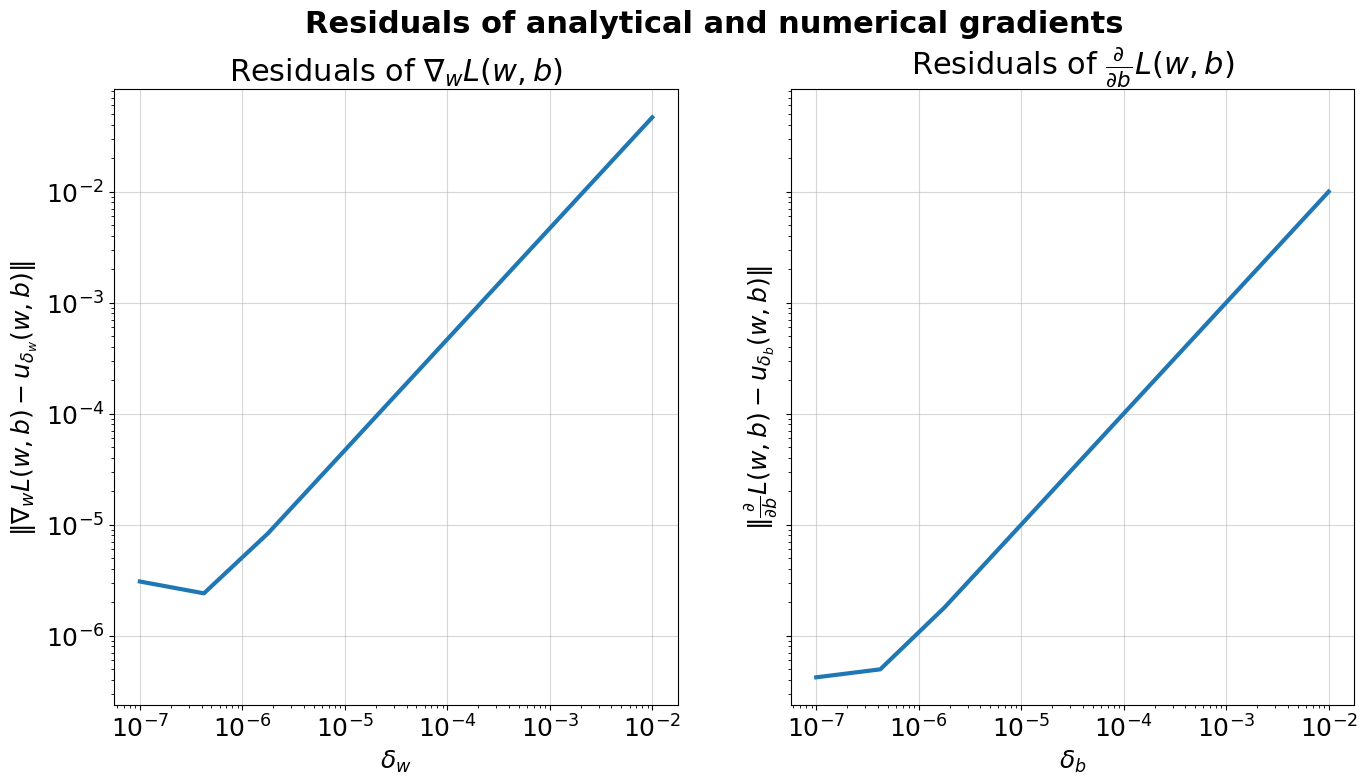
Major HW 3 – Regression

Orad Barel, 311288203, [oradbarel@campus.technion.ac.il](mailto:oradbarel@campus.technion.ac.il)

Ofir Manor, 316084623, [ofir.manor@campus.technion.ac.il](mailto:ofir.manor@campus.technion.ac.il)

**Q1.**

**Q2.**

****

Diagram

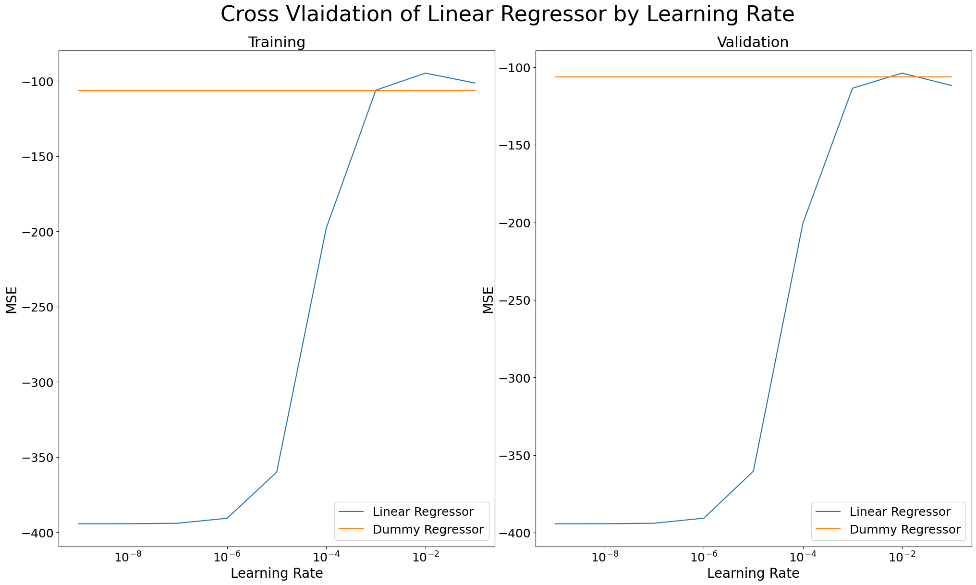
Description automatically generated**Q3.**

We see here that from the most part (excluding the higher the learning rate, the faster the descent of the loss is. This makes sense because small learning rates create a slow gradient descent, so that we arrive at the minimum of the loss only after a great amount of iterations. As for we can see that the loss is divergent after a few iterations, the large “jumps” that occur do not allow the gradient descent to arrive at a minimal loss.

Our best learning rate (the one that achieved the minimal validation loss) is , and it does not make sense to increase the number of gradient steps as it seems to achieve this minimum before step 1500 and slightly diverge around the minimum.

**Q4.**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Section | Train MSE | Valid MSE |
|  |  | Cross Validated | |
| Dummy | 2 | -105.82 | -106.19 |

**Q5.**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Section | Train MSE | Valid MSE |
|  |  | Cross Validated | |
| Dummy | 2 | -105.82 | -106.19 |
| Linear | 2 | -94.77 | -103.8 |

**Q6.**

Depends on the model. The dummy model, that always uses the average contamination level, would not change. The average stays the same regardless of any normalization we did to the features.

The Linear Regressor will change. This model uses SGD in order to find the minimal loss. SGD is sensitive to the values of the features. The partial derivative of a feature which holds “larger” (as in far greater or far lesser) data points than other features will pull the SGD stronger in it’s direction, thus making it more prominent, even if it’s data is less indicative of out target.