# Question 1

a) Let $f(m) = \frac{1}{n}\sum_{i=1}^{n}|Y_i - m|^2$, which is equivalent to saying $f(m) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - m)^2$. We want to find the value of $m$ such that $f(m)$ is minimized. Thus, we take the derivative with respect to $m$ and set it equal to 0:

$$\begin{aligned}
\frac{d}{dm}f(m) &= \frac{d}{dm}\frac{1}{n}\sum_{i=1}^{n}(Y_i - m)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{dm}(Y_i - m)^2 \\
&= -\frac{2}{n}\sum_{i=1}^{n}Y_i - m \\
&= -\frac{2}{n}(\sum_{i=1}^{n}Y_i - \sum_{i=1}^{n}m)
\end{aligned}$$

then by setting $\frac{d}{dm}f(m) = 0$, it is easy to see that $\sum_{i=1}^{n}Y_i = \sum_{i=1}^{n}m = nm$, which then implies that $m = \frac{1}{n}\sum_{i=1}^{n}Y_i = h_{avg}(\mathcal{D})$. To show that this is indeed the minimum, we apply the second derivative test:

$$\begin{aligned}
\frac{d^2}{dm^2}f(m = h_{avg}(\mathcal{D})) &= \frac{d}{dm} - \frac{2}{n}\sum_{i=1}^{n}Y_i - m \\
&= -\frac{2}{n}\sum_{i=1}^{n}\frac{d}{dm}(Y_i - m) \\
&= \frac{2}{n}\sum_{i=1}^{n}1 = 2 > 0
\end{aligned}$$

hence, $m = h_{avg}$ produces a minimum. Thus

$$h_{avg} = \frac{1}{n}\sum_{i=1}^{n}Y_i = \arg\min_{m\in\mathbb{R}}\frac{1}{n}\sum_{i=1}^{n}|Y_i - m|^2$$

as wanted.

b) We note that:

$$\begin{aligned}
E[h_{avg}(D)] &= E[\frac{1}{n}\sum_{i=1}^{n}Y_i] \\
&= \frac{1}{n}\sum_{i=1}^{n}E[Y_i] = \frac{1}{n}(n\mu) = \mu
\end{aligned}$$

Hence, bias is $(E[h_{avg}(D)] - u)^2 = 0$. For variance, we note that:

$$\begin{aligned}
\text{Var}(h_{avg}(D)) &= \text{Var}(\frac{1}{n}\sum_{i=1}^{n} Y_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(Y_i) \qquad \text{since each } Y_i \text{ is independent} \\
&= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}
\end{aligned}$$

c) Let $f(m) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - m|^2 + \lambda|m|^2$ , which is equivalent to saying $f(m) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - m)^2 + \lambda m^2$. We want to find the value of $m$ such that $f(m)$ is minimized. Thus, we take the derivative with respect to $m$ and set it equal to 0:

$$\begin{aligned}
\frac{d}{dm}f(m) &= \frac{d}{dm}\frac{1}{n}\sum_{i=1}^{n}(Y_i - m)^2 + \lambda m^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{dm}(Y_i - m)^2 + \frac{d}{dm}\lambda m^2 \\
&= -\frac{2}{n}(\sum_{i=1}^{n} Y_i - m) + 2\lambda m \\
&= -\frac{2}{n}(\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} m) + 2\lambda m \\
&= -\frac{2}{n}(\sum_{i=1}^{n} Y_i - nm) + 2\lambda m \\
&= -\frac{2}{n}(\sum_{i=1}^{n} Y_i) + 2m + 2\lambda m \\
&= -\frac{2}{n}(\sum_{i=1}^{n} Y_i) + 2m(1 + \lambda)
\end{aligned}$$

Then by setting $\frac{d}{dm}f(m) = 0$, we get that $m = (\frac{1}{1+\lambda})\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{h_{avg}}{1+\lambda}$. To see that this value is indeed the minimum we take the second derivative:

$$\begin{aligned}
\frac{d^2}{dm^2}f(m = \frac{h_{avg}}{1+\lambda}) &= \frac{d}{dm} - \frac{2}{n}(\sum_{i=1}^{n} Y_i) + 2m(1 + \lambda) \\
&= 0 + 2(1 + \lambda) = 2(1 + \lambda) > 0
\end{aligned}$$

Hence, $m = \frac{h_{avg}}{1+\lambda}$ is the explicit formula for the estimator.

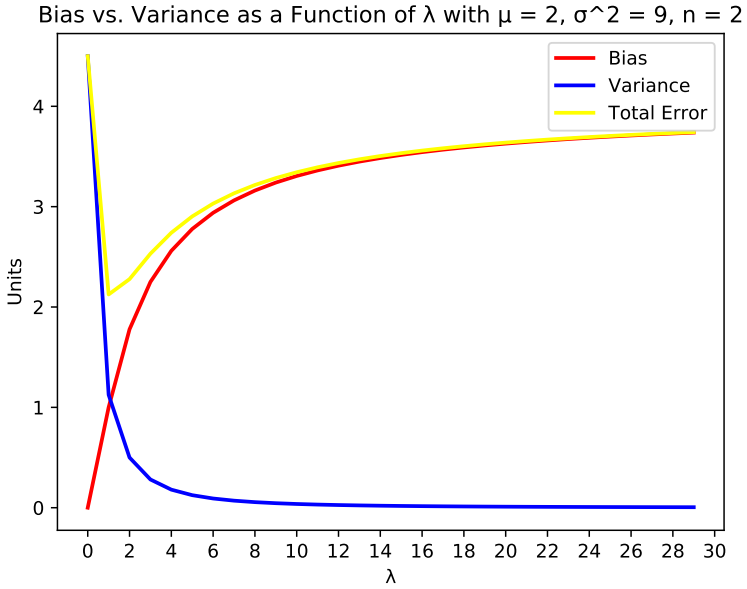
d) First, we note that the expectation of $h_\lambda(\mathcal{D})$ is the following:

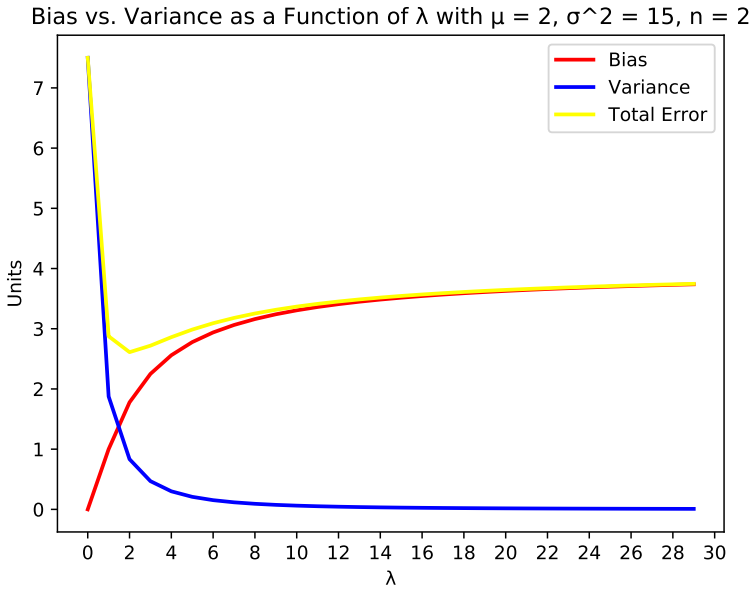
$$\begin{aligned} E[h_\lambda(\mathcal{D})] &= E[\frac{h_{avg}(\mathcal{D})}{1+\lambda}] \\ &= \frac{1}{1+\lambda} E[h_{avg}] \\ &= \frac{\mu}{1+\lambda} \qquad \text{by earlier calculations} \end{aligned}$$

Hence we get that the bias is $(\frac{\mu}{1+\lambda} - \mu)^2 = \frac{\lambda^2\mu^2}{(1+\lambda)^2}$.
We note that the variance of $h_\lambda(\mathcal{D})$ is the following:

$$\begin{aligned} \text{Var}(h_\lambda(\mathcal{D})) &= \text{Var}(\frac{h_{avg}(\mathcal{D})}{1+\lambda}) \\ &= \frac{1}{(1+\lambda)^2} \text{Var}(h_{avg}(\mathcal{D})) \\ &= \frac{\sigma^2}{n(1+\lambda)^2} \qquad \text{by earlier calculations} \end{aligned}$$

e) More graphs are available when running the code q1.py that was submitted. Here are a few:






f) As we increase the value of $\lambda$ we ensure that the variance is lowered by punishing high weights. This way the function has low variability in the values that it predicts. However, if we increase $\lambda$ too much then we put too much emphasis on low weight values and hence increase the bias of the estimator to drift away from the true mean value. We also note that on very low values of $\lambda$ we have very high variance and low bias. The reason we expect higher variance is because we allow our weight to increase, which means there is more variability in our prediction values even with slight changes in our data point value. Thus, we need to find an optimal value for $\lambda$ that will not only lower the variance of our estimator but limit how much it biases the estimator. This way we can minimize the expected square loss function, and thus optimizing our estimator. This "perfect" value for $\lambda$ can be seen as the argmin of the total error line in the above graphs. Another observation we make is that as $\lambda$ grows larger and larger, the variance approaches 0, and the bias approaches some positive number or perhaps infinity. This suggest that with increasing $\lambda$, we are forcing our model to make a certain prediction without considering the actual test point.

# Question 2

b) The number of data points in the Boston House data-set is 506, with each data point having 13 features (hence dimension is 13), and the target, $t^{(i)}$ value represents the median value of owner-occupied homes in 1000's given the features, $x^{(i)}$. We define what each of the features mean:
CRIM - per capita crime rate by town
ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS - proportion of non-retail business acres per town.
CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX - nitric oxides concentration (parts per 10 million)
RM - average number of rooms per dwelling
AGE - proportion of owner-occupied units built prior to 1940
DIS - weighted distances to five Boston employment centres
RAD - index of accessibility to radial highways
TAX - full-value property-tax rate per $10,000
PTRATIO - pupil-teacher ratio by town
B - $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT -% lower status of the population

c) The plot can be seen when running code q2.py. The graph is omitted from this report because of the space it would occupy.

e)

| Features | Weights |
|---|---|
| BIAS | 23.536 |
| CRIM | -6.907 |
| ZN | 2.943 |
| INDUS | 2.172 |
| CHAS | 2.99 |
| NOX | -11.811 |
| RM | 23.474 |
| AGE | 0.77 |
| DIS | -14.359 |
| RAD | 6.183 |
| TAX | -5.625 |
| PTRATIO | -9.047 |
| B | 4.306 |
| LSTAT | -18.336 |

The sign of the weight for INDUS is positive. This means there is a positive relationship between higher INDUS values and higher prices, more specifically, the more non-retail buisnesses around will result in higher prices for the house. However, the magnitude of the weight is relatively small, meaning there does not exist a strong positive relationship. This positive sign may not be something I would expect, simply because through my understanding of what people

want, I would expect houses closer to more retail businesses would likely be more expensive because of the location being so convenient to buy products fairly easily. However, the fact that the weight is relatively small is reassuring because I also understand that houses that are less surrounded by retail businesses are also favourable and usually the property will have more space to build a bigger house. Of course, I am no real estate expert, so my understanding of the housing market could be totally wrong and I acknowledge that.

g) The two additional metrics I have decided to use is $R^2$-score and MAE, or mean absolute error. $R^2$ is a statistical measure of how close the data is to the fitted regression line, and so will give us a good idea of how well our model is fitting the test data points. On the other hand, MAE measures the average magnitude of the errors in a set of predictions, without taking into account the direction of the error. This will give us an understanding of how accurate our model is in predicting target values. This is different from MSE in that it will give us a raw idea of the accuracy of our model, without punishing outlier or high residual predictions.

The following are the metric results of the linear regression model implemented:

$$MSE : 10.706$$
$$MAE : 3.405$$
$$R^2 : 0.725$$

h) I would say the three most important features are RM, LSTAT, and NOX. This is because they have amongst the largest weight magnitude out of all the features, meaning they have either a very strong positive or negative relationship towards the target as we change their respective values. The three features also intuitively makes sense. The number of rooms should obviously have a very strong positive relationship with the price of the house. Moreover, the price of a house is also decided by the community, and so the status of the people living their should have a direct correlation to the community, and hence the price of the house. Lastly, it also seems quite intuitive that places that have larger nitric oxides concentration will have lower housing prices. I note that DIS has a slightly higher weight magnitude than NOX, but it makes less intuitive sense the strong negative relationship that it has. Hence, I went with the NOX feature instead.

# Question 3

a) Let $\mathbf{z} = \mathbf{y} - \mathbf{X}\mathbf{w}$. Then consider the loss function:

$$\begin{aligned}\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N} z_i^2 a^{(i)} &= \frac{1}{2}\sum_{i=1}^{N} z_i \cdot a^{(i)} z_i \\ &= \frac{1}{2}\mathbf{z}^T\mathbf{A}\mathbf{z} \\ &= \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T\mathbf{A}(\mathbf{y} - \mathbf{X}\mathbf{w}) \\ &= \frac{1}{2}\mathbf{y}^T\mathbf{A}\mathbf{y} + \frac{1}{2}\mathbf{w}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{A}\mathbf{y}\end{aligned}$$

Then taking the gradient with respect to $\mathbf{w}$ and setting it to 0, we arrive at:

$$\mathcal{L}'(\mathbf{w}) = \mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{A}\mathbf{y} = 0$$

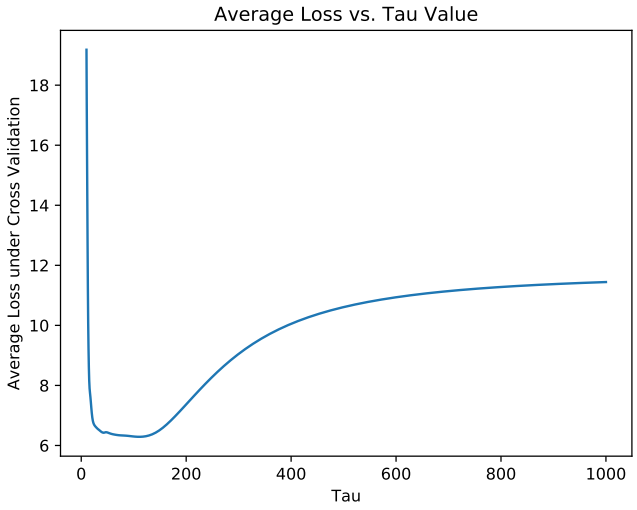
Then we get that $\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{A}\mathbf{y}$. We note that $\mathbf{X}^T\mathbf{A}\mathbf{X}$ is positive definite since $A$ has only positive entries in its diagonal. Hence the determinant is positive, and so invertible. Thus, we arrive at $\mathbf{w} = (\mathbf{X}^T\mathbf{A}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}\mathbf{y}$. To see that this is indeed a minimum, we take the second derivative:

$$\mathcal{L}''(\mathbf{w}) = \mathbf{X}^T\mathbf{A}\mathbf{X}$$

Since $\mathbf{X}^T\mathbf{A}\mathbf{X}$ is a positive definite matrix, all of its eigenvalues are positive, and so we get that $\mathbf{w}$ produces a minimum, as required. Thus

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{A}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}\mathbf{y}$$

c)



d) As $\tau$ approaches zero, it is clear to see that the average loss recorded under the cross-validation technique approaches infinity. This implies that with really small values of $\tau$, we get a very inaccurate model. On the other hand, as $\tau$ approaches $\infty$, we get that the average loss is gradually increasing, much like a logarithmic function. This also implies that the loss will approach infinity, and so suggesting that extremely large values of $\tau$ also lead to an inaccurate locally weighted linear regression model.

e) One of the disadvantages of locally weighted linear regression model is the fact that we require to keep in memory the entire training set, when making prediction on the test set. Also, it is much more computationally expensive, as we essentially build a regression model for each test point, and also the parameter increase linearly with the training set.
Some of the advantages of this model is that it performed much better on the Boston data set, achieving the lowest MSE of around 4, whereas, just the linear regression had a MSE of around 10. Furthermore, with locally weighted linear regression, an extra measure is taken to prevent over fitting of the data, as the test point is also taken into consideration when predicting target value. Moreover, it does not require our model to fit to a single function for all of the test points, thus making it more flexible to model more complex data with greater accuracy.