Assignment 1 in Machine Learning FMAN45

Eliot Montesino Petrén, el6183mo-s, 990618-9130, LU Faculty of Engineering F19 , eliot.mp99@gmail.com

I. Exercises

Exercise 1

To verify the first line of Equation (3) by solving Equation (2) for $w_i \neq 0$, we take the derivative of the objective function with respect to w_i, and set it to zero.

The objective function in Equation (2) is:

$$\text{minimize}_{w_i} \quad \frac{1}{2}\|r_i - x_i w_i\|_2^2 + \lambda|w_i|$$

We compute the derivative as follows, noting that for $w_i \neq 0$, the derivative of $|w_i|$ is $\frac{w_i}{|w_i|}$:

$$\frac{d}{dw_i}\left(\frac{1}{2}\|r_i - x_i w_i\|_2^2 + \lambda|w_i|\right) = -x_i^T(r_i - x_i w_i) + \lambda\frac{w_i}{|w_i|}$$

Setting the sum of these derivatives equal to zero gives:

$$-x_i^T(r_i - x_i\hat{w}_i) + \lambda\frac{\hat{w}_i}{|\hat{w}_i|} = 0$$

Solving for $x_i^T r_i$:

$$x_i^\top r_i = x_i^\top x_i\hat{w}_i + \lambda\frac{\hat{w}_i}{|\hat{w}_i|}$$

$$x_i^\top r_i = \hat{w}_i(x_i^\top x_i + \lambda\frac{1}{|\hat{w}_i|}) \tag{1}$$

Taking the absolute value of the expression yields:

$$|x_i^\top r_i| = |\hat{w}_i||x_i^\top x_i + \lambda\frac{1}{|\hat{w}_i|}|$$

Both terms $x_i^\top x_i$ and λ are positive, yielding:

$$x_i^\top x_i + \lambda\frac{1}{|\hat{w}_i|} > 0 \tag{2}$$

Solving for $\hat{w}_i$ in the previous equation where the absolute value was introduced gives us:

$$|\hat{w}_i| = \frac{1}{x_i^T x_i}\left(|x_i^T r_i| - \lambda\right)$$

Since the sign of $\hat{w}_i$ aligns with $x_i^\top r_i$ from eq 1 considering the condition in eq 2, we can express $\hat{w}_i$ as:

$$\hat{w}_i = \text{sign}(x_i^\top r_i)\frac{1}{x_i^T x_i}\left(|x_i^T r_i| - \lambda\right)$$

$$\hat{w}_i = \frac{x_i^T r_i}{x_i^T x_i|x_i^T x_i|}\left(|x_i^T r_i| - \lambda\right)$$

This is the final result of this exercise given the conditions from the instructions, marking the end of this exercise.

Exercise 2

Given the orthogonal regression matrix condition, $X^\top X = I$, the coordinate descent solver's update for the i-th weight, $\hat{w}_i$, in equation (3) can be simplified as follows:

Starting with the update equation:

$$\hat{w}_i^{(j)} = \frac{x_i^T r_i^{(j-1)}}{x_i^\top x_i}\left(|x_i^T r_i^{(j-1)}| - \lambda\right) \tag{3}$$

Given $X^\top X = I$, we have $x_i^\top x_i = 1$, simplifying the expression to:

$$\hat{w}_i^{(j)} = x_i^T r_i^{(j-1)} - \lambda\text{sign}(x_i^T r_i^{(j-1)}) \tag{4}$$

Next, using the definition of the residual $r_i^{(j-1)}$ from equation (4):

$$r_i^{(j-1)} = t - \sum_{k<i} x_k\hat{w}_k^{(j)} - \sum_{k>i} x_k\hat{w}_k^{(j-1)} \tag{5}$$

Since $X^\top X = I$, the dot product of x_i with any x_k (where $k \neq i$) is zero. Therefore, we have:

$$x_i^\top r_i^{(j-1)} = x_i^\top\left(t - \sum_{\ell<i} x_\ell\hat{w}_\ell^{(j)} - \sum_{\ell>i} x_\ell\hat{w}_\ell^{(j-1)}\right) \tag{6}$$

$$= \begin{cases} x_i^\top x_\ell = 0 & \text{when } \ell \neq i, \\ x_i^\top x_i = 1 \end{cases} \tag{7}$$

$$= x_i^\top t \tag{8}$$

By substituting $x_i^\top r_i^{(j-1)} = x_i^\top t$ back into equation 4, we obtain:

$$\hat{w}_i^{(j)} = x_i^\top t - \lambda\text{sign}(x_i^\top t) \tag{9}$$

This shows that the coordinate descent update for $\hat{w}_i$ is invariant to previous estimates and only depends on t, x_i, and λ. Therefore, the coordinate descent solver will converge after at most one full pass over the weights in $\mathbf{w}$, i.e. $\hat{w}_i^{(2)} - \hat{w}_i^{(1)} = 0$ for every i.

$$\hat{w}_i^{(2)} - \hat{w}_i^{(1)} = \left(x_i^\top t - \lambda\cdot\text{sign}(x_i^\top t)\right) - \left(x_i^\top t - \lambda\cdot\text{sign}(x_i^\top t)\right) = 0$$

Exercise 3

From the previous exercise, we have the result for the LASSO estimate:

$$\hat{w}_i = x_i^\top t - \lambda \operatorname{sgn}(x_i^\top t) \tag{10}$$

Assuming the data t is generated as:

$$t = Xw^* + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I) \tag{11}$$

We need to show the bias for the LASSO estimate as $\lambda \to 0$. Retaining the expectation operator, we consider:

$$E(\hat{w}_i - w_i^*) = E(x_i^\top t - \lambda \operatorname{sgn}(x_i^\top t) - w_i^*) \tag{12}$$

Given that $t = Xw^* + \epsilon$:

$$E(\hat{w}_i - w_i^*) = E(x_i^\top (Xw^* + \epsilon) - \lambda \operatorname{sgn}(x_i^\top (Xw^* + \epsilon)) - w_i^*) \tag{13}$$

This simplifies to:

$$E(\hat{w}_i - w_i^*) = E(x_i^\top Xw^* + x_i^\top \epsilon - \lambda \operatorname{sgn}(x_i^\top Xw^* + x_i^\top \epsilon) - w_i^*) \tag{14}$$

Since $X^\top X = I$ and $x_i^\top X = e_i^\top$ where e_i is the i-th standard basis vector:

$$E(\hat{w}_i - w_i^*) = E(w_i^* + x_i^\top \epsilon - \lambda \operatorname{sgn}(w_i^* + x_i^\top \epsilon) - w_i^*) \tag{15}$$

Thus,

$$E(\hat{w}_i - w_i^*) = E(x_i^\top \epsilon - \lambda \operatorname{sgn}(w_i^* + x_i^\top \epsilon)) \tag{16}$$

Next, we consider three cases for w_i^*:
1. For $w_i^* > \lambda$:

$$E(\hat{w}_i - w_i^*) = E(x_i^\top \epsilon - \lambda) \tag{17}$$

Since $E(x_i^\top \epsilon) = 0$,

$$E(\hat{w}_i - w_i^*) = -\lambda \tag{18}$$

2. For $w_i^* < -\lambda$:

$$E(\hat{w}_i - w_i^*) = E(x_i^\top \epsilon + \lambda) \tag{19}$$

Since $E(x_i^\top \epsilon) = 0$,

$$E(\hat{w}_i - w_i^*) = \lambda \tag{20}$$

3. For $|w_i^*| \leq \lambda$:

$$E(\hat{w}_i - w_i^*) = E(0 - w_i^*) \tag{21}$$

Thus,

$$E(\hat{w}_i - w_i^*) = -w_i^* \tag{22}$$

These conditions (13), (14), and (15) derive directly from the expected value analysis of $\hat{w}_i$ under the LASSO penalty as $\lambda \to 0$.

The Least Absolute Shrinkage and Selection Operator (LASSO) promotes sparsity in the model by penalizing the absolute values of the coefficients. This leads to fewer non-zero coefficients, which is useful in preventing overfitting, especially in high-dimensional settings where many predictors may not be relevant. By shrinking coefficients, LASSO reduces model variance while maintaining interpretability and mitigating the risk of including noise as predictors. This balance between bias and variance makes LASSO an attractive choice for models where high variance is a concern, but the risk of underfitting due to high bias is managed by selecting an appropriate λ.

A. Exercise 4

In this exercise, a coordinate descent solver for the LASSO solution was implemented by filling the provided skeleton and it was applied to reconstruct a signal using different regularization strengths (λ). The original signal is a linear combination of two sinusoids with different frequencies, contaminated by noise.

Reconstruction plots were produced for three different λ values: 0.1, 10, and 1.0. Figure 1 shows these reconstructions. Let me comment them:

1) $\lambda = 0.1$): With a small λ, the model shows clear signs of overfitting. The reconstructed signal (thin blue line) closely follows the predicted data points (blue crosses), indicating that the model is fitting the noise rather than capturing the underlying signal structure. The result is a jagged reconstruction rather than a smooth sinusoidal of the true noise-free signal. The noisy signal provided is seen in red circles.

2) $\lambda = 10$): With a large λ, we observe significant underfitting. The reconstructed signal is straight up a smooth and simple sinusoidal that fails to capture the variability in the data and the linear combination of two sinusoids. The thin blue line clearly deviates substantially from the noisy data points (red circles) and is almost as if it was not based in reality.

3) $\lambda = 1$): This λ value provides a more balanced reconstruction, at the least closer to the sought after "sweet spot"'. The reconstructed signal (thin blue line) is a better trade-off this time. In my opinion, it maintains a reasonable distance from the noisy data points, filtering them out, and it captures the overall nature of a linear combination of two sinusoids, thus the choice of $\lambda = 1$, after testing around with some values for λ.

I calculated the number of non-zero weights (coordinates) in the weight vector for each λ value:

TABLE I
NUMBER OF NON-ZERO WEIGHTS FOR DIFFERENT λ VALUES.

λ	Number of non-zero weights
0.1	227
10	7
1	65

These results demonstrate a relationship between λ and the number of non-zero weights. A smaller λ (0.1) results in many non-zero coefficients (227), leading to overfitting. A larger λ (10) aggressively shrinks coefficients to zero, leaving only 7 non-zero weights and causing underfitting. The intermediate

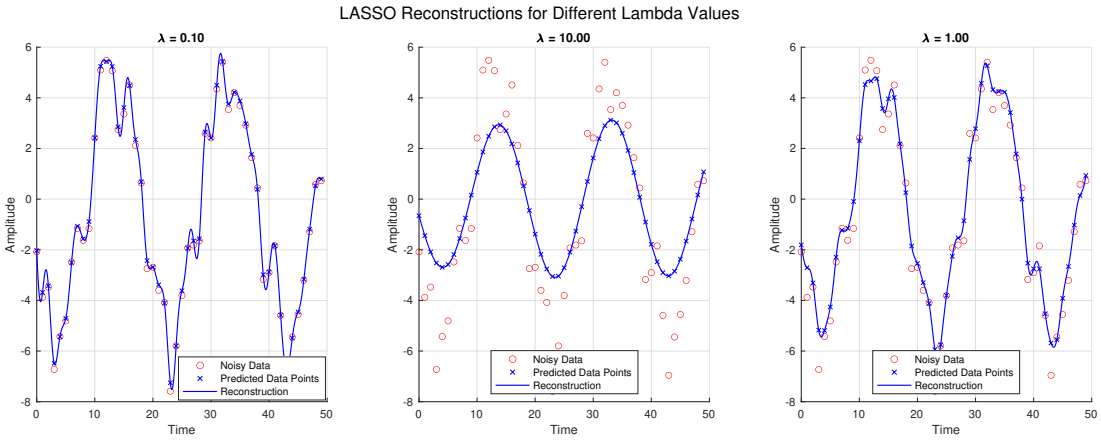

Fig. 1. Reconstruction plots for different λ values. The different regularization strengths (λ) showcase overfitting, underfitting and a more

λ (1) strikes a balance, retaining 65 non-zero coefficients to capture the signal.

Even the balanced reconstruction ($\lambda = 1$) uses more non-zero weights (65) than the true signal, which consists of only 4 non-zero weights. This suggests that some degree of model complexity is needed to account for noise and imperfections in the data, even when the underlying signal is sparse.

B. Exercise 5

The purpose of this exercise is to find the optimal value for the hyperparameter λ by minimizing RMSE. A K-fold cross-validation scheme with K = 10 for the number of folds and a grid of 100 λ values was created and they were logarithmically spaced between 0.01 and $\max(|X^T t|)$. The data and regression matrix provided in A1_data was used. The results are illustrated in figures 2 and 3.

The reconstruction seems fairly effective observing it by the naked eye, my previous guess was $\lambda = 1$ but apparently this value should be slightly more effective. Minimizing the $RMSE_{val}$ (on the validation dataset) indicates striking a balance between underfitting and overfitting when applying the trained model on real data.

C. Exercise 6

To find the optimal λ across all frames of the audio excerpt, I implemented a multi-frame 3-fold cross-validation scheme in the provided skeleton using code from earlier. The RMSE for each λ was calculated as the mean RMSE across all frames. The optimal λ was then determined by identifying the value that corresponded to the smallest validation RMSE.

For this task, I used a range of λ values: $\lambda \in [10^{-5}, \max(|X^T t_i|)\ \forall i]$, where i is the frame index.

The plot in Figure 4 demonstrates the value of regularization to produce a clear reconstruction of the audio. In small λ values the regularization is weak and the validation RMSE is high, suggesting a noisy and overfitted reconstruction.

In contrast, as λ increases, we initially see a decrease in RMSE, indicating improved model performance. However, beyond a certain point, the RMSE begins to increase again. This is no surprise, as we are familiar to the concept of too much regularization leads to underfitting, where the model fails to capture important features of the audio signal.

The optimal λ value was found to be approximately 0.0043. This value represents the best trade-off between model complexity and generalization performance.

These results showcase once again the importance of careful hyperparameter tuning in LASSO regression. Let us move on to how the final result sounds.

D. Exercise 7

Using the optimal $\hat{\lambda} = 0.0043$ from Exercise 6, I denoised the test data Ttest and saved the results as instructed.

Listening to the original and denoised audio revealed an improvement in sound quality. The denoised version had a noticeably less background noise while preserving the clarity of the piano music. The noise is more present in between notes, which was an interesting result, I think. This result is because silence means that there is only noise left and there is no other underlying process that can be identified, so that the noise can be filtered out.

I quickly tried a slightly higher and lower λ value, but both resulted in somewhat poorer audio quality (in defense of the hypothesized global minimum).

The cross-validation approach in Exercise 6 successfully identified an optimal, or at least a sensible, λ value for noise reduction.

Extra: See Figure 5. I quickly made a plot of the noisy and the denoised signal. There is no noticeable difference by the naked eye of the two samples, but interestingly enough we can hear the difference!

II. References

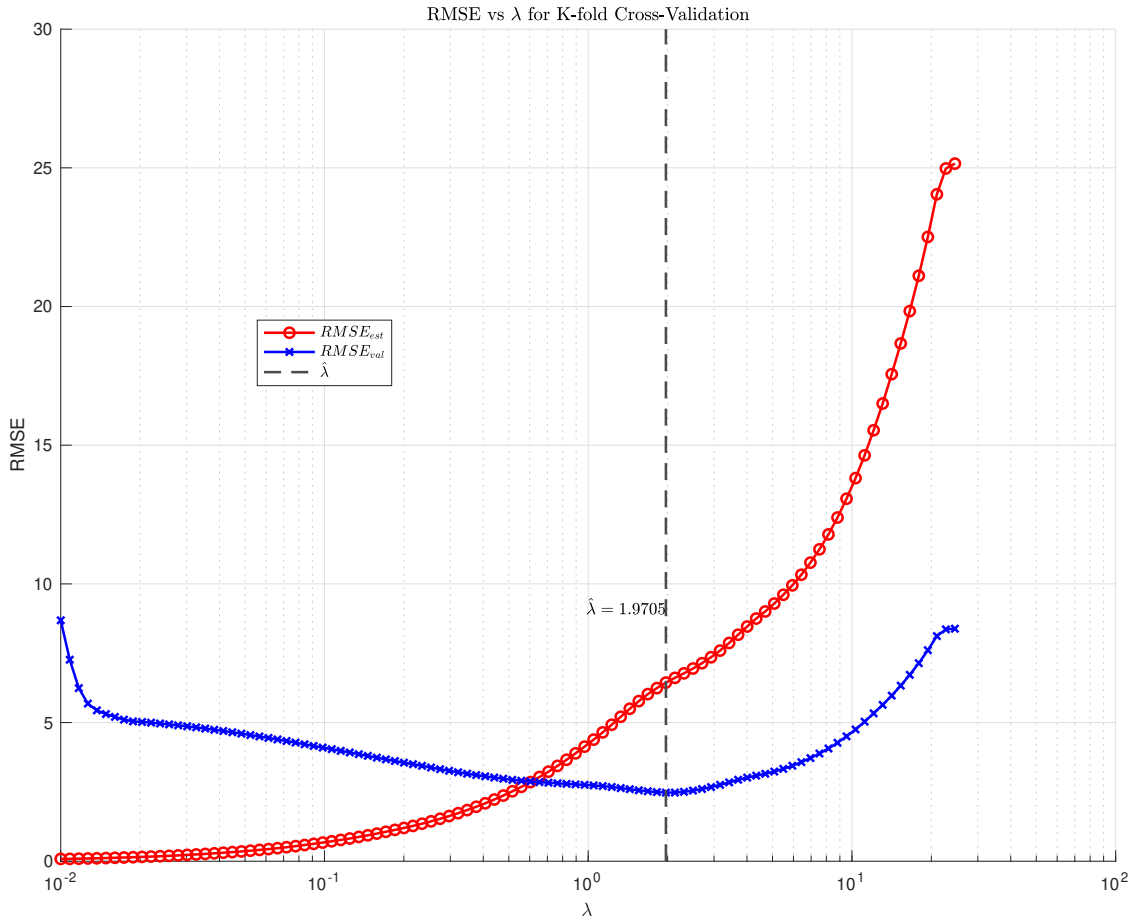

Fig. 2. RMSE for validation (blue crosses) and estimation (red circles) data across different λ values, marked line with minimizing and optimal lambda value.

III. Appendix

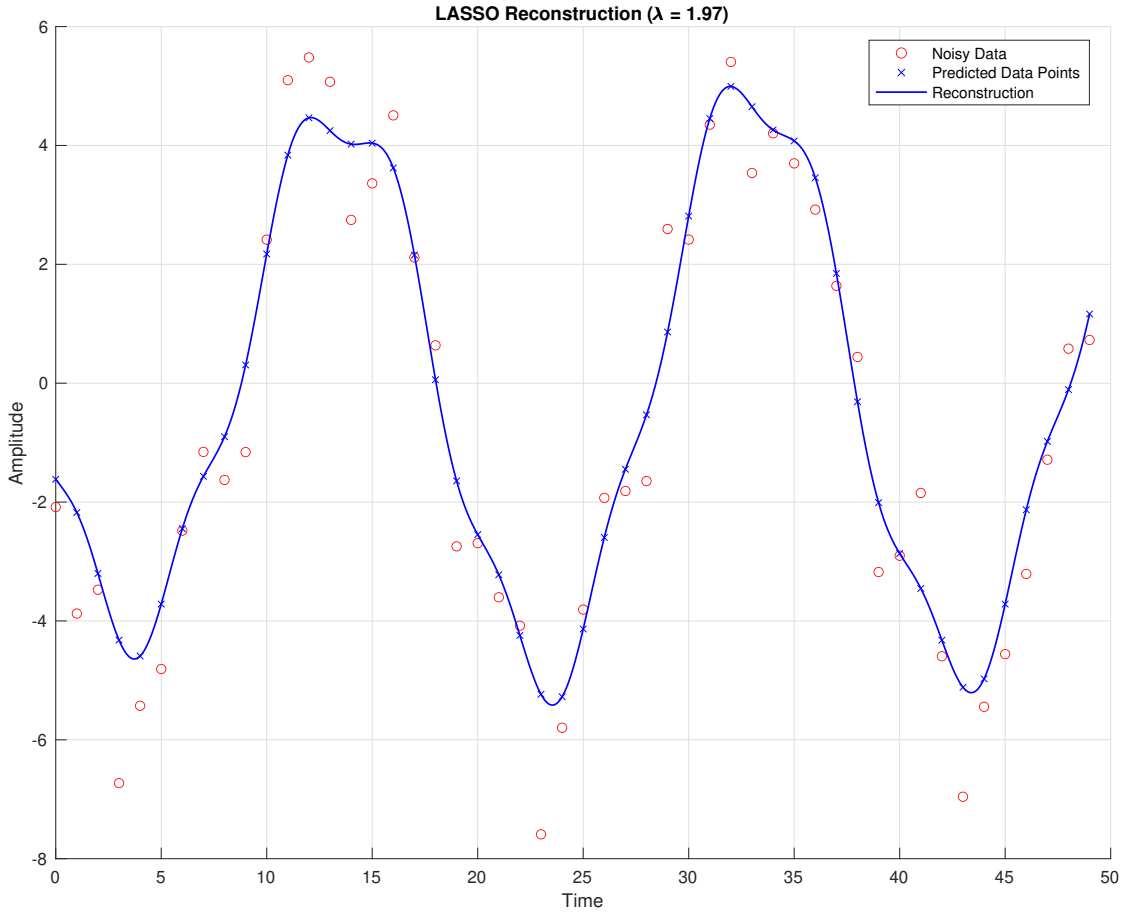

Fig. 3. Reconstruction plot using optimal $\lambda = 1.97$.

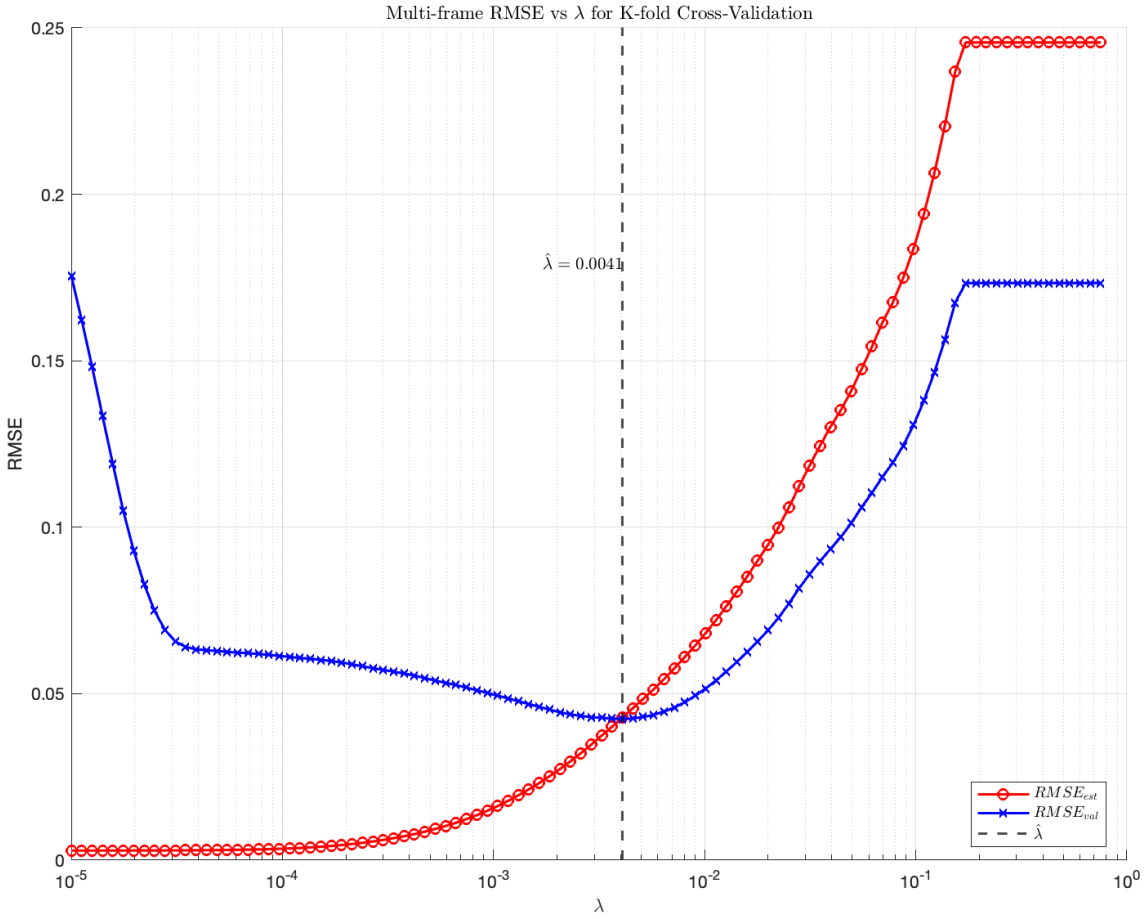

Fig. 4. Validation (blue crosses) and estimation (red circles) RMSE for different λ values on the audio data.

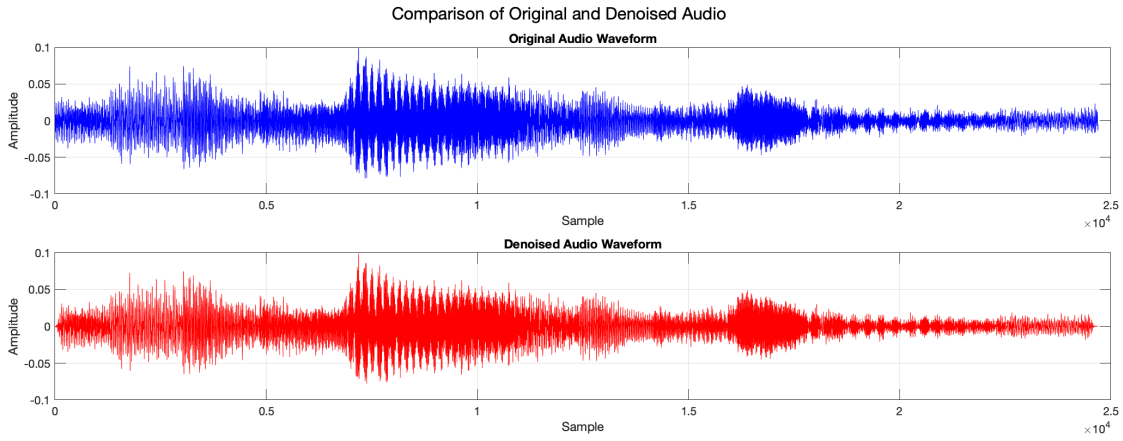

Fig. 5. **(Extra!)** A simple plot comparing the noisy and the denoised waveforms with amplitude by time. I cannot see any difference between them, but it is possible to hear the difference.