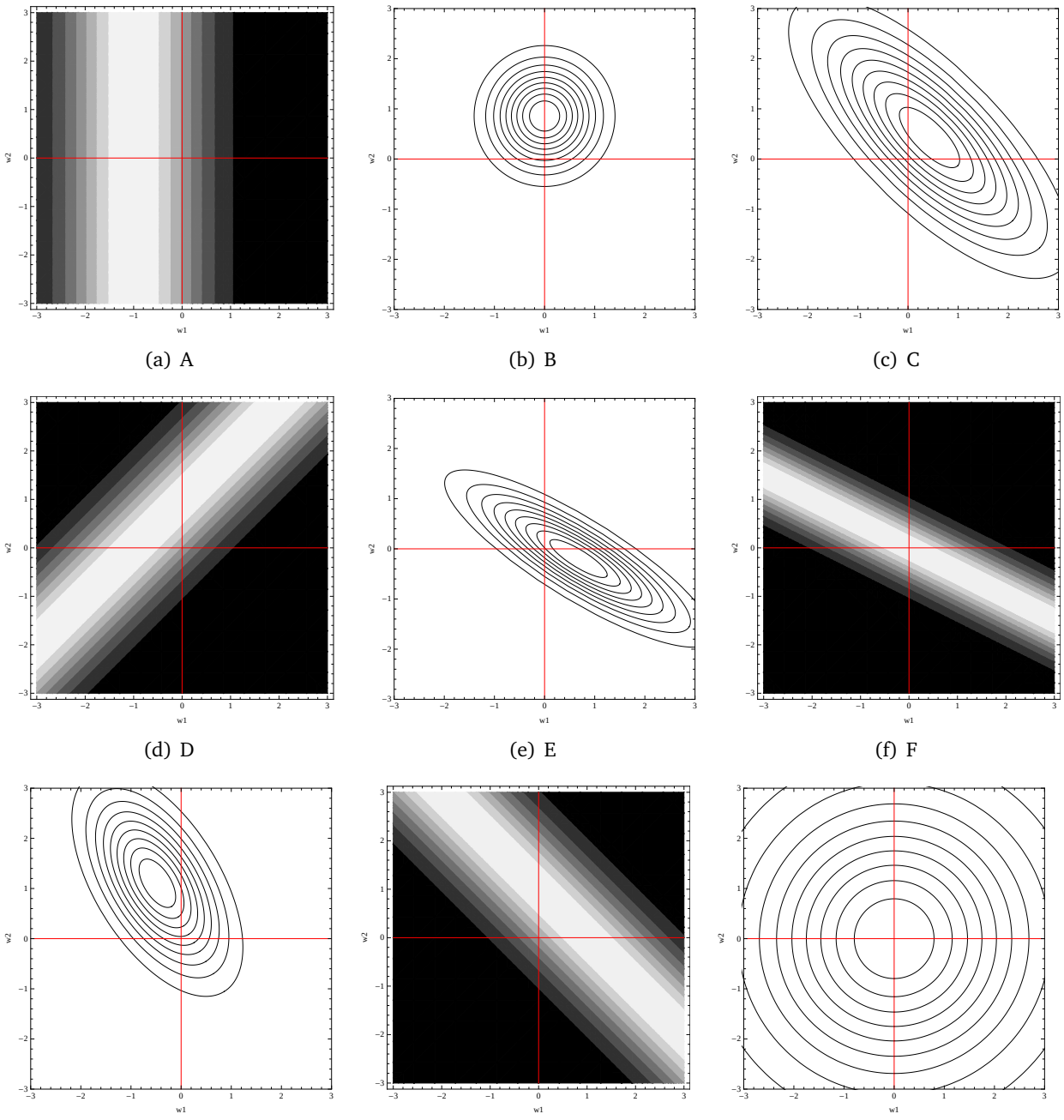
(c) $\Pr(\mathcal{D} = \{((1,-1),-1)\} \mid W)$ (Likelihood of one data point)
◯ A ◯ B ◯ C ◯ D ◯ E ◯ F ◯ G ◯ H ◯ I ◯ None

(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?

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?

(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?

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

(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?

(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 $\mu_{\mathbf{ml}}$?

(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?

(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.

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 $b/2$.

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

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

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?

(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.

(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?

(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?

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

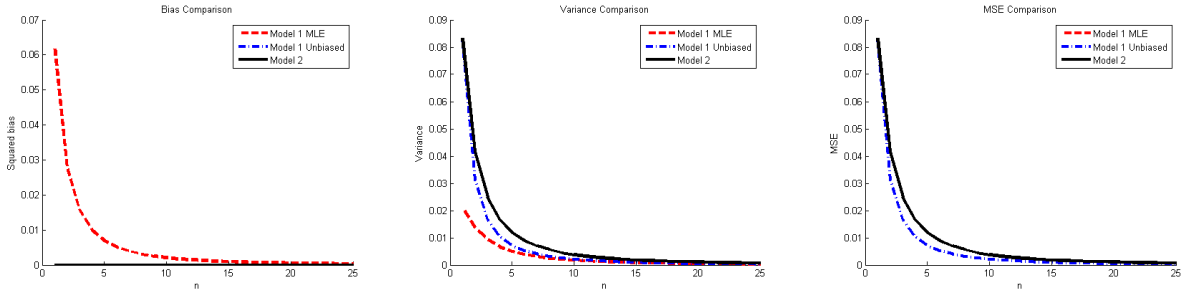
$$\begin{aligned} V[\mu_{\mathbf{ml}}] &= \int\int \left(\frac{x+y}{2}\right)^2 f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x, y)\, dxdy - [E[\mu_{\mathbf{ml}}]]^2 \\ &= \int_a^b \int_a^y \frac{(x+y)^2}{4} n(n-1) \frac{(y-x)^{n-2}}{(b-a)^n}\, dxdy - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{2(n+1)(n+2)} \end{aligned}$$

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

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.

(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 $\mu_{\mathbf{ml}} = x^{[n]}(n+1)/(2n)$ on data drawn from an interval $[a, b]$.

Here is the expected value of $\mu_{\mathbf{ml}}$:

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

Explain in English why this answer makes sense.

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

(d) (2 points) The variance of this $\mu_{\mathbf{ml}}$ is:

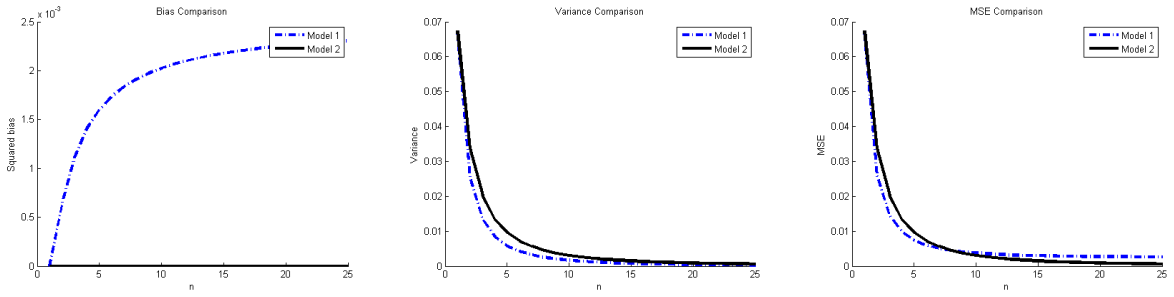
$$\begin{aligned} V[\mu_{\mathbf{ml}}] &= \int\int \left(\frac{y(n+1)}{2n}\right)^2 f_{a_{\mathbf{ml}},b_{\mathbf{ml}}}(x,y)\,dxdy - [E[\mu_{\mathbf{ml}}]]^2 \\ &= \int_a^b \int_a^y \frac{y^2(n+1)^2}{4n^2} n(n-1)\frac{(y-x)^{n-2}}{(b-a)^n}\,dxdy - \frac{(a+bn)^2}{4n^2} = \frac{(b-a)^2}{4n(n+2)} \end{aligned}$$

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)

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

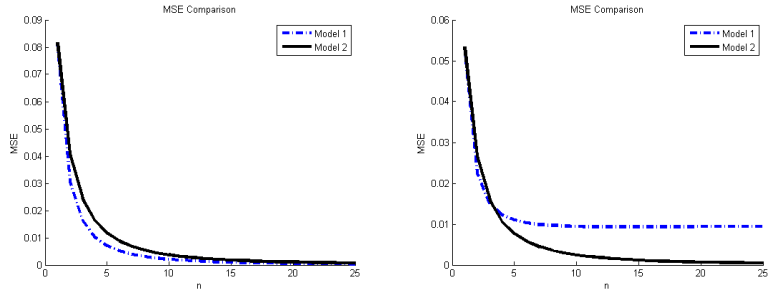
Figure 3 has plots that compare the bias, variance, and MSE of this estimator with the regular model 2 estimator on data drawn from $[0.1, 1]$, as a function of n.

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



(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.

(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.

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: $(x^{(i)}, y^{(i)})$, $i = 1, \ldots, n$. 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?

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.

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.

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

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

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 $\hat{w}^{(Ridge)}$

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

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

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.

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

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

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

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.

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

(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?

[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.