FMAN45 - Machine Learning - Assignment 1

William Lindskog

March/April 2021

Introduction

For this assignment, 7 tasks will be solved.

Task 1

Starting with the coordinate-wise problem

$$\min_{\omega_i} \frac{1}{2}||\mathbf{r}_i - \mathbf{x}_i\omega_i||_2^2 + \lambda|\omega_i| \tag{1}$$

it is possible to test the first case when $\lambda = 0$. The objective function is then

$$\min_{\omega_i} \frac{1}{2}||\mathbf{r}_i - \mathbf{x}_i\omega_i||_2^2 \tag{2}$$

Differentiation gives with Fermat's rule that

$$0 = -x_i^T(r_i - x_i\hat{\omega}_i) \tag{3}$$

$$\iff \hat{\omega}_i = \frac{x_i^T r_i}{x_i^T x_i} = \frac{x_i^T r_i}{||x_i||_2^2} \tag{4}$$

when $x \neq 0$. If $\lambda > 0$ and $\omega_i \neq 0$ then

$$0 \in -x_i^T(r_i - x_i\omega_i) + \lambda\frac{\omega_i}{|\omega_i|} \tag{5}$$

Optimal $\omega_i = \hat{\omega}_i$ is found to be

$$0 = -x_i^T r_i + x_i^T x_i\hat{\omega}_i + \lambda\frac{\hat{\omega}_i}{|\hat{\omega}_i|} \iff \tag{6}$$

$$x_i^T r_i = x_i^T x_i\hat{\omega}_i + \lambda\frac{\hat{\omega}_i}{|\hat{\omega}_i|} = \hat{\omega}_i\big(x_i^T x_i + \lambda\frac{1}{|\hat{\omega}_i|}\big) \tag{7}$$

Letting both sides be considered by their absolute value equals

$$|x_i^T r_i| = |\hat{\omega}_i\big(x_i^T x_i + \lambda\frac{1}{|\hat{\omega}_i|}\big)| \tag{8}$$

$$= |\hat{\omega}_i||\big(x_i^T x_i + \lambda\frac{1}{|\hat{\omega}_i|}\big)| \tag{9}$$

$$\iff |\hat{\omega}_i| = \frac{1}{x_i^T x_i}(|x_i^T r_i| - \lambda) \tag{10}$$

$\text{sgn}(\omega_i)$ can be derived from equation (7) since the parentheses contain a positive value.

$$\text{sgn}(\omega_i) = \text{sgn}(x_i^T r_i) = \frac{x_i^T r_i}{|x_i^T r_i|} \tag{11}$$

This eventually gives that

$$\omega_i = \text{sgn}(\omega_i)\omega_i = \frac{x_i^T r_i}{|x_i^T r_i|}\frac{1}{x_i^T x_i}(|x_i^T r_i| - \lambda) = \frac{x_i^T r_i}{x_i^T x_i |x_i^T r_i|}(|x_i^T r_i| - \lambda) \tag{12}$$

Which is what we wanted to show.

Task 2

Using that $X^T X = I_N$, it is presented that

$$\hat{\omega}_2^{(2)} - \hat{\omega}_2^{(1)} = 0, \ \forall i \tag{13}$$

It is also presented in information about assignment that one can rewrite $x_i^T r_i^{(j-1)}$ as

$$x_i^T r_i^{(j-1)} = x_i^T(t - \sum_{l<i} x_l\hat{\omega}_l^{(j)} - \sum_{l>i} x_l\hat{\omega}_l^{(j-1)}) = x_i^T t \tag{14}$$

This is due to the fact that $x_i^T x_l = 0, \forall l \neq i$. Moreover, $x_i^T x_i = 1$. We may for $|x_i^T r_i^{(j-1)} > \lambda|$ find that

$$\hat{\omega}_i^{(j)} = \frac{x_i^T r_i^{(j-1)}}{x_i^T x_i |x_i^T r_i^{(j-1)}|}\Big(|x_i^T r_i^{(j-1)}| - \lambda\Big) \tag{15}$$

$$= \frac{x_i^T r_i^{(j-1)}}{|x_i^T r_i^{(j-1)}|}\Big(|x_i^T r_i^{(j-1)}| - \lambda\Big) \tag{16}$$

$$= x_i^T r_i^{(j-1)} - \lambda\text{sgn}\big(x_i^T r_i^{(j-1)}\big) \tag{17}$$

$$= x_i^T t - \lambda\text{sgn}\big(x_i^T t\big) \tag{18}$$

$\hat{\omega}_i^{(j)}$ does therefore not depend on previous estimates

$$\hat{\omega}_2^{(2)} - \hat{\omega}_2^{(1)} = x_i^T t - \lambda\text{sgn}\big(x_i^T t\big) - (x_i^T t - \lambda\text{sgn}\big(x_i^T t\big) = 0 \tag{19}$$

Also, for $|x_i^T r_i^{(j-1)} < \lambda|$, $\hat{\omega}_i^{(j)} = 0$ and this concludes task 2.

Task 3

For this task, we wish to show that

$$\lim_{\sigma \to 0} E\Big(\hat{\omega}_i^{(1)} - \omega_i^\star\Big) = \begin{cases} -\lambda, & \omega_i^\star > \lambda \\ -\omega_i^\star, & |\omega_i^\star| \leq \lambda \\ \lambda, & \omega_i^\star < -\lambda \end{cases} \tag{20}$$

Using the hint presented, investigating the 2 cases, we have

Case (1) $x_i^T r_i^{(j-1)} > \lambda$

From equation (18) the limit can be found as

$$\lim_{\sigma \to 0} x_i^T t = x_i^T X \omega^\star = \omega_i^\star \tag{21}$$

Expected value of the limit can now be calculated as

$$\lim_{\sigma \to 0} E\Big(\hat{\omega}_i^{(1)} - \omega_i^\star\Big) = \lim_{\sigma \to 0} E\Big(x_i^T t - \lambda \text{sgn}(x_i^T t) - \omega_i^\star\Big) \tag{22}$$

$$= \lim_{\sigma \to 0} E\Big(\omega_i^\star - \lambda \text{sgn}(\omega_i^\star) - \omega_i^\star\Big) \tag{23}$$

$$= -\lambda \text{sgn}(\omega_i^\star) = -\lambda \tag{24}$$

The last step is possible since $x_i^T r_i^{(j-1)} \to \omega_i^\star$ when $\sigma \to 0$.

Case (2) $x_i^T r_i^{(j-1)} < -\lambda$

Utilizing similar derivation but with changed sign for $\omega_i^\star$

$$\lim_{\sigma \to 0} E\Big(\hat{\omega}_i^{(1)} - \omega_i^\star\Big) = \lim_{\sigma \to 0} E\Big(x_i^T t - \lambda \text{sgn}(x_i^T t) - \omega_i^\star\Big) \tag{25}$$

$$= \lim_{\sigma \to 0} E\Big(\omega_i^\star - \lambda \text{sgn}(\omega_i^\star) - \omega_i^\star\Big) \tag{26}$$

$$= -\lambda \text{sgn}(\omega_i^\star) = \lambda \tag{27}$$

However, we must look at a third case

Case (3) $|x_i^T r_i^{(j-1)}| \leq \lambda$

Using that $\hat{\omega}_i^{(j)} = 0$ for $|x_i^T r_i^{(j-1)}| \leq \lambda$ presented in second line of equation (3) in instructions. Therefore,

$$\lim_{\sigma \to 0} E\Big(\hat{\omega}_i^{(1)} - \omega_i^\star\Big) = \lim_{\sigma \to 0} E\Big(0 - \omega_i^\star\Big) = -\omega_i^\star \tag{28}$$

for $|\omega_i^\star| \leq \lambda$. The cases in equation (21) are therefore fullfilled. One can identify that as λ increases, the bias of LASSO estimates increases. One can therefore argue that there is an evident compensation of increased bias in form of reduced variance. This, one must consider when creating the LASSO model.

Task 4

The graphs presented for this task include 50 original data points together with 50 synthesised points. It is worth mentioning that I have been using the updated dataset. The first graph shows interpolated values reconstructed using $\lambda = 0.1$.

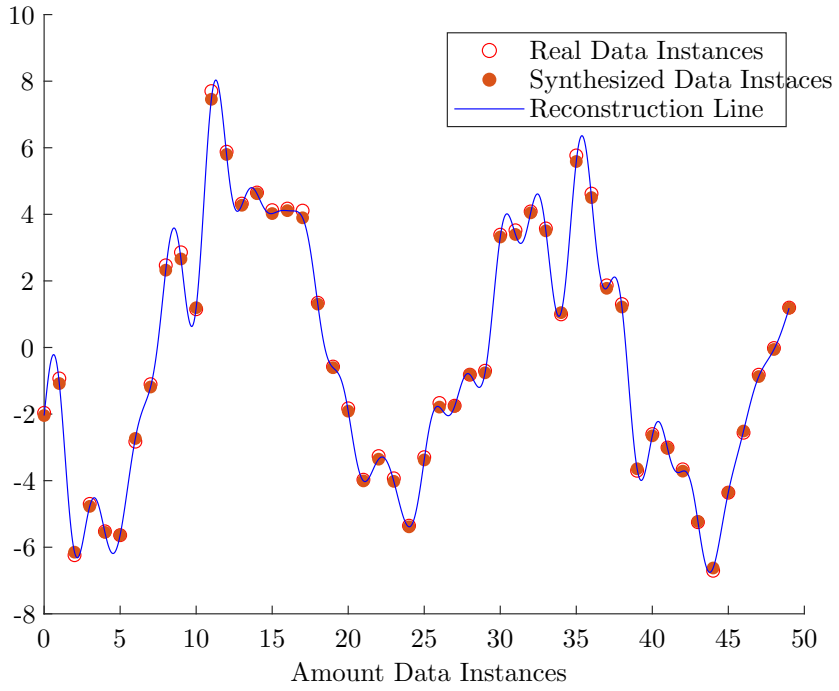

Figure 1: Interpolated reconstruction of the data using $\lambda = 0.1$

Blue line in previous figure shows obvious sign of overfitting. It does not generalize, but seeks to capture all data points. This consequently means that this interpolated reconstruction of the data would not perform well on new values as it relies heavily on the values of current data. For $\lambda = 10$

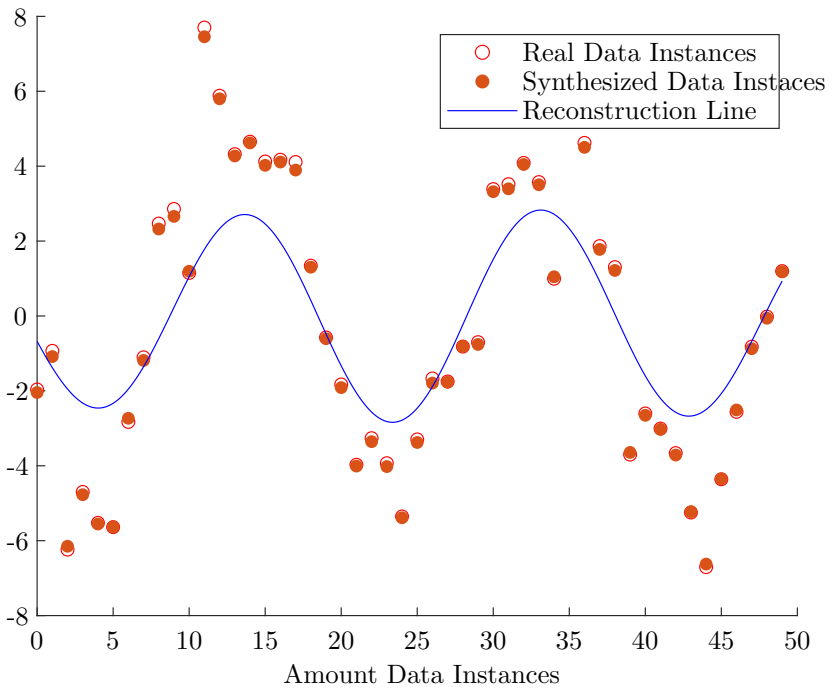

Figure 2: Interpolated reconstruction of the data using $\lambda = 0.1$

we can clearly identify signs of underfitting. It is to generalized and does not capture the peaks well. It spots that there seems to be some sort of periodicity in the data set but may be too biased in its core. Using another λ set to 2 we receive

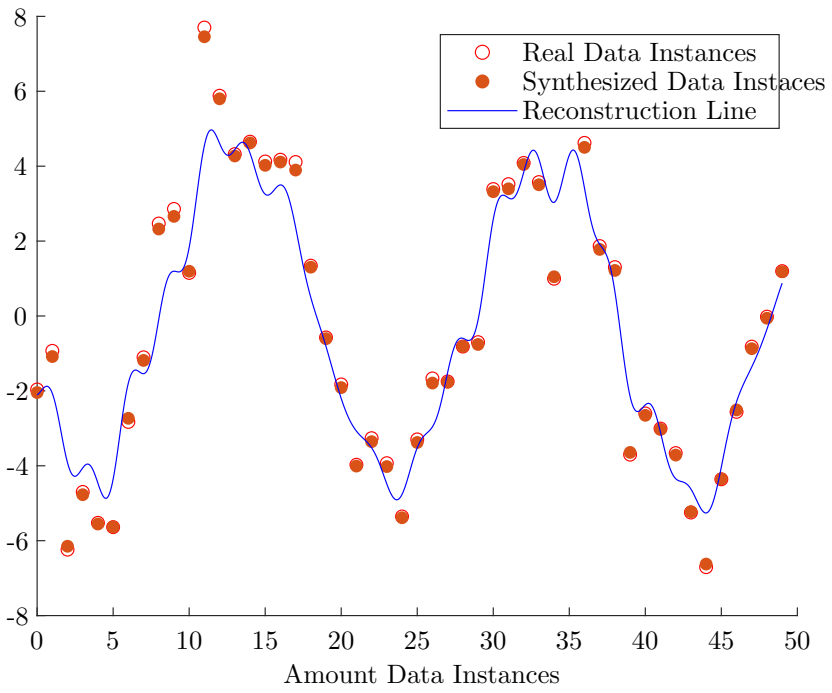

Figure 3: Interpolated reconstruction of the data using $\lambda = 2$

This interpolated reconstruction captures better the peaks and the periodicity. It can be argued that the variance is somewhat high but better than for figure 1. The errors are not that great.

As for part two of this task, it is found that number of non-zero coordinates are

- $\lambda = 0.1$, 223 non-zero coordinates
- $\lambda = 10$, 7 non-zero coordinates
- $\lambda = 2$, 15 non-zero coordinates

Concluding what this means is that for greater values of λ we need a greater number of coordinates than the four we are using now.

Task 5

Following figures are visualizations of root mean squared errors (RMSEs) for estimation and validation data sets. Moreover, using 10 folds for cross-validation and LASSO solver used in Task 4, lambdas are set from 0.1 to $\max(|X^T t|) \approx$

23.259, and evenly distributed in this interval on a log scale. Moreover, for computation set of lambdas are 200.

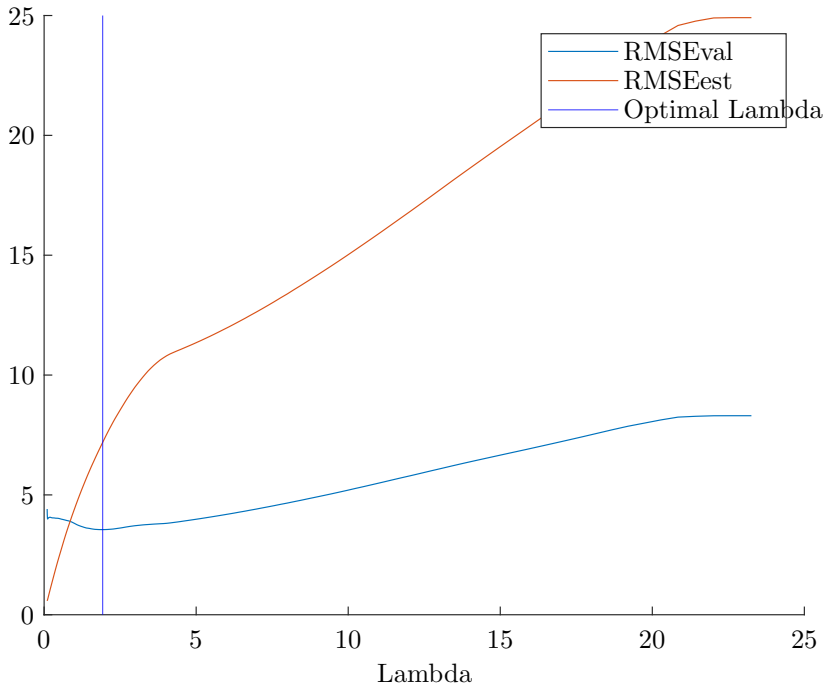

Figure 4: $RMSE_{Validation}$ and $RMSE_{Estimation}$ for various lambdas.

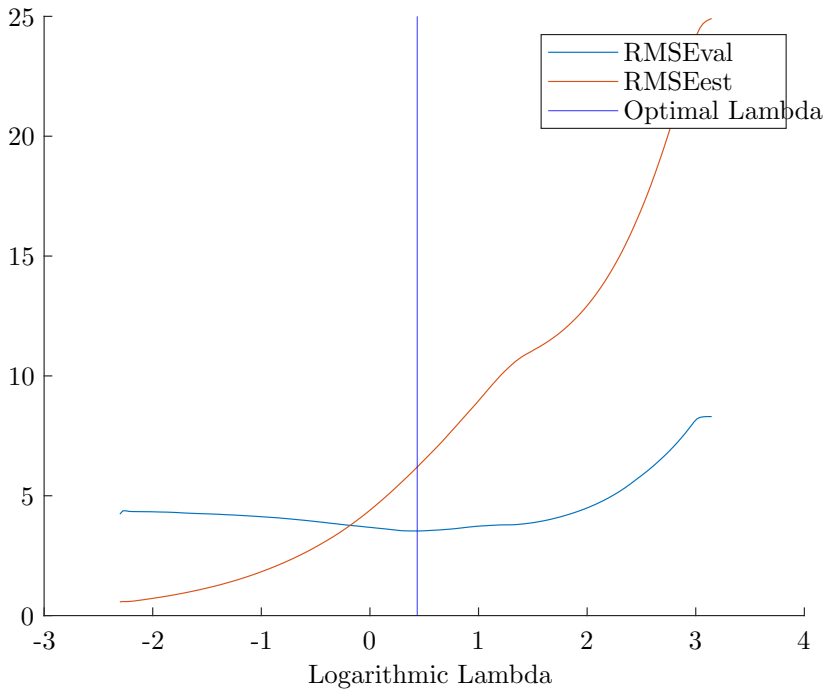

Figure 5: $RMSE_{Validation}$ and $RMSE_{Estimation}$ for various lambdas (log scale).

Evidently, the RMSEs are quite low for low lambda values which is in line with what we saw in task 4 when it overfitted. Towards the end of the graph these values increases as the model is uable to explain the underlying data, therefore underfitting. Optimal lambda is found to be ≈ 1.55. Using this lambda and reconstructing the data, following figure is received

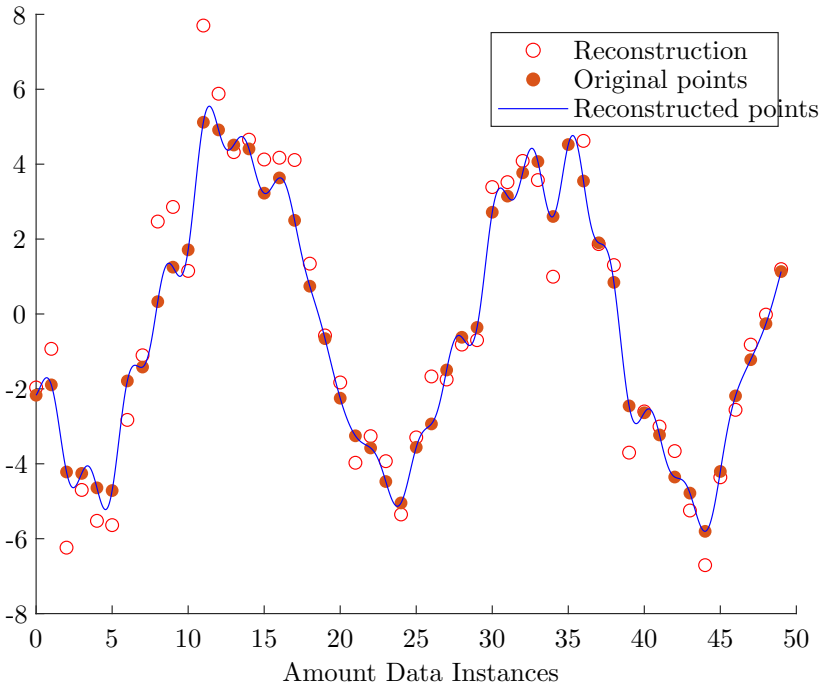

Figure 6: Reconstruction of data using optimal lambda.

Task 6

Utilizing K-fold cross-validation, we can analyze an excerpt of audio data. It is possible to approach the problem as in Task 5. Retrieving a set of 100 values λ, we also set the lowest value to 0.001 and greatest to $\max |X^T t_i|, \forall i$, where i is the number of frames. For the frames one can select the greatest value.

Having applied a 3-fold LASSO regression to the frames for λ values, it is possible to calculate the squared error and the sum for all frames. Thereafter, the $RMSE_{val}$ and $RMSE_{est}$ were calculated for all λ values. The optimal λ was found where $\min_\lambda \ RMSE_{val}(\lambda)$. Optimal lambda in this case is approx. 0.0045.

Figure below shows $RMSE_{val}$ and $RMSE_{est}$ for different λ and figure (8) includes the log-scale.

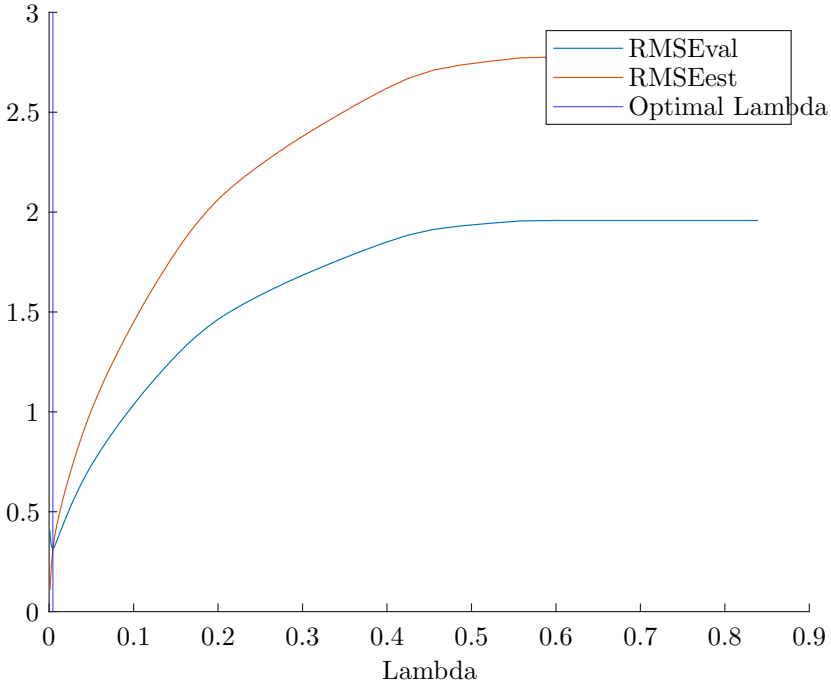

Figure 7: RMSE value for estimation and validation using different lambda

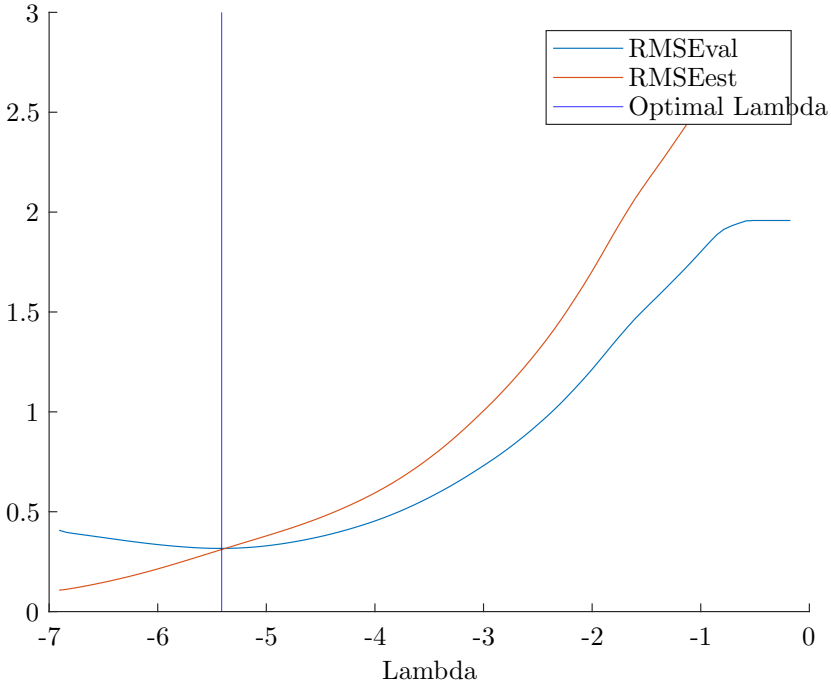

Figure 8: RMSE value for estimation and validation using different lambda (log scale)

Task 7

Having retrieved an optimal λ from task 6, we can create new denoised audio data. Having listened to it, it is possible to compare with the original test data. It is possible to hear a reduction of background noise.

Further testing, includes greater values such as $\lambda = 0.02$. Background noise is reduced even more. However, the sound coming from the piano "goes away". For somewhat lower values, e.g. 0.0001, one hears more of the background sounds and little less from the piano.