1.1. Number of samples for uniqueness: How large does n need to be compared to d for it to be *possible* for X^TX to have full rank d? When this condition is not satisfied, how should we modify the formulation (1)?

We have $\mathrm{rank}(X^TX) \le n$, so we need $n \ge d$ for it to be possible for $\mathrm{rank}(X^TX) = d$. When this condition is not satisfied, we can add regularization, solving

$$\min_w \sum_{i=1}^n (x_i^T w - y_i)^2 + \rho(w).$$

1.2. Unbiased estimates: Suppose that $y = Xw_{\text{true}} + z$, with $z \sim \mu$ a random (noise) vector with probability distribution μ. Under what conditions on z, is the least squares solution *unbiased*, i.e.,

$$\mathbb{E}\Big[w_{\text{LS}}\Big] = w_{\text{true}}? \tag{5}$$

Note: in lecture we showed that if $z \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$, w_{LS} is unbiased. However, this holds more broadly – please make your conditions as broad as possible for full credit!

Note that

$$\begin{aligned}
\mathbb{E}\Big[w_{\text{LS}}\Big] &= \mathbb{E}\Big[(X^TX)^{-1}X^Ty\Big] \\
&= \mathbb{E}\Big[(X^TX)^{-1}X^T(Xw_{\text{true}} + z)\Big] \\
&= w_{\text{true}} + (X^TX)^{-1}X^T\mathbb{E}[z].
\end{aligned}$$

The estimator is unbiased as long as the noise is zero mean, i.e., $\mathbb{E}[Z] = 0$.[1]

1.3. Least squares by gradient descent: We can also compute w_{LS} iteratively, using gradient descent – namely, by choosing some initial guess w^0 and updating w as

$$w^{k+1} = w^k - t\nabla\mathcal{L}_{\text{LS}}(w^k). \tag{6}$$

Part A. Show that

$$w^{k+1} - w_{\text{LS}} = \left(I - 2tX^TX\right)\left(w^k - w_{\text{LS}}\right) \tag{7}$$

Hint: use our expressions for ∇L_{LS} and w_{LS}.

We have

$$\begin{aligned}
w_{\text{LS}} &= (X^TX)^{-1}X^Ty \\
\nabla\mathcal{L}_{\text{LS}}(w) &= 2X^T(Xw - y) \\
&= 2(X^TX)w - 2X^Ty \\
&= 2(X^TX)w - 2(X^TX)w_{\text{LS}} \\
&= 2(X^TX)(w - w_{\text{LS}})
\end{aligned}$$

So,

$$\begin{aligned}
w^{k+1} - w_{\text{LS}} &= (w^k - t\nabla\mathcal{L}(w^k)) - w_{\text{LS}} \\
&= w^k - w_{\text{LS}} - 2t(X^TX)(w^k - w_{\text{LS}}) \\
&= (I - 2tX^TX)(w^k - w_{\text{LS}}).
\end{aligned}$$

Part B. Let $\lambda_1 \geq \cdots \geq \lambda_n$ denote the eigenvalues of the symmetric matrix $I - 2tX^TX$. Argue that if t is chosen such that $|\lambda_i| < 1$ for all i, then as $k \to \infty$, $w^k \to w_{\text{LS}}$.

Hint: You can use the following inequality. For a symmetric matrix M with eigenvalues $\lambda_1, \ldots \lambda_n$, and a vector x,

$$\|Mx\|_2 \leq \left(\max_i |\lambda_i|\right) \|x\|_2.$$

We have that

$$\begin{aligned} \|w^{k+1} - w_{\text{LS}}\|_2 &\leq \left(\max_i |\lambda_i|\right) \|w^k - w_{\text{LS}}\|_2 \\ &\leq \left(\max_i |\lambda_i|\right)^2 \|w^{k-1} - w_{\text{LS}}\|_2 \\ &\leq \left(\max_i |\lambda_i|\right)^{k+1} \|w^0 - w_{\text{LS}}\|_2. \end{aligned}$$

Because $\max_i |\lambda_i|$ is strictly less than 1, the right hand side above goes to zero as $k \to \infty$, and so $\|w^k - w_{\text{LS}}\|_2 \to 0$ and $w^k \to w_{\text{LS}}$.

2.1. Demonstrate that the ridge regression solution w_{RR} is given by

$$w_{\text{RR}} = \left(\lambda I + X^TX\right)^{-1} X^Ty.$$

Hint: differentiate the objective function in (8) and set the derivative equal to zero.

Let

$$\mathcal{L}_{\text{RR}}(w) = \sum_{i=1}^{n} (x_i^T w - y_i)^2 + \lambda w^T w.$$

We have

$$\begin{aligned} \nabla \mathcal{L}_{\text{RR}} &= 2X^T(Xw - y) + 2\lambda w \\ &= 2(X^TX + \lambda I)w - 2X^Ty. \end{aligned}$$

We have $\nabla \mathcal{L}_{\text{RR}} = 0$ if and only if

$$(X^TX + \lambda I)w = X^Ty,$$

i.e., $w = (X^TX + \lambda I)^{-1}X^Ty$.

2.2. Consider two potential values of λ: $\lambda = 0.1$ and $\lambda = 1,000$. For which value of λ do you expect the ridge regression solution w_{RR} to exhibit greater *bias*? For which value of λ do you expect it to exhibit greater *variance*?

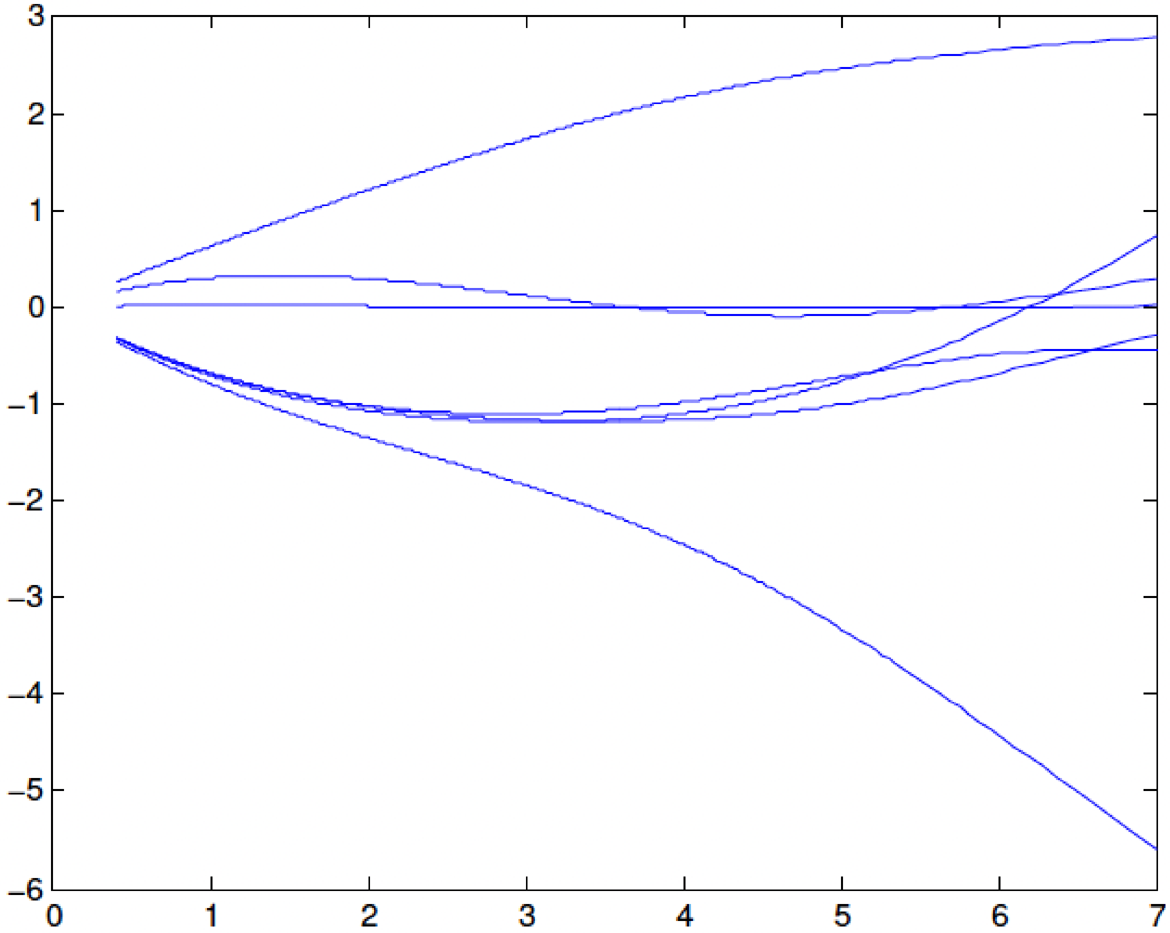
$\lambda = 1,000$ will exhibit greater bias – the coefficient vector w exhibits more shrinkage towards zero. $\lambda = 0.1$ will exhibit greater variance.

2.3. Describe a procedure for choosing λ in practice.

λ can be chosen by cross-validation: one randomly splits the given dataset into training and validation sets, using the training set calculates the ridge regression solution $w_{\mathrm{RR}}^{\lambda_i}$ for a range of values $\lambda_1 < \lambda_2 < \cdots < \lambda_N$, and choses the λ_i that minimizes the mean square error over the validation set. This procedure can be extended to k-fold cross validation, in which one partitions the training set into k subsets, and chooses the value of λ which minimizes the average error, over all choices of the validation set, when the model is fit using the remaining $k - 1$ subsets.

Problem 3

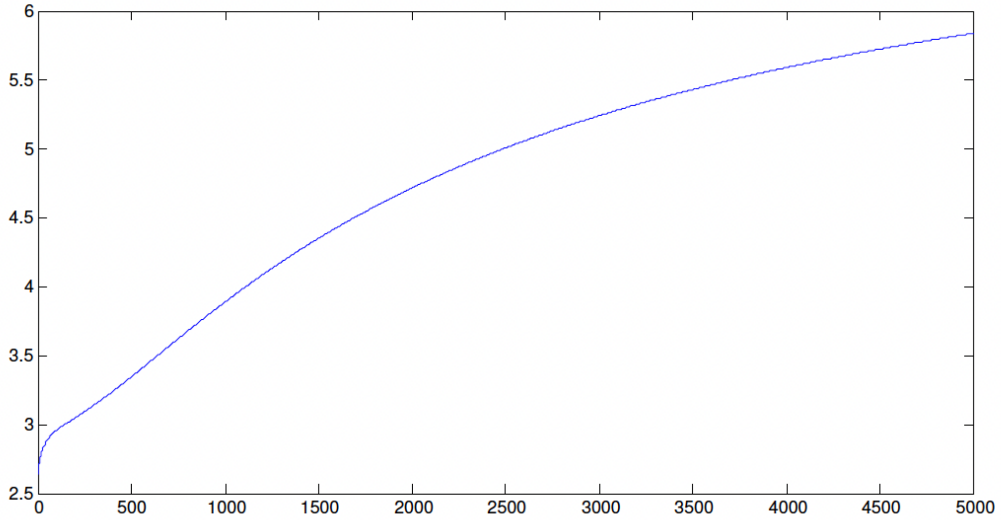
(a)



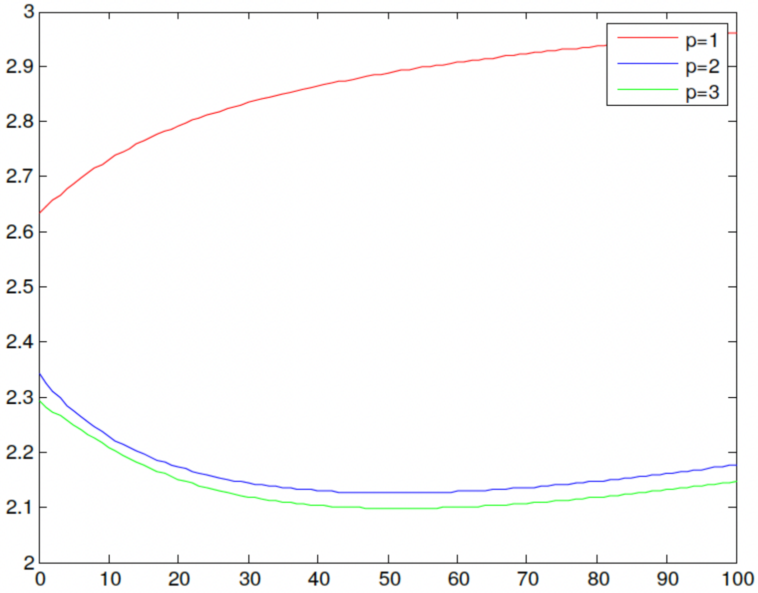
(a) Correct figure as function of $df(\lambda)$

(b) The value of the 4th dimension is consistently very negative. This indicates that as car weight increases, gas mileage decreases. The value of the 6th dimension is consistently very positive. This indicates that newer cars tend to have much better gas mileage.

(c) The plot indicates that $\lambda = 0$ performs the best. As a result we can say that least squares is better than ridge regression for this linear regression setup because $\lambda = 0$ is least squares and $\lambda > 0$ is linear regression.



(d) The values of p we should choose is $p = 3$ because the performance is best for $p = 3$ at the best value of λ. For $p = 3$, the best value is $\lambda \approx 51$.


p=1
p=2
p=3
3
2.9
2.8
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2
0
10
20
30
40
50
60
70
80
90
100