Problem 1

Imagine you have sequence of N observations $(x_1, \cdots, x_N)$, where each $x_i \in \{0,1,2,\cdots,\infty\}$. You model this sequence as i.i.d from a Poisson distribution with unknown parameter $\lambda \in \mathbb{R}_+, where$

$$p(X \mid \lambda) = \frac{\lambda^X}{X!} e^{-\lambda}$$

(a) **Joint likelihood of the data** $(x_1, \cdots, x_N)$

$$\begin{aligned} p(x_1, \cdots, x_n \mid \lambda) &= \prod_{i=1}^{n} p(x_i \mid \lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \\ &= \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \cdots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} \\ &= \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} e^{-\lambda N} \end{aligned}$$

(b) **Maximum likelihood estimate** λ_{ML}

$$\begin{aligned} \lambda_{ML} &= \arg\max_{\lambda} \prod_{i=1}^{n} p(x_i \mid \lambda) \\ &= \arg\max_{\lambda} \ln\left(\prod_{i=1}^{n} p(x_i \mid \lambda)\right) \\ &= \arg\max_{\lambda} \sum_{i=1}^{n} \ln p(x_i \mid \lambda) \end{aligned}$$

Setting the derivative of the joint likelihood with respect to λ equal to 0,

$$
\begin{aligned}
\nabla_\lambda \sum_{i=1}^{n} \ln p(x_i \mid \lambda) &= 0 \\
\sum_{i=1}^{n} \nabla_\lambda \ln \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} &= 0 \\
\sum_{i=1}^{n} \nabla_\lambda \ln \frac{\lambda^{x_i}}{x_i!} + \ln e^{-\lambda} &= 0 \\
\sum_{i=1}^{n} \nabla_\lambda \left(x_i \ln \lambda - \ln x_i! - \lambda \right) &= 0 \\
\sum_{i=1}^{n} \frac{x_i}{\lambda} - 1 &= 0 \\
\frac{1}{\lambda} \sum_{i=1}^{n} x_i - N &= 0 \\
\lambda_{ML} &= \frac{1}{N} \sum_{i=1}^{n} x_i
\end{aligned}
$$

(c) **Maximum a posteriori (MAP) estimate** λ_{MAP}

To help learn λ, we use a prior distribution $p(\lambda) = gamma(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$.

λ_{MAP} seeks the most probable value of λ according to its posterior distribution $p(\lambda \mid X)$,

$$
\begin{aligned}
\lambda_{MAP} &= \arg\max_\lambda \ \ln p(\lambda \mid X) \\
&= \arg\max_\lambda \ \ln \frac{p(X \mid \lambda)\, p(\lambda)}{p(X)} \\
&= \arg\max_\lambda \ \ln p(X \mid \lambda) + \ln p(\lambda) \\
&= \arg\max_\lambda \ \ln \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} e^{-\lambda N} + \ln \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \\
&= \arg\max_\lambda \ \sum_{i=1}^{n} x_i \ln \lambda - \ln \prod_{i=1}^{n} x_i! - \lambda N + \ln \frac{\beta^\alpha}{\Gamma(\alpha)} + \alpha - 1 \ln \lambda - \beta\lambda
\end{aligned}
$$

Taking the derivative ∇_λ and setting it equal to 0,

$$
\begin{aligned}
\frac{\sum_{i=1}^{n} x_i}{\lambda} - N + \frac{\alpha - 1}{\lambda} - \beta &= 0 \\
(N + \beta)\lambda &= \sum_{i=1}^{n} x_i + (\alpha - 1) \\
\lambda_{MAP} &= \frac{\sum_{i=1}^{n} x_i + (\alpha - 1)}{N + \beta}
\end{aligned}
$$

(d) **Posterior distribution** $p(\lambda \mid X)$

$$\begin{aligned} p(\lambda \mid X) &\propto \; p(X \mid \lambda)p(\lambda) \\ &\propto [\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} e^{-\lambda N}] \, [\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}] \\ &\propto \frac{\beta^\alpha}{\prod_{i=1}^{n} x_i! \Gamma(\alpha)} \; \lambda^{(\sum_{i=1}^{n} x_i + \alpha)-1} \; e^{-(\beta+N)\lambda} \end{aligned}$$

This posterior distribution is essentially $p(\lambda \mid X) = gamma(\sum_{i=1}^{n} x_i + \alpha, \beta + N)$ with $\alpha_{new} = \sum_{i=1}^{n} x_i + \alpha$ and $\beta_{new} = \beta + N$.

(e) **Mean and variance of** λ

Mean of gamma distribution $\mathbb{E}(\lambda) = \frac{\alpha}{\beta}$ and $Var(\lambda) = \frac{\alpha}{\beta^2}$

$$\begin{aligned} \mathbb{E}(\lambda) &= \frac{\sum_{i=1}^{n} x_i + \alpha}{\beta + N} \\ Var(\lambda) &= \frac{\sum_{i=1}^{n} x_i + \alpha}{(\beta + N)^2} \end{aligned}$$

We can observe that when $\alpha = 0$ and $\beta = 0$, $\mathbb{E}(\lambda)$ and λ_{ML} would be equal. Also, under the posterior distribution, $\mathbb{E}(\lambda) > \lambda_{MAP}$.

Problem 2

Ridge regression $w_{RR} = (\lambda I + X^T X)^{-1} X^T y$

(a) $\mathbb{E}(w_{RR})$

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

$$\begin{aligned} \mathbb{V}[w_{RR}] &= \mathbb{E}[(w_{RR} - \mathbb{E}[w_{RR}])(w_{RR} - \mathbb{E}[w_{RR}])^T] \\ &= \mathbb{E}[w_{RR} {w_{RR}}^T] - \mathbb{E}[w_{RR}]\mathbb{E}[w_{RR}]^T \\ &= \mathbb{E}[(\lambda I + X^T X)^{-1} X^T y y^T X (\lambda I + X^T X)^{-1}] - (\lambda I + X^T X)^{-1} X^T X w w^T X^T X (\lambda I + X^T X)^{-1} \\ &= (\lambda I + X^T X)^{-1} X^T \mathbb{E}[y y^T] X (\lambda I + X^T X)^{-1} - (\lambda I + X^T X)^{-1} X^T X w w^T X^T X (\lambda I + X^T X)^{-1} \\ &= (\lambda I + X^T X)^{-1} X^T (\sigma^2 I + X w w^T X^T) X (\lambda I + X^T X)^{-1} - (\lambda I + X^T X)^{-1} \cdots \\ &= (\lambda I + X^T X)^{-1} X^T \sigma^2 I X (\lambda I + X^T X)^{-1} \\ &= \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} \end{aligned}$$

(b) Relationship of w_{RR} to w_{LS}

$$\begin{aligned} w_{RR} &= (\lambda I + X^T X)^{-1} X^T y \\ &= (\lambda I + X^T X)^{-1} (X^T X)(X^T X)^{-1} X^T y \\ &= [(X^T X)(\lambda (X^T X)^{-1} + I)]^{-1} (X^T X) w_{LS} \\ &= (\lambda (X^T X)^{-1} + I)^{-1} (X^T X)^{-1} (X^T X) w_{LS} \\ &= (\lambda (X^T X)^{-1} + I)^{-1} w_{LS} \end{aligned}$$

Using SVD, $(X^T X)^{-1} = V S^{-2} V^T$

$$\begin{aligned} &= (\lambda ((V S^{-2} V^T) + I)^{-1} w_{LS} \\ &= V (\lambda S^{-2} + I)^{-1} V^T w_{LS} \end{aligned}$$

Problem 3

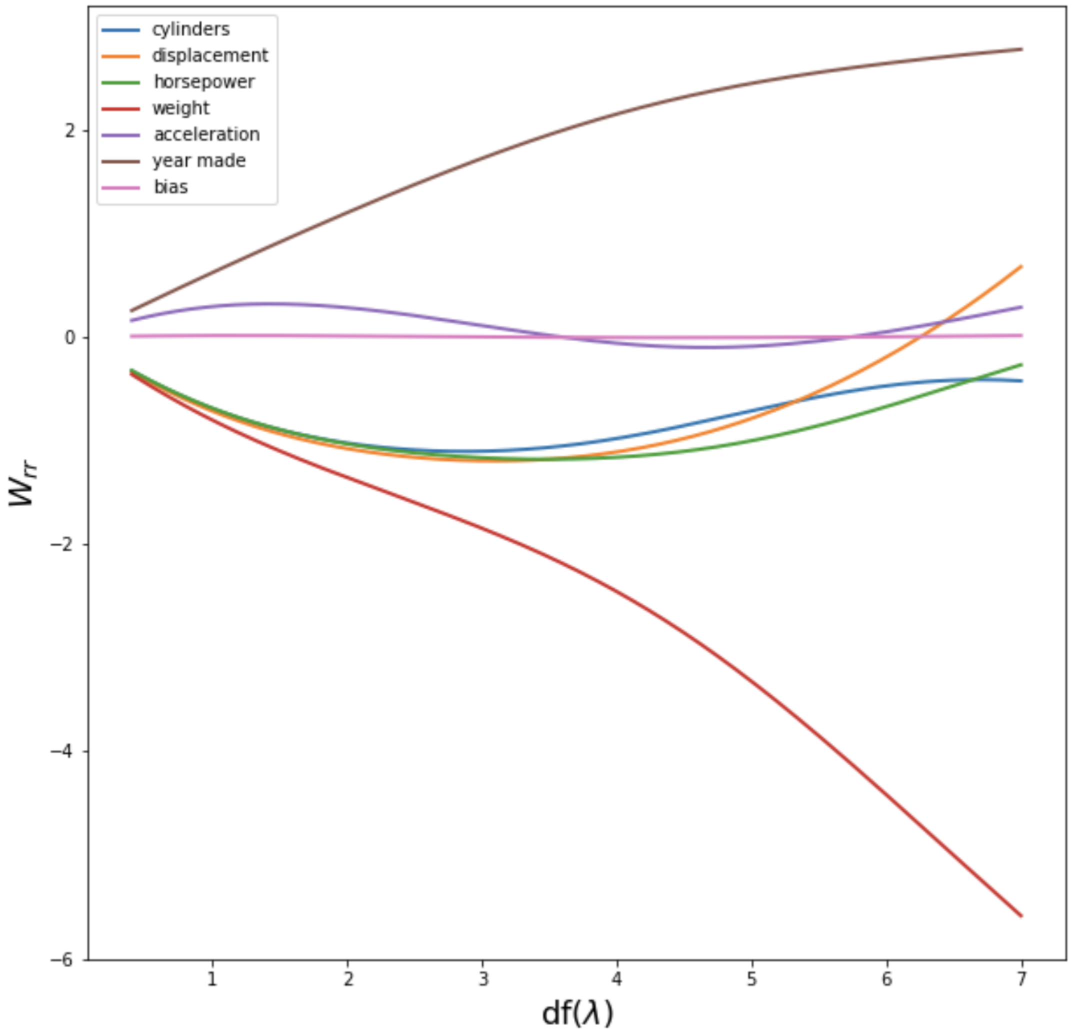

Figure 1: Ridge Regression

(a) w_{RR} vs. $df(\lambda)$

(b) The two features that stand out are ”year made” and ”weight” and they play an important role in determining the miles per gallon of the car. We can also observe that there is positive relationship between year-made of the car and its mileage. This is reasonable as new cars have higher mileage. Also, there is negative relationship that heavier (large weight) cars have smaller mileage. As λ increases and $df(\lambda)$ decreases, these two features' relationship to mileage remains unchanged as they cross the zero axis.

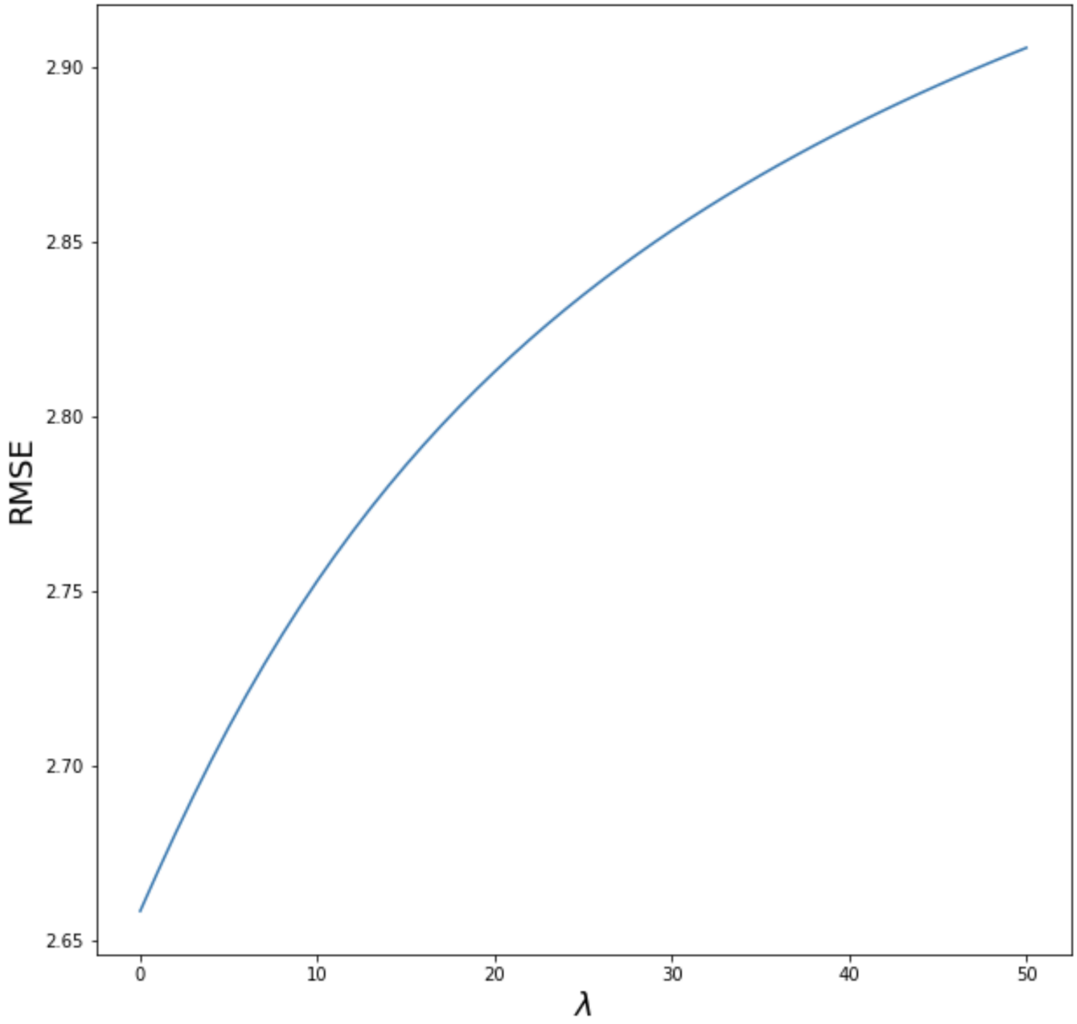

Figure 2: RMSE vs λ

(c) We observe that as λ increases, the RMSE also increases. The minimum value RMSE appears when $\lambda = 0$ which is the least squares solution. Therefore, we choose the least squares solution instead of ridge regression.

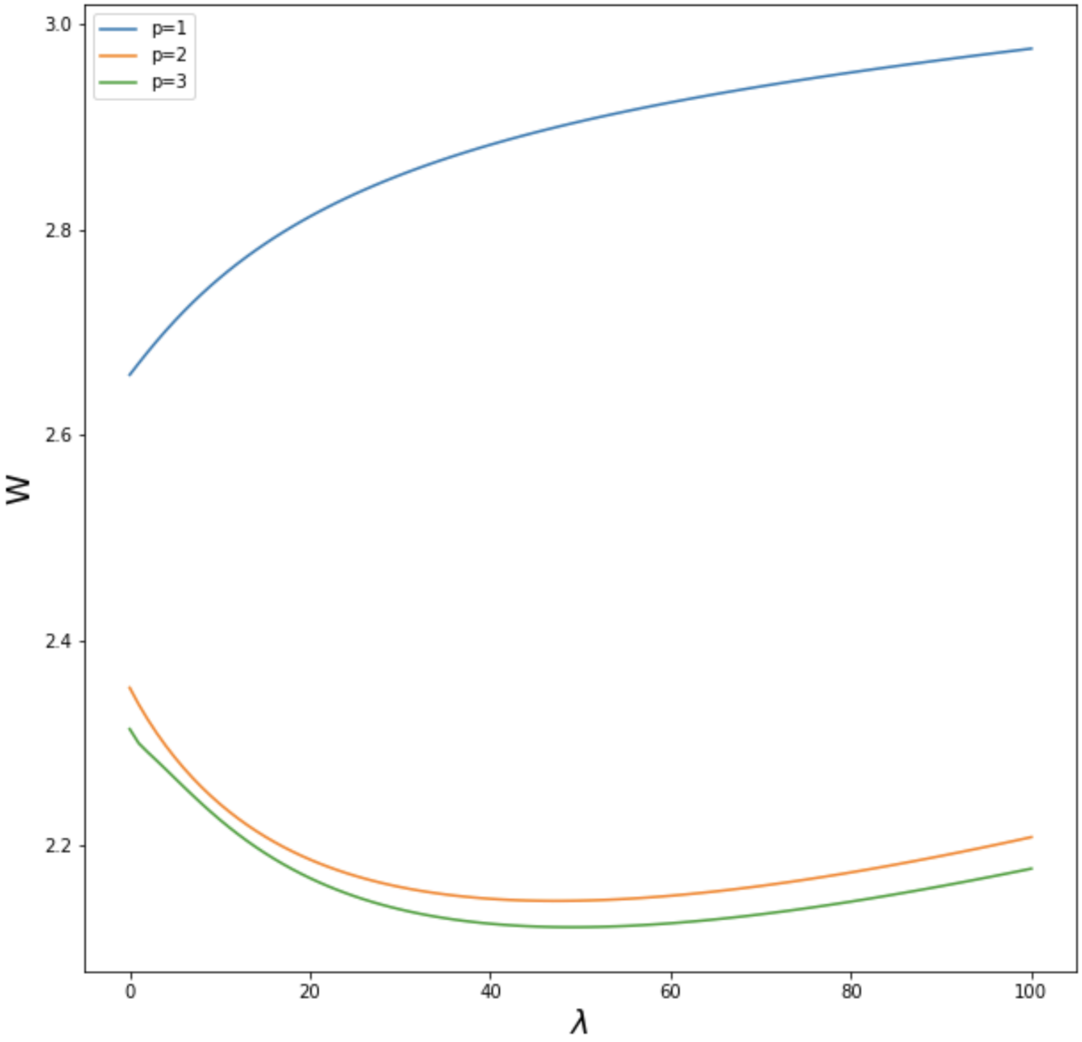

Figure 3: pth-order polynomials

(d) Unlike for $p = 1$, the RMSE decreases with increase in λ for both $p = 2$ and $p = 3$. Least squares is no longer the desired solution because the new minimum RMSE can be obtained when choosing $p = 3$ at around a value of $\lambda = 42$.