HOMEWORK 1 SOLUTIONS

Problem 1

(a)

$$p(x_1, \ldots, x_N|\lambda) = \prod_{i=1}^{N} p(x_i|\lambda) = \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} \mathrm{e}^{-\lambda}$$

(b)

$$\lambda_{\text{ML}} = \arg\max_{\lambda} \ln p(x_1, \ldots, x_N|\lambda) = \sum_{i=1}^{N} x_i \ln \lambda - \lambda - \ln x_i!$$

$$\nabla_\lambda \ln p(x_1, \ldots, x_N|\lambda) = \sum_{i=1}^{N} (\frac{x_i}{\lambda} - 1) = 0$$

$$\lambda_{\text{ML}} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

(c)

$$\lambda_{\text{MAP}} = \arg\max_{\lambda} \ln p(x_1, \ldots, x_N|\lambda) + \ln p(\lambda) = (a-1) \ln \lambda - b\lambda + \sum_{i=1}^{N} x_i \ln \lambda - \lambda + const.$$

$$\nabla_\lambda \ln p(x_1, \ldots, x_N, \lambda) = \frac{a-1}{\lambda} - b + \sum_{i=1}^{N} (\frac{x_i}{\lambda} - 1) = 0$$

$$\lambda_{\text{MAP}} = \frac{1}{b+N} \left(a - 1 + \sum_{i=1}^{N} x_i \right)$$

(d)

$$p(\lambda|x_1, \ldots, x_N) \propto p(x_1, \ldots, x_N|\lambda) p(\lambda) \propto \lambda^{a-1} \mathrm{e}^{-b\lambda} \prod_{i=1}^{N} \lambda^{x_i} \mathrm{e}^{-\lambda}$$

We can recognize that

$$\lambda^{a-1+\sum_{i=1}^{N} x_i} \mathrm{e}^{-(b+N)\lambda} \propto \text{gamma}(a', b')$$

So the posterior is a gamma distribution with $a' = a + \sum_{i=1}^{N} x_i$ and $b' = b + N$.

(e) Using the definition of a' and b' above, $\mathbb{E}[\lambda] = \frac{a'}{b'}$ and $Var(\lambda) = \frac{a'}{b'^2}$. The key observation is how these things change as functions of N. The expectation equals MAP in this case, and as $N \to \infty$ MAP, ML and the expectation converge to the same value. The variance decreases as N increases, meaning we are more confident in the posterior as we get more data. This last point is a major advantage of a Bayesian method.

Problem 2

$$\mathbb{E}[w_{\text{RR}}] = \mathbb{E}[(\lambda I + X^T X)^{-1} X^T y] = (\lambda I + X^T X)^{-1} X^T \mathbb{E}[y] = (\lambda I + X^T X)^{-1} X^T X w$$

$$Var(w_{\text{RR}}) = \mathbb{E}[w_{\text{RR}} w_{\text{RR}}^T] - \mathbb{E}[w_{\text{RR}}]\mathbb{E}[w_{\text{RR}}]^T$$
$$= (\lambda I + X^T X)^{-1} X^T \mathbb{E}[yy^T] X (\lambda I + X^T X)^{-1} - (\lambda I + X^T X)^{-1} X^T \mathbb{E}[y]\mathbb{E}[y]^T X (\lambda I + X^T X)^{-1}$$

Since $\mathbb{E}[yy^T] = \sigma^2 I + Xww^T X^T$ and $\mathbb{E}[y] = Xw$, we have a cancellation of two terms involving $Xww^T X^T$ and we are left with

$$Var(w_{\text{RR}}) = \sigma^2 (\lambda I + X^T X)^{-1} X^T X (\lambda I + X^T X)^{-1}$$

A final observation is that

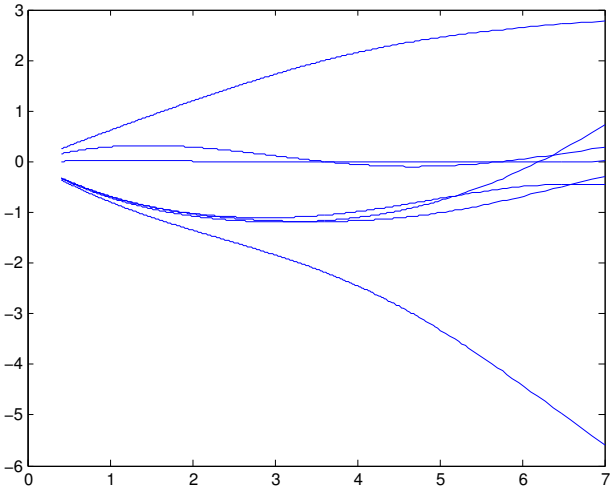
$$(\lambda I + X^T X)^{-1} = (\lambda (X^T X)^{-1} + I)^{-1} (X^T X)^{-1} = (X^T X)^{-1} (\lambda (X^T X)^{-1} + I)^{-1}$$

Therefore,

$$Var(w_{\text{RR}}) = \sigma^2 (\lambda (X^T X)^{-1} + I)^{-1} (X^T X)^{-1} (\lambda (X^T X)^{-1} + I)^{-1}$$

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 $\lambda \approx 51$.

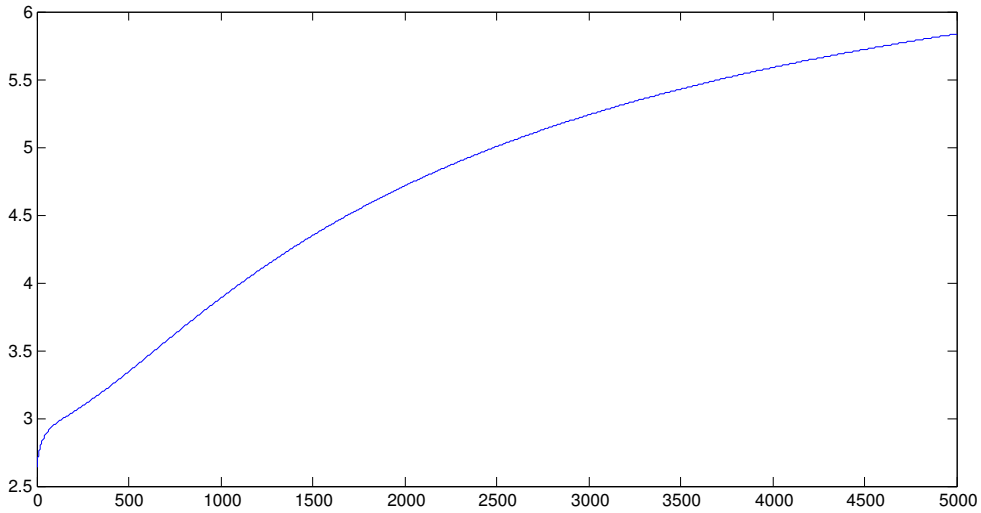


Figure 1: Figure for (c)

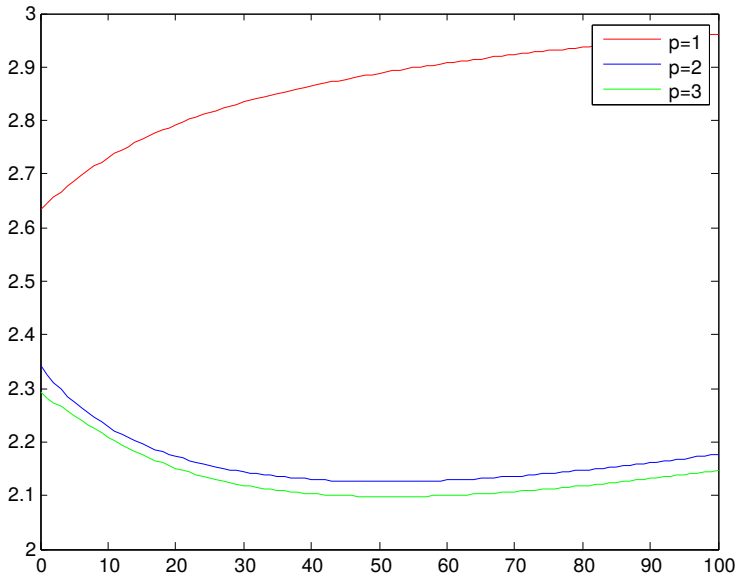


Figure 2: Figure for (d)