Rasmus Helander

# Assignment 1



## Task 1

Task is to minimize

$$L(\mathbf{x},\mathbf{w}) = \frac{1}{2}\left\|\mathbf{r}_i - \mathbf{x}_i w_i\right\|_2^2 + \lambda\left|w_i\right| \tag{1}$$

with respect to the parameter $w_i$. First we simplify:

$$\begin{aligned} L = \frac{1}{2}(\mathbf{r}_i - \mathbf{x}_i w_i)^T(\mathbf{r}_i - \mathbf{x}_i w_i) + \lambda\left|w_i\right| = \frac{1}{2}(\mathbf{r}_i^T - \mathbf{x}_i^T w_i)(\mathbf{r}_i - \mathbf{x}_i w_i) + \lambda\left|w_i\right| = \\ \frac{1}{2}\left(\mathbf{r}_i^T\mathbf{r}_i - w_i\mathbf{x}_i^T\mathbf{r}_i - w_i\mathbf{r}_i^T\mathbf{x}_i + w_i^2\mathbf{x}_i^T\mathbf{x}_i\right) + \lambda\left|w_i\right| = \frac{1}{2}\left(\mathbf{r}_i^T\mathbf{r}_i - 2w_i\mathbf{x}_i^T\mathbf{r}_i + w_i^2\mathbf{x}_i^T\mathbf{x}_i\right) + \lambda\left|w_i\right| \end{aligned} \tag{2}$$

Note that the last step uses the fact that $\mathbf{x}_i^T\mathbf{r}_i$ is a scalar, and thus $\mathbf{x}_i^T\mathbf{r}_i = (\mathbf{x}_i^T\mathbf{r}_i)^T = \mathbf{r}_i^T\mathbf{x}_i$.

To minimize $L$ with respect to $w_i$ we must first differentiate:

$$\frac{\partial L}{\partial w_i} = \frac{1}{2}(2w_i\mathbf{x}_i^T\mathbf{x}_i - 2\mathbf{x}_i^T\mathbf{r}_i) + \lambda\frac{w_i}{|w_i|} \tag{3}$$

where we have used that the derivative of $|w_i|$ is the sign of $w_i$ for all $w_i \neq 0$.

We can then find eventual optima by solving

$$\frac{\partial L}{\partial w_i} = 0 \Leftrightarrow w_i = \frac{1}{\mathbf{x}_i^T\mathbf{x}_i}\left(\mathbf{x}_i^T\mathbf{r}_i - \lambda\frac{w_i}{|w_i|}\right) \tag{4}$$

Equation 4 can also be written as follows:

$$w_i\left(\mathbf{x}_i^T\mathbf{x}_i + \frac{\lambda}{|w_i|}\right) = \mathbf{x}_i^T\mathbf{r}_i \tag{5}$$

Since $\mathbf{x}_i^T\mathbf{x}_i = \|\mathbf{x}_i\|_2^2 \geq 0$ and $\frac{\lambda}{|w_i|} \geq 0$, the sign of $w_i$ must be the same as the sign of $\mathbf{x}_i^T\mathbf{r}_i$, or

$$\frac{w_i}{|w_i|} = \frac{\mathbf{x}_i^T\mathbf{r}_i}{|\mathbf{x}_i^T\mathbf{r}_i|} \tag{6}$$

which finally gives the updating rule

$$w_i = \frac{1}{\mathbf{x}_i^T\mathbf{x}_i}\left(\mathbf{x}_i^T\mathbf{r}_i - \lambda\frac{\mathbf{x}_i^T\mathbf{r}_i}{|\mathbf{x}_i^T\mathbf{r}_i|}\right) = \frac{\mathbf{x}_i^T\mathbf{r}_i}{\mathbf{x}_i^T\mathbf{x}_i\,|\mathbf{x}_i^T\mathbf{r}_i|}\left(\left|\mathbf{x}_i^T\mathbf{r}_i\right| - \lambda\right) \ \square \tag{7}$$

# Task 2

When the regression matrix is an orthonormal basis we have

$$\mathbf{x}_j^T \mathbf{x}_i = \delta_{ij} \tag{8}$$

where $\delta_{ij}$ is the Kronecker delta.

Thus, we have

$$\mathbf{x}_i^T \mathbf{r}_i^{(j-1)} = \mathbf{x}_i^T \left( \mathbf{t} - \sum_{l<i} \mathbf{x}_l \hat{w}_l^{(j)} - \sum_{l>i} \mathbf{x}_l \hat{w}_l^{(j-1)} \right) = \mathbf{x}_i^T \mathbf{t} \tag{9}$$

which gives

$$\hat{w}_i^{(j)} = \frac{\mathbf{x}_i^T \mathbf{t}}{1 \cdot |\mathbf{x}_i^T \mathbf{t}|} \left( \left| \mathbf{x}_i^T \mathbf{t} \right| - \lambda \right) \tag{10}$$

Since the update is not dependent any previous parameter values we conclude that only one full pass is needed for convergence. □

# Task 3

Like previously $\mathbf{X}$ is an orthogonal basis, so

$$\mathbf{x}_i^T \mathbf{r}_i^{(j-1)} = \mathbf{x}_i^T \mathbf{t} = \mathbf{x}_i^T \left( \mathbf{X}\mathbf{w}^* + \mathbf{e} \right) = w_i^* + \mathbf{x}_i^T \mathbf{e} \tag{11}$$

since $\mathbf{x}_i^T \mathbf{X}$ is a $1 \times N$ vector with zeros on all elements except the i:th element, which is 1. Since $\mathbf{e} \sim \mathcal{N}(\mathbf{0}_N, \boldsymbol{\sigma}_N)$, we have $\lim_{\sigma \to 0} \mathbf{e} = \mathbf{0}$, which gives

$$\lim_{\sigma \to 0} \mathbf{x}_i^T \mathbf{r}_i^{(j-1)} = w_i^* \tag{12}$$

and thus, in the limit:

$$\left| \mathbf{x}_i^T \mathbf{r}_i^{(j-1)} \right| > \lambda \Rightarrow |w_i^*| > \lambda \tag{13}$$

Using 7, we have (for $|w_i^*| > \lambda$)

$$\hat{w}_i^{(1)} = \frac{w_i^*}{|w_i^*|} \left( |w_i^*| - \lambda \right) = w_i^* - \frac{w_i^*}{|w_i^*|} \lambda = \begin{cases} w_i^* - \lambda, & w_i^* > \lambda \\ w_i^* + \lambda, & w_i^* < -\lambda \end{cases} \tag{14}$$

where $\mathbf{X}$ is assumed to be ortho*normal* (if just orthogonal there will be a factor $\frac{1}{\mathbf{x}_i^T \mathbf{x}_i}$ in the solution). For $|w_i^*| \leq \lambda$, the updating rule gives

$$\hat{w}_i^{(1)} = 0 \tag{15}$$

Combining it all gives

$$\lim_{\sigma \to 0} E(\hat{w}_i^{(1)} - w_i^*) = \begin{cases} w_i^* - \lambda - w_i^*, & w_i^* > \lambda \\ w_i^* + \lambda - w_i^*, & w_i^* < -\lambda \\ 0 - w_i^*, & |w_i^*| \leq \lambda \end{cases} = \begin{cases} -\lambda, & w_i^* > \lambda \\ \lambda, & w_i^* < -\lambda \\ -w_i^*, & |w_i^*| \leq \lambda \end{cases} \tag{16}$$

□

## Task 4

The resulting plots are shown in figure 1. It is apparent that higher regularization leads to less noisy predictions, that does not fit the measured data as good as with lower regularization. Table 1 gives the number of non-zero weights for the different values of $\lambda$

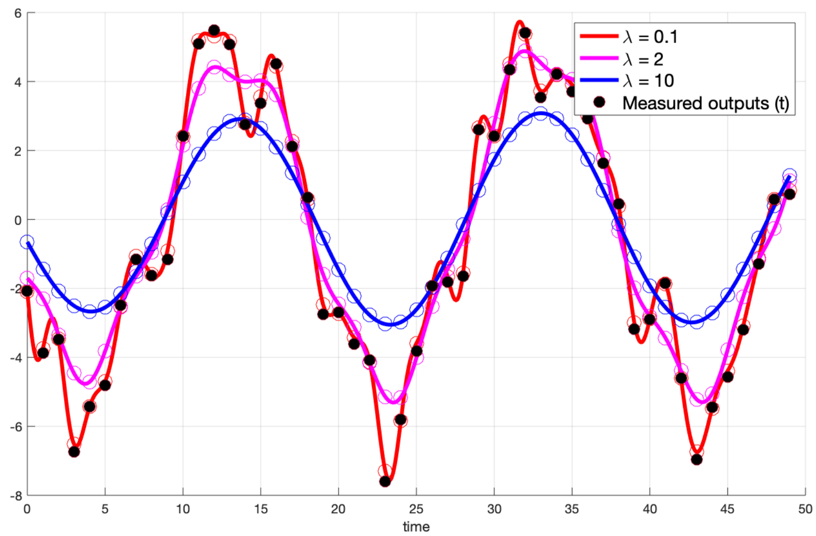


Figure 1: Fitted results for three values of the regularization parameter. Crosses are predicted points and lines are predicted with the interpolated times and regression matrix.

Table 1: Number of non-zero parameters for different regularization strengths.

| $\lambda$ | # non-zero $w_i$:s |
|---|---|
| 0.1 | 230 |
| 2 | 17 |
| 10 | 4 |

## Task 5

10 fold cross validation of gives an optimal regularization parameter $\lambda = 2.1544$, after testing 100 values of $0.1 \leq \lambda \leq 10$ (evenly spaced on a logarithmic scale) - see figure 2 which depicts training and validation error as a function of regularization. The fitted line with optimal regularization is shown in figure 3

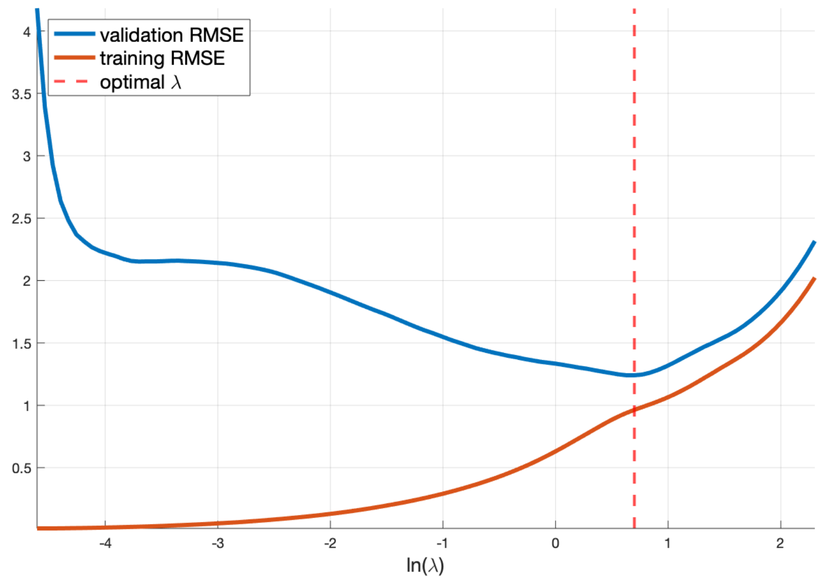


Figure 2: 10-fold cross validation RMSE results.

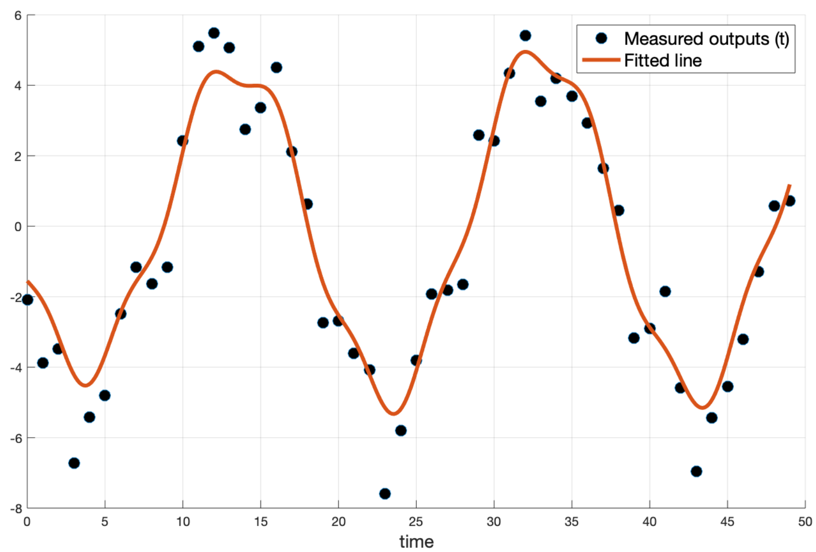


Figure 3: Measurements and line predicted with optimal regularization.

# Task 6

The optimal regularization parameter for the audio file was significantly smaller than for the previous data (see figure 4). This is not very surprising, perhaps, since the average signal intensity in the previous data was 3.32, while it was 0.0176 for the audio file. This means that the "real" significant weights in the audio file are small, and thus that the estimated weights must be allowed to be small as well.

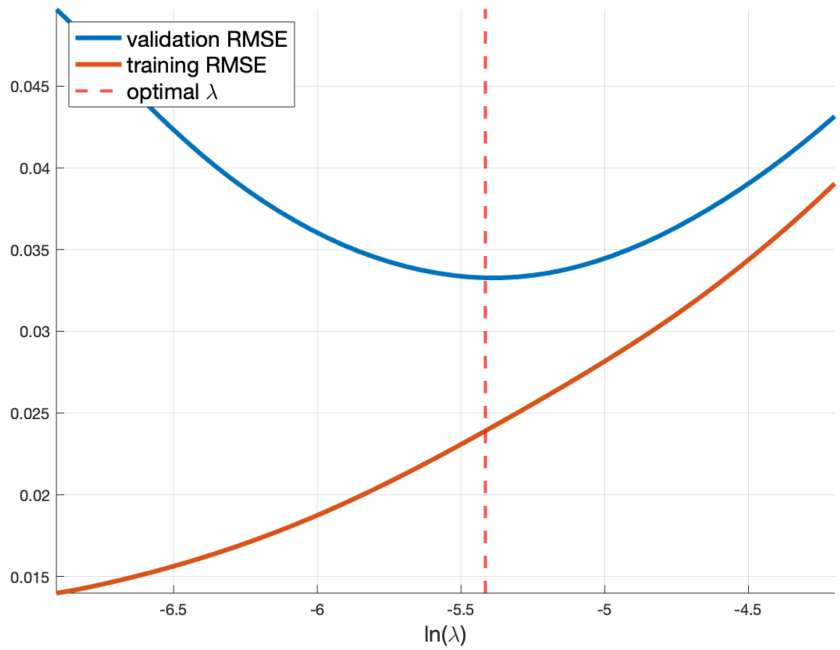


Figure 4: 10-fold cross validation RMSE results.

# Task 7

The audio after cleaning is still very noisy, although there is a significant difference to the uncleaned audio. In my opinion, using a slightly higher regularization (0.006 rather than the optimized 0.0044 works fine for me) gives a better result with less noise, and the piano is still clear. Higher values of the regularization removes most noise, but also distorts the sound of the piano, indicating that frequencies have been removed from the piano.