FMAN45 - Assignment 1

Jingmo Bai

April 25, 2023

1 Task 1

The LASSO solves the minimization problem, by considering a coordinate-wise approach. The objective is to

$$\min_{\omega_i} \frac{1}{2}||r_i - x_i\omega_i||_2^2 + \lambda|\omega_i| \tag{1}$$

By differentiating the expression with respect to ω_i and setting it to 0

$$\frac{d}{d\omega_i}\big(\frac{1}{2}||r_i - x_i\omega_i||_2^2 + \lambda|\omega_i|\big) = 0$$

Given that $\omega_i \neq 0$, we can get

$$\begin{aligned}
\frac{1}{2} \cdot 2(r_i - x_i\omega_i)(-x_i) + \lambda\frac{\omega_i}{|\omega_i|} = 0 \iff \\
-x_i^T(r_i - x_i\omega_i) + \lambda\frac{\omega_i}{|\omega_i|} = 0 \iff \\
-x_i^T r_i + x_i^T x_i\omega_i + \lambda\frac{\omega_i}{|\omega_i|} = 0 \iff \\
x_i^T r_i = x_i^T x_i\omega_i + \lambda\frac{\omega_i}{|\omega_i|} \iff \\
\big(x_i^T x_i + \frac{\lambda}{|\omega_i|}\big)\omega_i = x_i^T r_i
\end{aligned} \tag{2}$$

Since $x_i^T x_i \geq 0$ and $\lambda \geq 0$, we take the absolute value of both sides

$$\begin{aligned}
\big(x_i^T x_i + \frac{\lambda}{|\omega_i|}\big)|\omega_i| = |x_i^T r_i| \iff \\
|\omega_i| = \frac{|x_i^T r_i| - \lambda}{x_i^T x_i}
\end{aligned}$$

And we can also find from (2) that the sign of ω_i is the same as the sign of $x_i^T r_i$, so we can get

$$\begin{aligned}
\omega_i = sgn(x_i^T r_i)|\omega_i| \iff \\
\omega_i = \frac{x_i^T r_i}{|x_i^T r_i|}\frac{|x_i^T r_i| - \lambda}{x_i^T x_i} \iff \\
\omega_i = \frac{x_i^T r_i(|x_i^T r_i| - \lambda)}{|x_i^T r_i| x_i^T x_i}
\end{aligned} \tag{3}$$

Which is the first line of the equation we need to verify.

2 Task 2

Given the orthogonal regression matrix, we know that $X^T X = I_N$, we want wo show that

$$\hat{\omega_i}^{(2)} - \hat{\omega_i}^{(1)} = 0, \forall i$$

$X^T X = I_N$ means that $x_i^T x_l = 0$ when $l \neq i$. Therefore, we can get

$$r_i^{(j-1)} = (t - \sum_{l<i} x_l \hat{w}_l^{(j)} - \sum_{l>i} x_l \hat{w}_l^{(j-1)}) \Longleftrightarrow$$

$$x_i^T r_i^{(j-1)} = x_i^T (t - \sum_{l<i} x_l \hat{w}_l^{(j)} - \sum_{l>i} x_l \hat{w}_l^{(j-1)}) = x_i^T t - 0 - 0 = x_i^T t$$

By replacing $x_i^T r_i^{(j-1)}$ by $x_i^T t$ in the equation, we can get

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

$$\hat{w}_i^{(j)} = \frac{x_i^T t}{x_i^T x_i |x_i^T t|}(|x_i^T t| - \lambda)$$

Since it is independent of j, we have proved that the coordinate descent solver will converge in at most 1 full pass.

3 Task 3

Given the orthogonal regression matrix, we know that $X^T X = I_N$, hence

$$\hat{w}_i^{(j)} = \frac{x_i^T t}{x_i^T x_i |x_i^T t|}(|x_i^T t| - \lambda) \Longrightarrow$$

$$\hat{w}_i^{(j)} = x_i^T t - \lambda sgn(x_i^T t)$$

When $\sigma \to 0$

$$t = X w^*$$
$$x_i^T r_i^{(j-1)} \to w_i^*$$
$$\lim_{\sigma \to 0} x_i^T t = \lim_{\sigma \to 0} x_i^T X w^* = w^*$$

The first case when $x_i^T r_i^{(j-1)} > \lambda$

$$\lim_{\sigma \to 0} E(\hat{w_i}^{(1)} - w_i^*) = \lim_{\sigma \to 0} E(x_i^T t - \lambda sgn(x_i^T t) - w_i^*)$$

$$= E(w_i^* - \lambda sgn(w_i^*) - w_i^*)$$
$$= -\lambda E(sgn(w_i^*)) = -\lambda$$

The second case when $x_i^T r_i^{(j-1)} < -\lambda$

$$\lim_{\sigma \to 0} E(\hat{w_i}^{(1)} - w_i^*) = E(w_i^* - \lambda sgn(w_i^*) - w_i^*)$$

$$= -\lambda E(sgn(w_i^*)) = \lambda$$

The third case when $|x_i^T r_i^{(j-1)}| \leq \lambda$, we note that $\hat{w}_i^{(j)} = 0$, hence

$$\lim_{\sigma \to 0} E(\hat{w_i}^{(1)} - w_i^*) = \lim_{\sigma \to 0} E(0 - w_i^*) = -w_i^*$$

Finally, we have shown the expression of the LASSO estimate's bias.
The full name of LASSO is Least Absolute Shrinkage and Selection Operator. As we can see, the bias of the estimate stops increasing when $|w_i^*| > \lambda$, and the maximum value of bias keeps as λ. This operator can shrink the weights with the help of the hyper-parameter λ, as a form of the least absolute of the weights, preventing the weights to be too large so the model will be overfitting.

4 Task 4

In this task, we implement three variants of coordinate descent solver using LASSO with different values of λ to reconstruct a sum of two sinusoids with different frequencies with an additive Gaussian noise.
Below are reconstruction plots using different values of λ.

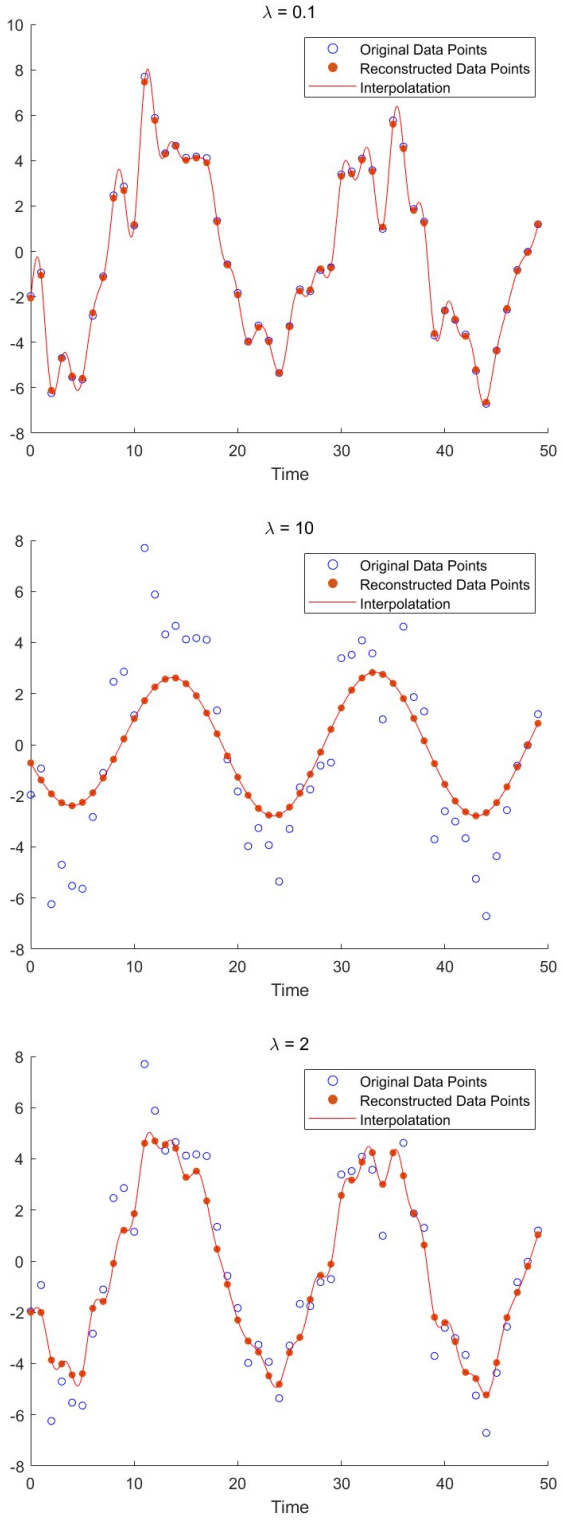

Figure 1: Plots of reconstructions using different values of λ.

As the figure shows, for $\lambda = 0.1$, the model is overfitting, it tends to fit all the noise so it can not generalize well. For $\lambda = 10$, the model is underfitting, it fails to fit the true signal and missing too many details. For $\lambda = 2$, the reconstruction plot seems to be suitable and reasonable, as a sum of two sinusoids with different frequencies.

For the second part of this task, the numbers of non-zero coordinates for different values of λ are shown in the table below.

λ	Number of non-zero coordinates
0.1	273
10	5
2	33

Table 1: Number of non-zero coordinates using different values of hyperparameter.

When comparing these to the actual number of nonzero coordinates needed to model the data given the true frequencies, which is 4 coordinates, we can find that we need more non-zero coordinates than the actual number to get a suitable model. The number of non-zero coordinates is inversely proportional to λ.

5 Task 5

In this task, we implement a K-fold cross-validation scheme for the LASSO solver implemented in Task 4. It is a 10-fold cross-validation with 100 different λ values ranging from 0.01 to $\max(|X^T t|)$.

To find the optimal λ for reconstruction, we minimize the RMSE of the validation dataset. The figure below shows the RMSE of estimation data and validation data. The dashed vertical line shows the optimal λ we select.

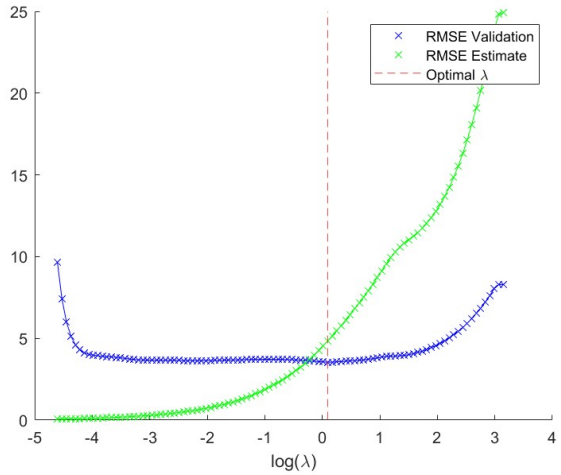

Figure 2: Estimation and validation RMSE using K-fold cross-validation

The plot indicates that the smaller the value of λ, the smaller the estimation RMSE is. Because the model tends to be overfitting. On the contrary, with the larger value of λ, the

estimation RMSE goes larger and the model is underfitting. We can find the optimal value of λ when the validation RMSE is minimized.

The optimal λ we select is 1.1. Below is the reconstruction plot using this optimal λ.

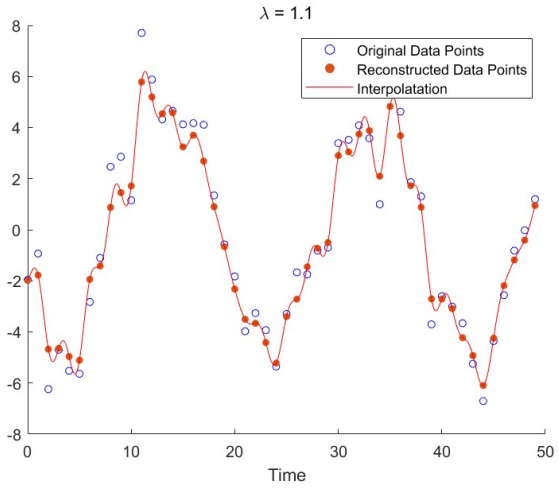

Figure 3: Plots of reconstructions using optimal λ

Compared with the λ value we choose before, which is 2, the plot of this optimal value is more fitted to the data and captures more details.

6 Task 6

In this task, we attempt to perform denoising of a noisy excerpt of piano music, by implementing a multi-frame K-fold cross-validation scheme to obtain the optimal λ for all frames. It is a 3-fold cross-validation with 100 different λ values ranging from 0.0001 to $\max_i(|X^T t_i|)$.

To find the optimal λ for reconstruction, we minimize the RMSE of the validation dataset. The figure below shows the RMSE of estimation data and validation data. The dashed vertical line shows the optimal λ we select. The optimal λ which can minimize the validation RMSE is 0.0049.

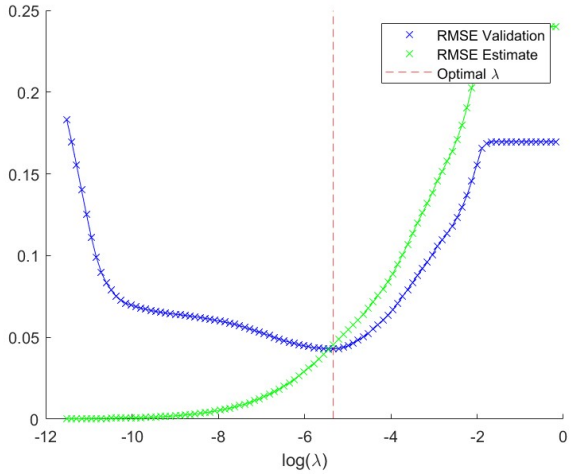

Figure 4: Estimation and validation RMSE for different λ

7 Task 7

After using the optimal λ we obtained from the last task, we denoise the test data. In comparison to the original recording, the noise is cancelled to some extent.

Trying to change the λ to a smaller value makes the denoising effect weakened. The result of using the value of 0.0001 is almost the same as the original noised recording.

Trying to change the λ to a larger value makes the denoising effect strengthened. The result of using the value of 0.01 is that the noise has been cancelled almost completely. But the music seems to be "far away", it loses some original details.