FMAN45 Machine Learning
-Assignment 1.

Nils Romanus

April 2022

1 Penalized regression via the LASSO

Task 1.

To verify the first line in the closed form solution of the coordinate descent minimizer one may solve expression (1) with respect to w_i under the condition of $|\mathbf{x}_i^T \mathbf{r}_j^{j-1}| > \lambda$.

$$\underset{w_i}{\text{minimize}} \frac{1}{2}||\mathbf{t} - \mathbf{X}\mathbf{w}||_2^2 + \lambda||w_i|| \quad , w_i \neq 0 \tag{1}$$

As is customary in optimization problems we differentiate expression (1) with respect to w_i and equal it to 0, yielding expression (2).

$$-\lambda \mathbf{x}_i^T(\mathbf{r}_i - \mathbf{x}_i w_i) + \lambda \frac{w_i}{|w_i|} = 0 \quad , w_i \neq 0 \tag{2}$$

$$w_i = \frac{\mathbf{x}_i^T \mathbf{r}_i - \lambda sign(w_i)}{\mathbf{x}_i^T \mathbf{x}_i} \quad , w_i \neq 0 \tag{3}$$

Using the condition of $|\mathbf{x}_i^T \mathbf{r}_j^{j-1}| > \lambda$ it is known that the sign of w_i is purely determined by the sign of $\mathbf{x}_i^T \mathbf{r}_i = \frac{\mathbf{x}_i^T \mathbf{r}_i}{|\mathbf{x}_i^T \mathbf{r}_i|}$. Using that one can substitute $sign(w_i)$ in expression (3) and arrive at expression (4).

$$w_i = \frac{\mathbf{x}_i^T \mathbf{r}_i - \lambda \frac{\mathbf{x}_i^T \mathbf{r}_i}{|\mathbf{x}_i^T \mathbf{r}_i|}}{\mathbf{x}_i^T \mathbf{x}_i} \quad , w_i \neq 0 \tag{4}$$

Which can be rewritten as the closed form solution of the coordinate descent minimizer in expression (5), completing the proof.

$$w_i = \frac{\mathbf{x}_i^T \mathbf{r}_i}{\mathbf{x}_i^T \mathbf{x}_i |\mathbf{x}_i^T \mathbf{r}_i|}(|\mathbf{x}_i^T \mathbf{r}_i| - \lambda) \quad , w_i \neq 0 \tag{5}$$

Task 2

By substituting the residual vector $\mathbf{r}$ with its definition $\mathbf{r}_i = \mathbf{t} - \sum_{l \neq i} \mathbf{x}_l w_l$ in the expression that defines the closed form solution of the coordinate descent minimizer, one arrives at expression (6).

$$w_i^{(j)} = \frac{\mathbf{x}_i^T(\mathbf{t} - \sum_{l \neq i} \mathbf{x}_l w_l^{(j-1)})}{\mathbf{x}_i^T \mathbf{x}_i |\mathbf{x}_i^T(\mathbf{t} - \sum_{l \neq i} \mathbf{x}_l w_l^{(j-1)}))|}(|\mathbf{x}_i^T(\mathbf{t} - \sum_{l \neq i} \mathbf{x}_l w_l^{(j-1)})| - \lambda) \tag{6}$$

Given that the regression matrix $\mathbf{X}$ is orthogonal, the $\mathbf{x}_i$s are pairwise orthogonal. This fact makes *all* the terms in the sums evaluate to zero due to the fact that $\mathbf{x}_i^T \mathbf{x}_l = 0$ when $i \neq l$. Also remembering that $\mathbf{x}_i$ has unit norm we arrive at the simplified expression (7).

$$w_i^{(j)} = \frac{\mathbf{x}_i^T \mathbf{t}}{|\mathbf{x}_i^T \mathbf{t}|}(|\mathbf{x}_i^T \mathbf{t}| - \lambda) \quad \forall_{i,j} \tag{7}$$

Expression (7) is independent of j and hence $\hat{w}_i^{(2)} - \hat{w}_i^{(1)} = (7) - (7) = 0$.

Task 3.

Given that the regression matrix $\mathbf{X}$ is orthogonal one may formulate the update equation for the coordinate descent minimizer as expression (8) by using the simplified expression(7).

$$w_i^{(j)} = \begin{cases} \frac{\mathbf{x}_i^T\mathbf{t}}{|\mathbf{x}_i^T\mathbf{t}|}(|\mathbf{x}_i^T\mathbf{t}| - \lambda) \quad ,|\mathbf{x}_i^T\mathbf{r}_i^{(j-1)}| > \lambda \\ 0 \quad ,|\mathbf{x}_i^T\mathbf{r}_i^{(j-1)}| \leq \lambda \end{cases} \tag{8}$$

Given our data $\mathbf{t} = \mathbf{Xw^*} + \mathbf{e}$, where $\mathbf{e}$ is zero-mean N dimensional Gaussian Noise, we arrive at the update equation (9) for the first iteration.

$$w_i^{(1)} = \begin{cases} \frac{\mathbf{x}_i^T(\mathbf{x}_i\mathbf{w^*}_i+\mathbf{e})}{|\mathbf{x}_i^T(\mathbf{x}_i\mathbf{w^*}_i+\mathbf{e})|}(|\mathbf{x}_i^T(\mathbf{x}_i\mathbf{w^*}_i + \mathbf{e})| - \lambda) \quad ,|\mathbf{x}_i^T\mathbf{r}_i^{(j-1)}| > \lambda \\ 0 \quad ,|\mathbf{x}_i^T\mathbf{r}_i^{(j-1)}| \leq \lambda \end{cases} \tag{9}$$

By evaluating the estimate bias in the limit, i.e $\lim\limits_{\sigma\to 0} E[\hat{w}^{(1)} - w*_i]$, using expression (9), one arrives arrive at equation (10). The Gaussian noise is deterministic in the limit $\sigma \to 0$ and since its expectation is zero the expression gets simplified. Once again we utilize the fact that $\mathbf{x}_i$ has unit norm.

$$\lim_{\sigma\to 0} E[\hat{w}^{(1)} - w_i^*] = \frac{w_i^*}{|w_i^*|}(|w_i^*| - \lambda) - w_i^* = -sign(w_i^*)(\lambda) \quad ,|\mathbf{x}_i^T\mathbf{r}_i^{(j-1)}| > \lambda \tag{10}$$

In the edge case where $|\mathbf{x}_i^T\mathbf{r}_i^{(j-1)}| \leq \lambda$ we use the bottom half of (9) and arrive at expression (11) for the limit.

$$\lim_{\sigma\to 0} E[\hat{w}^{(1)} - w_i^*] = \lim_{\sigma\to 0} E[0 - w_i^*] = -w_i^*, \quad |\mathbf{x}_i^T\mathbf{r}_i^{(j-1)}| \leq \lambda \tag{11}$$

By combining (10) and (11) we arrive at the full expression 12.

$$\lim_{\sigma\to 0} E[\hat{w}^{(1)} - w_i^*] = \begin{cases} -\lambda, \quad w_i^* > \lambda \\ -w_i^*, \quad w_i^* \leq \lambda \\ \lambda, \quad w_i^* < \lambda \end{cases} \tag{12}$$

Expression (10) and (12) show some of the intuition behind the acronym *LASSO*: least absolute shrinkage and selection operator. The bottom half of (10) shows that a parameter close to zero, where λ decides what is considered "close", will be set to zero. I.e we are (S)electing which parameters to exclude/include. If a parameter is further than λ away from zero we (S)hrink it one λ step towards zero. This is more clearly shown in expression (12) that describes the limit of the estimate bias. If the real parameter is greater than λ, we will underestimate it by an amount λ, and vice versa if it is smaller than $-\lambda$. If it smaller than λ we will set the estimated parameter to zero, causing us to be wrong by an amont equal to the actual parameter value.

2 Hyperparameter-learning via K-fold cross-validation

Task 4

a)

After having implemented the cyclical coordinate solver one may apply LASSO regression to the noisy data **t** and produce reconstruction plots. The visualizations in figures 1, 2 and 3 show these reconstruction plots for different values of the regularization parameter λ using $N = 50$ reconstructed data points. Figure 1 corresponds to $\lambda = 0.1$, Figure 2 to $\lambda = 10$ and Figure 3 to $\lambda = \lambda_{user} = 1$

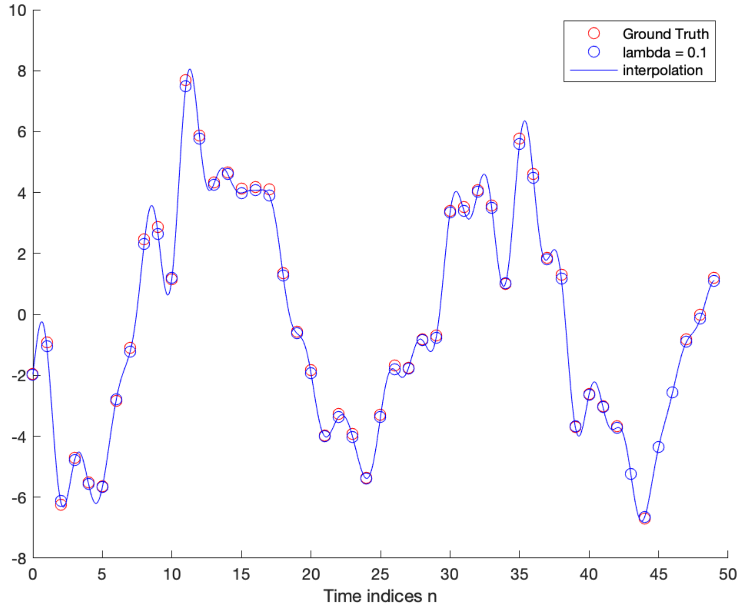

Figure 1: Original data points (ground truth), reconstructed data points and interpolation of reconstruction against time, using $\lambda = 0.1$. Model is clearly over fitted

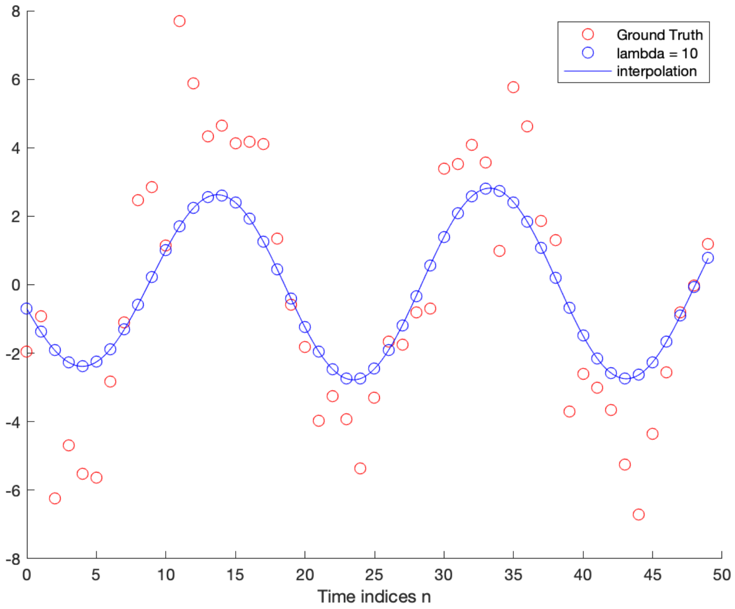

Figure 2: Original data points (ground truth), reconstructed data points and interpolation of reconstruction against time, using $\lambda = 10$. Model is clearly under fitted

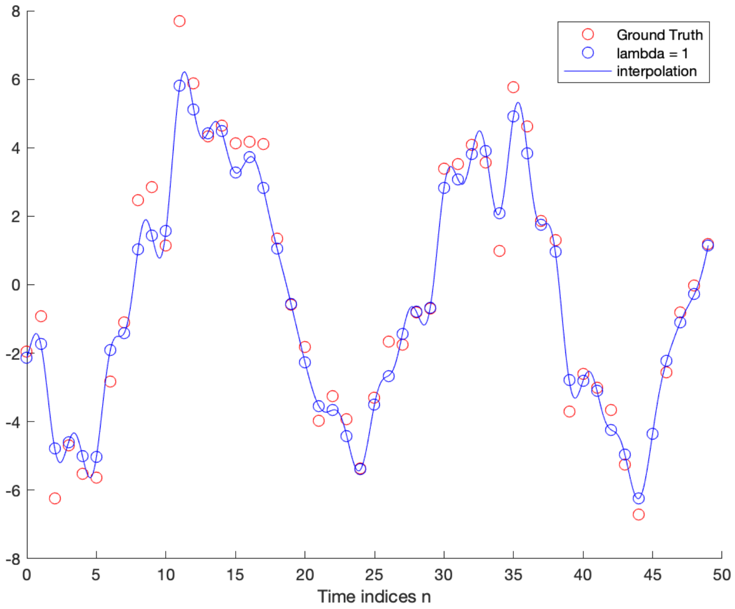

Figure 3: Original data points (ground truth), reconstructed data points and interpolation of reconstruction against time, using $\lambda = \lambda_{user} = 1$. Model is descently fitted

The figures show the intuition behind the regularization hyper parameter λ as the reconstruction plots with a lower value of λ value are more closely fitted to the data points, and vice versa. With a λ of 10 the reconstruction points fall right on the ground truth points and the model will hence not generalize well, the model is over fitted. With a λ of 0.1 the model becomes under fitted. The user guess of $\lambda = 1$ provides a descent fit.

b)

In the second sub-task we investigate the amount of non-zero coordinates in the parameter vector. These are listed in table 1 where k denotes the number of non-zero coordinates.

Non zero elements	
λ	k
10	6
1	96
0.1	234

Table 1: Number of non-zero coordinates for different λs

From looking at table 1 one may note that the number of non-zero elements decreases as λ increases. Table 1 gives further intuition behind the name 'regularization' parameter, the larger the λ the more regularization we impose. The regularization will zero out parameters deemed less important, larger λ therefore leads to less non-zero elements. Zeroing out less important parameters is in some cases a good thing, since we prevent overfitting, however in some instances we might zero out a more important parameter leading to an under fit. Since the actual signal is a sum of 2 complex valued sinusoids the actual amount of non zero elements should be 4. This is much smaller than the amount of non zero elements in the decently fitted case using $\lambda = 1$. Despite theoretically needing only 4 non-zero coordinate to perfectly reconstruct the signal we are unable to do it using many more. This fact stems from their being noise in the data which impacts the estimation of our parameters.

Task 5

To choose the optimal value for the regularization parameter one may implement a K fold cross validation grid search scheme. The cross validation is done using $K = 50$ random folds with a search space consisting of an evenly spaced grid with $N_\lambda = 50$ entries between $log(1)$ and $log(10)$. The search can be vizualised by plotting the objective function against the value of the regularization parameter during estimation and validation. Figure 4 shows the value of the objective function during estimation $RMSE_{est}$, and validation $RMSE_{val}$ for different values of λ.

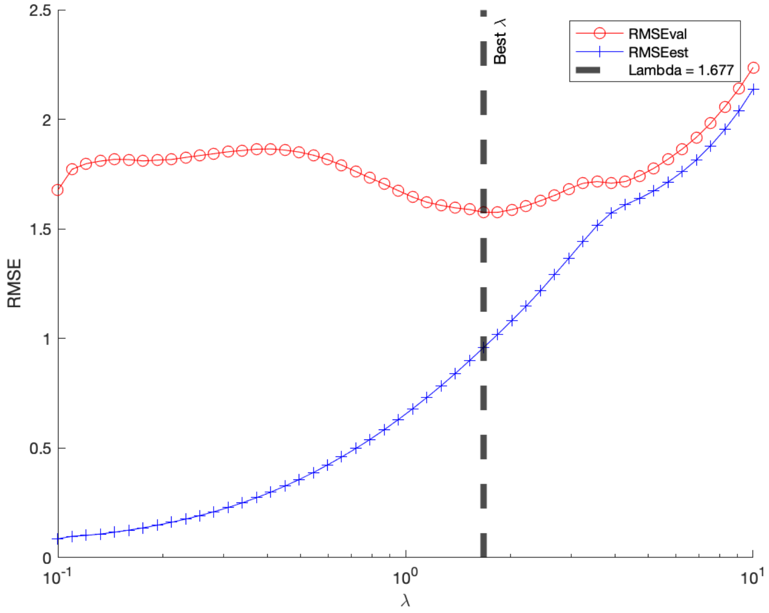

Figure 4: Test- and validation RMSE for different values of λ. The horizontal line indicates the best value of λ, i.e the value that minimizes validation error.

Unsurprisingly one can note that the estimation error displayed in figure 4 increases with increasing λ. This is because a larger amount of regularization will force the model to be fitted less closely to the training data during estimation of parameters and hence our $RMSE_{est}$ will increase the more we regularize. When assesing performance we are more intersted in the error during validation since we will make inference on new data. The validation error reaches a minimum for $\lambda_{opt} = 1.68$. Using the optimal choice of regularization parameter we can create a better reconstruction plot as shown in figure 5.

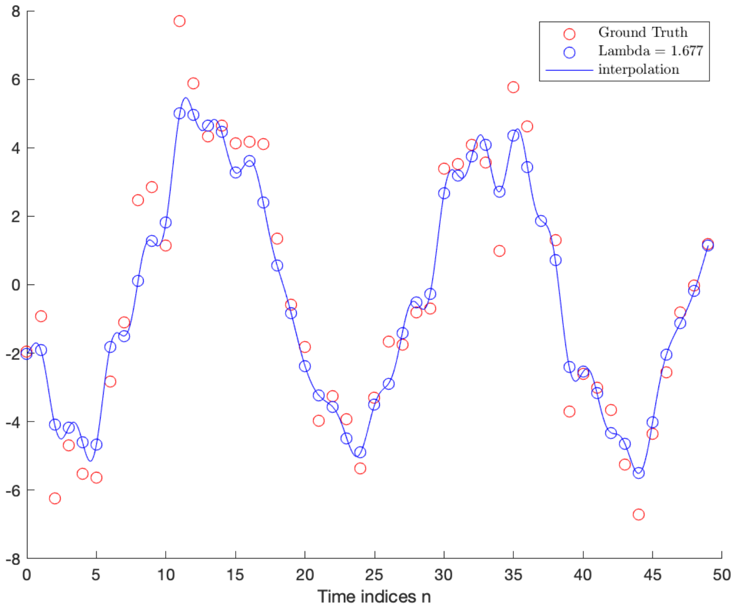

Figure 5: Original data points (ground truth), reconstructed data points and interpolation of reconstruction against time, using $\lambda = \lambda_{opt} = 1.68$.

3 Denoising of an audio excerpt

Task 6.

The K-fold cross validation grid search scheme can be applied frame by frame to find the optimal λ per frame. Using this approach we can tune the the model for the task of denoising an audio excerpt. The results from this scheme is visualized, using the same methodology as task 5 in, figure 6. The optimal λ for denoising the audio excerpt is $\hat{\lambda} = 0.005$.

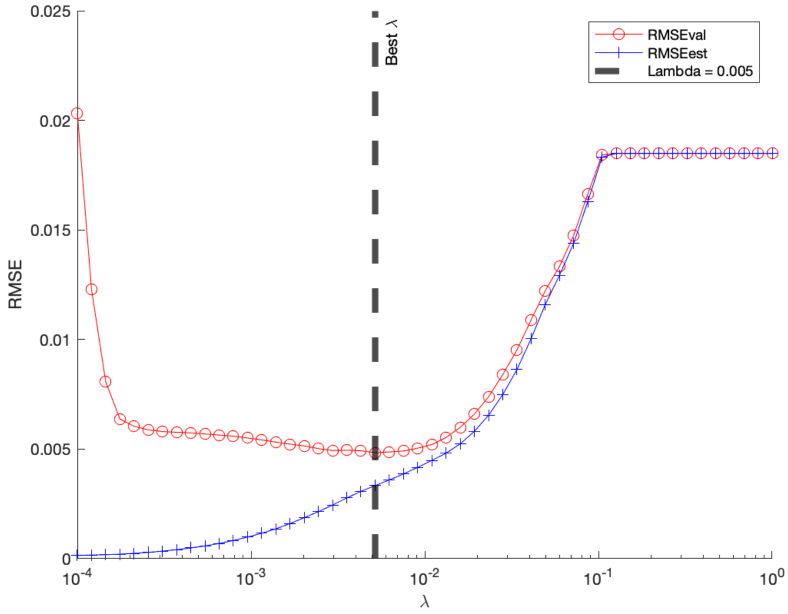

Figure 6: Test- and validation RMSE for different values of λ with a horizontal line indicating the optimal λ

Task 7.

A listener notices the difference in noise when comparing the original noisy sample to the output of the optimized denoising model. The output sounds better, however there is still quite a bit of noise. Testing with $\lambda = 0.1$ actually produces an ouptut that 'sounds' subjectively less noisy. A larger λ produces 'better' sounding output since the increased regularization cleans out more of the noise. However this larger λ also clears out more of the signal. By playing around with the regularization parameter the user can get an auditory intuition for the bias/variance trade-off. An output using a larger λ will become more under-fitted and hence less noisy(less variance) since the parameters of the model will be less impacted by the noise in the training set, however if significantly under fitted the output looses all of the signal's information(it becomes biased). For example, using $\lambda = 0.5$ the output becomes 2 constant tones. Using a λ that is to small, on the other hand, keeps to much of the noise (variance) of the original signal since the parameters of the model has been fitted to capture the noise in the training set.