Assignment 1 - MDL 07.03.2024

Q 1

.1) $f(Y, \hat{Y}) = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2$

$$\frac{\partial f(Y,\hat{Y})}{\partial a} = -2 \sum (y_i - a - bx_i) = 0$$

$$\Leftrightarrow \quad na = \sum y_i - b \sum x_i$$

$$\Leftrightarrow \quad a = \bar{y} - b\bar{x}, \quad \text{where } \bar{y} = \frac{1}{n}\sum y_i \text{ and } \bar{x} = \frac{1}{n}\sum x_i$$

.2) Yes, there is only a unique global minimizer of $f(Y, \hat{Y})$ as it is a polynomial of second degree with positive coefficents. Hence, by convex format it holds that any local minimum is a global minimum and that there only exists one such point.

No, as then the coefficients become $\frac{1}{n}$ instead of 1, which is still positive and, hence, has no impact on the location of the global minimizer.

Q2.2) The true function is curved (sin(x)) while the plotted line is linear.

Q2.3) The coefficients are virtually the same and also the residuals are very similar, which shows that for the polynomial of degree 1 we have essentially a linear ordinary least squares estimator.

Q3) As we are training our model with the training set, we would expect the loss on our train set to be less than on the test set. In practice we observe that the test loss is about twice as large as the train loss, which makes sense, as we use the coefficients that minimize the loss of the train set to estimate the test data.

Q4) By increasing n we obtain higher residuals and mse, as this implies that we sum over more quadratic terms, each one eventually adding to the sum of squared residuals. Similarly, by increasing epsilon we introduce more variation in our ys, eventually creating higher residuals, which also implies higher mse for both the train and the test set. Hence, the loss function increases both in n and in ε.

Q5.2) "Polyfit may be poorly conditioned":
This might indicate that the degree of the polynomial ist too high, as a small change in the input data would lead to significant changes in its coefficients, i.e., the numerical stability is not guaranteed. This comes due to the fact that the number of coefficients is very close to the number of observations.

Q 6

1) Judging by the MSE loss function plotted against the Degree of Polynomials, we can observe that from around the polynomials of degree 5 the MSE of the train set starts to diverge from the test set one, suggesting that from this point onwards the model is overfitted. This conclusion is fortified by the fact that the loss function of the test set increases in degree of polynomials from this point onwards

2) Observing the plots in Q 5.3, the polynomial of degree 5 is the best compromise between in sample prediction and out of sample performance. This is also confirmed by the result in Q 5.2, suggesting that the polynomial of degree 5 has the best overall performance

3) By increasing the number of observation to 200, the overfitting starts on a later point of the axis containing the degrees of the polynomials. By the result in 5.2 this point would now be around 12, as from this point onwards the loss function of the test set starts to increase as can be seen in 5.3.

4) Yes, as this yields more options of well conditioned models, for the point at which the number of degrees is relatively close to the number of observations is at a larger number, essentially creating numerical stability.