Penalized regression via the LASSO

FMAN45: ML Assignment-1

Praveenkumar Hiremath

August 17, 2023

Exercise 1: Coordinate descent minimizer solution

The minimization problem at hand is given by

$$\underset{\mathbf{w}}{minimize}\frac{1}{2}\|\mathbf{t\text{-}Xw}\|_2^2+\lambda\|\mathbf{w}\|_1. \tag{1}$$

Here, $\mathbf{t}\in\mathbb{R}^N$, $\mathbf{X}\in\mathbb{R}^{NXM}$, and $\mathbf{w}\in\mathbb{R}^M$ are the response variable (target), regression matrix, and explanatory variable or weights, respectively. $\lambda\geq 0$ is a capacity (or regularization) hyperparameter. To obtain the i^{th} coordinate w_i in the coordinate-wise approach, it is needed to

$$\underset{\mathbf{w}}{minimize}\frac{1}{2}\|\mathbf{r}_i-\mathbf{x}_i w_i\|_2^2+\lambda|w_i|, \tag{2}$$

where, $\mathbf{x}_i$ is the i^{th}-vector of the regression matrix and $\mathbf{r}_i=\mathbf{t}-\sum_{l\neq i}\mathbf{x}_l w_l$ is the residual vector without the effect of i^{th} regressor. To achieve the goal of computing updated weight, w_i (an element of $\mathbf{w}$), equation (2) will be differentiated with respect to w_i and is equated to zero i.e.

$$\begin{aligned}
&\frac{d}{dw_i}\left[\frac{1}{2}\|\mathbf{r}_i-\mathbf{x}_i w_i\|_2^2+\lambda|w_i|\right]=0\\
&\implies \mathbf{x}_i^T\mathbf{x}_i w_i-\mathbf{x}_i^T\mathbf{r}_i+\lambda\frac{w_i}{|w_i|}=0\\
&\implies \mathbf{x}_i^T\mathbf{x}_i w_i=\mathbf{x}_i^T\mathbf{r}_i-\lambda sign(w_i)\\
&\implies w_i=\frac{\mathbf{x}_i^T\mathbf{r}_i-\lambda sign(w_i)}{\mathbf{x}_i^T\mathbf{x}_i}
\end{aligned} \tag{3}$$

Equation (3) is valid for $\mathbf{x}_i^T\mathbf{x}_i>0.0$ and $\lambda>0$ (given). Therefore, $sign(w_i)=sign(\mathbf{x}_i^T\mathbf{r}_i)=\frac{\mathbf{x}_i^T\mathbf{r}_i}{|\mathbf{x}_i^T\mathbf{r}_i|}$. Now, equation (3) can be re-written as

$$w_i=\frac{\mathbf{x}_i^T\mathbf{r}_i}{\mathbf{x}_i^T\mathbf{x}_i|\mathbf{x}_i^T\mathbf{r}_i|}\left(\left|\mathbf{x}_i^T\mathbf{r}_i\right|-\lambda\right). \tag{4}$$

However, at j^{th} iteration, $\mathbf{r}_i$ is defined using weights from previous iteration $(j-1)$. For that reason, at j^{th} iteration, i^{th} weight becomes,

$$w_i^j=\frac{\mathbf{x}_i^T\mathbf{r}_i^{j-1}}{\mathbf{x}_i^T\mathbf{x}_i\left|\mathbf{x}_i^T\mathbf{r}_i^{j-1}\right|}\left(\left|\mathbf{x}_i^T\mathbf{r}_i^{j-1}\right|-\lambda\right). \tag{5}$$

Here, $\left|\mathbf{x}_i^T\mathbf{r}_i^{j-1}\right|>\lambda$ must hold good. Hence proved.

Exercise 2: Show that $w_i^{(1)} = w_i^{(2)}$, $\forall i$.

The given regression matrix $\mathbf{X}$ is orthonormal. This means that $M = N$, $\mathbf{X}^T\mathbf{X} = \mathbf{I}_N$. $\mathbf{I}_N$ is an identity matrix of size N. Therefore, with $\mathbf{x}_i$ being the i^{th} vector of the regression matrix indicates that,

$$\begin{aligned} \mathbf{x}_i^T\mathbf{x}_j &= 1, \qquad \forall i = j \\ \mathbf{x}_i^T\mathbf{x}_j &= 0, \qquad \forall i \neq j \end{aligned} \tag{6}$$

$\mathbf{x}_i^T\mathbf{x}_j = 1$ and $\mathbf{x}_i^T\mathbf{x}_j = 0$, $\forall i = j$ and $\forall i \neq j$, respectively. Additionally, it is given that,

$$\mathbf{r}_i^{j-1} = \mathbf{t} - \sum_{l<i} \mathbf{x}_l w_l^{(j)} - \sum_{l>i} \mathbf{x}_l w_l^{(j-1)} \tag{7}$$

For $\left|\mathbf{x}_i^T\mathbf{r}_i^{j-1}\right| > \lambda$, plugging equation (7) in equation (5) from exercise 1 yields,

$$w_i^j = \frac{\mathbf{x}_i^T(\mathbf{t} - \sum_{l<i}\mathbf{x}_l w_l^{(j)} - \sum_{l>i}\mathbf{x}_l w_l^{(j-1)})}{\mathbf{x}_i^T\mathbf{x}_i\left|\mathbf{x}_i^T(\mathbf{t} - \sum_{l<i}\mathbf{x}_l w_l^{(j)} - \sum_{l>i}\mathbf{x}_l w_l^{(j-1)})\right|}\left(\left|\mathbf{x}_i^T(\mathbf{t} - \sum_{l<i}\mathbf{x}_l w_l^{(j)} - \sum_{l>i}\mathbf{x}_l w_l^{(j-1)})\right| - \lambda\right) \tag{8}$$

With the help of equation (6), equation (8) can be re-written as,

$$w_i^{(j)} = \frac{\mathbf{x}_i^T\mathbf{t}}{\mathbf{x}_i^T\mathbf{x}_i|\mathbf{x}_i^T\mathbf{t}|}\left(\left|\mathbf{x}_i^T\mathbf{t}\right| - \lambda\right). \tag{9}$$

For $\left|\mathbf{x}_i^T\mathbf{r}_i^{j-1}\right| \leq \lambda$,

$$w_i^{(j)} = 0. \tag{10}$$

Thus, w_i^j only depends on $\mathbf{t}$, $\mathbf{x}_i$ and λ and does not depend on previous (j^{th}) estimates for w_i^j. This implies that,

$$w_i^j = w_i^{j-1} \implies w_i^{(2)} - w_i^{(1)} = 0, \tag{11}$$

and convergence is achieved after only iteration ($j = 1$).

Exercise 3: A noisy data t, given by, $\mathbf{t} = \mathbf{X}\mathbf{w}^* + \mathbf{e}, \mathbf{e} \sim \boldsymbol{N}(\mathbf{0}_N, \sigma\mathbf{I}_N)$.

Given that,

$$\mathbf{t} = \mathbf{X}\mathbf{w}^* + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}_N, \sigma\mathbf{I}_N), \tag{12}$$

where, $\mathbf{w}^*$ is the particular $\mathbf{w}$ defining the used to generate the noisy data. $\boldsymbol{N}(\cdot)$, σ, and $\mathbf{0}_N$ are the Gaussian distribution, standard deviation, and a zero-column vector of size N, respectively. Using equation (8) in exercise 2 and $\mathbb{E}(A + B) = \mathbb{E}(A) + \mathbb{E}(B)$, we have,

$$\mathbb{E}(w_i^{(1)} - w_i^*) = \mathbb{E}(w_i^{(1)}) - \mathbb{E}(w_i^*)$$

$$\implies \mathbb{E}(w_i^{(1)} - w_i^*) = \mathbb{E}\left(\frac{\mathbf{x}_i^T\mathbf{r}_i^0}{\mathbf{x}_i^T\mathbf{x}_i|\mathbf{x}_i^T\mathbf{r}_i^0|}\left(\left|\mathbf{x}_i^T\mathbf{r}_i^0\right| - \lambda\right)\right) - \mathbb{E}(w_i^*), \qquad for\ \left|\mathbf{x}_i^T\mathbf{r}_i^{j-1}\right| > \lambda. \tag{13}$$

However, $\mathbf{X}$ is an orthonormal matrix. Therefore using equation (7) from exercise 2,

$$\mathbf{x}_i^T\mathbf{r}_i^{(j-1)} = \mathbf{x}_i^T(\mathbf{t} - \sum_{l<i}\mathbf{x}_l w_l^{(j)} - \sum_{l>i}\mathbf{x}_l w_l^{(j-1)}) = \mathbf{x}_i^T\mathbf{t}. \tag{14}$$

Using equation (14) in equation (13) yields,

$$\mathbb{E}(w_i^{(1)} - w_i^*) = \mathbb{E}\left(\frac{\mathbf{x}_i^T\mathbf{t}}{\mathbf{x}_i^T\mathbf{x}_i|\mathbf{x}_i^T\mathbf{t}|}\left(\left|\mathbf{x}_i^T\mathbf{t}\right| - \lambda\right)\right) - \mathbb{E}(w_i^*). \tag{15}$$

This result (equation (15)) is already used to solve exercise 2. Now using equation (12), $\mathbf{x}_i^T\mathbf{t}$ can be expanded as,

$$\mathbf{x}_i^T\mathbf{t} = \mathbf{x}_i^T(\mathbf{X}\mathbf{w}^* + \mathbf{e}) = \mathbf{x}_i^T\mathbf{X}\mathbf{w}^* + \mathbf{x}_i^T\mathbf{e}.$$

Because $\mathbf{x}_i^T\mathbf{X}$ gives a row vector of size N with only i^{th} element being a nonzero value i.e. unity, we get,

$$\mathbf{x}_i^T\mathbf{t} = \mathbf{w}^* + \mathbf{x}_i^T\mathbf{e}. \tag{16}$$

Further, from equation (12), $\mathbf{e} \sim N(\mathbf{0}_N, \sigma\mathbf{I}_N)$. This implies that,

$$\lim_{\sigma\to 0}\mathbf{x}_i^T\mathbf{r}_i^{(j-1)} = \lim_{\sigma\to 0}\mathbf{x}_i^T\mathbf{r}_i^0 = \lim_{\sigma\to 0}\mathbf{x}_i^T\mathbf{t} = \lim_{\sigma\to 0}\mathbf{x}_i^T\left(\mathbf{X}\mathbf{w}^* + \mathbf{e}\right) = w_i^* \tag{17}$$

The last step of equation (17) comes about because of the orthonormality of $\mathbf{X}$ and $\lim_{\sigma\to 0}\mathbf{e} = 0$. Consequently,

$$\left|\mathbf{x}_i^T\mathbf{r}_i^{(j-1)}\right| > \lambda \implies \left|\mathbf{x}_i^T\mathbf{r}_i^0\right| > \lambda \implies \left|\mathbf{x}_i^T\mathbf{t}\right| > \lambda \implies |w_i^*| > \lambda \tag{18}$$

Assuming w_i^* is not a random variable i.e. $\mathbb{E}(w_i^*) = w_i^*$, $\lim_{\sigma\to 0}\mathbb{E}(w_i^{(1)} - w_i^*)$ can be written as

$$\begin{aligned}
\lim_{\sigma\to 0}\mathbb{E}(w_i^{(1)} - w_i^*) &= \lim_{\sigma\to 0}\mathbb{E}(w_i^{(1)}) - \lim_{\sigma\to 0}\mathbb{E}(w_i^*) \\
&= \lim_{\sigma\to 0}\mathbb{E}\left(\frac{\mathbf{x}_i^T\mathbf{t}}{\mathbf{x}_i^T\mathbf{x}_i|\mathbf{x}_i^T\mathbf{t}|}\left(\left|\mathbf{x}_i^T\mathbf{t}\right| - \lambda\right)\right) - \mathbb{E}(w_i^*) \\
= \frac{w_i^*}{|w_i^*|}(|w_i^*| - \lambda) - w_i^* &= w_i^* - \frac{w_i^*}{|w_i^*|}\lambda - w_i^* = -sign(w_i^*)\lambda
\end{aligned}$$

$$\begin{aligned}
\lim_{\sigma\to 0}\mathbb{E}(w_i^{(1)} - w_i^*) &= -\lambda \qquad w_i^* > \lambda \\
\lim_{\sigma\to 0}\mathbb{E}(w_i^{(1)} - w_i^*) &= \lambda \qquad w_i^* < -\lambda
\end{aligned} \tag{19}$$

Additionally for $\left|\mathbf{x}_i^T\mathbf{r}_i^{(j-1)}\right| \leq \lambda \implies \left|\mathbf{x}_i^T\mathbf{r}_i^0\right| \leq \lambda \implies \left|\mathbf{x}_i^T\mathbf{t}\right| \leq \lambda \implies |w_i^*| \leq \lambda,\ w_i^{(1)} = 0$. Therefore,

$$\lim_{\sigma \to 0} \mathbb{E}(w_i^{(1)} - w_i^*) = 0 - w_i^* \qquad |w_i^*| \leq \lambda. \tag{20}$$

Combining equations (19) and (20), we have,

$$\lim_{\sigma \to 0} \mathbb{E}(w_i^{(1)} - w_i^*) = \begin{cases} -\lambda, & \text{for } w_i^* > \lambda \\ \lambda, & \text{for } w_i^* < -\lambda \\ -w_i^*, & \text{for } |w_i^*| \leq \lambda. \end{cases} \qquad \forall i \tag{21}$$

The Least Absolute Shrinkage and Selection Operator (LASSO) [1] method results in a sparse $\mathbf{w}$ vector. The method sets relatively smaller w_is to 'zero' (shrinkage) and selects w_i that are significant (selection). LASSO seems to be good for a model to be less prone to high variance problem i.e. overfitting of the model (as indicated by the "$\lim_{\sigma \to 0}$" in the above equation (21)). But the bias in equation (21) is sensitive to λ. Larger the λ, the larger the bias (i.e. could lead to underfitting). Therefore, to strike a balance between overfitting and underfitting, an optimal λ should be used.

Exercise 4: K-fold cross-validation scheme for the LASSO.

Implementing cyclic coordinate descent solver

The given data t in the file "A1_data.mat" contains N = 50 data points generated as below.

$$\begin{gathered} t(n) = f(n) + \sigma \cdot e(n) \\ f(n) = \mathrm{Re}\Big\{ 5e^{i2\pi(\frac{n}{20}+\frac{1}{3})} + 2e^{12\pi(\frac{n}{5}-\frac{1}{4})} \Big\} \qquad \text{for } n = 0,...,N-1 \\ e(n) \sim N(0,1) \end{gathered} \tag{22}$$

The regression matrix $\mathbf{X}$ contains 500 candidate sine and cosine pairs as shown below.

$$\begin{gathered} \mathbf{X} = [\mathbf{X}_1 ... \mathbf{X}_{500}] \\ \mathbf{X}_i = [\mathbf{u}_i \mathbf{v}_i] \\ \text{where, } \mathbf{u}_i = [\sin\,(2\pi f_i 0)\ \sin\,(2\pi f_i 1)\ ...\ \sin\,(2\pi f_i (N-1))]^{\mathrm{T}} \text{ and} \\ \mathbf{v}_i = [\cos\,(2\pi f_i 0)\ \cos\,(2\pi f_i 1)\ ...\ \cos\,(2\pi f_i (N-1))]^{\mathrm{T}} \end{gathered} \tag{23}$$

The cyclic coordinate descent solver for the coordinate-wise LASSO solution in equation (5) (from exercise 1) is implemented in the file "lasso_ccd.m".

Subtask 1

To evaluate the influence of the hyperparameter, λ (regularization parameter) on the data points' reconstruction, several values λ = [0.1,1,2,3,4,5,6,7,8,9,10] are tried. The reconstructed data points for all these values of λ are plotted in Figures 1 and 2 in the section appendix. Here, in the main report, plots are provided for only λ = 0.1, 1, 2, 4, and 10, see Figures 1-5. The reconstructed data points (represented by blue star markers) are connected by interpolated lines in black.

A value of $\lambda = 0$ would result in a non-regularized regression model. It can be seen that for λ = 0.1, the model reconstructs all the input (training) data points. This indicates a case of overfitting.

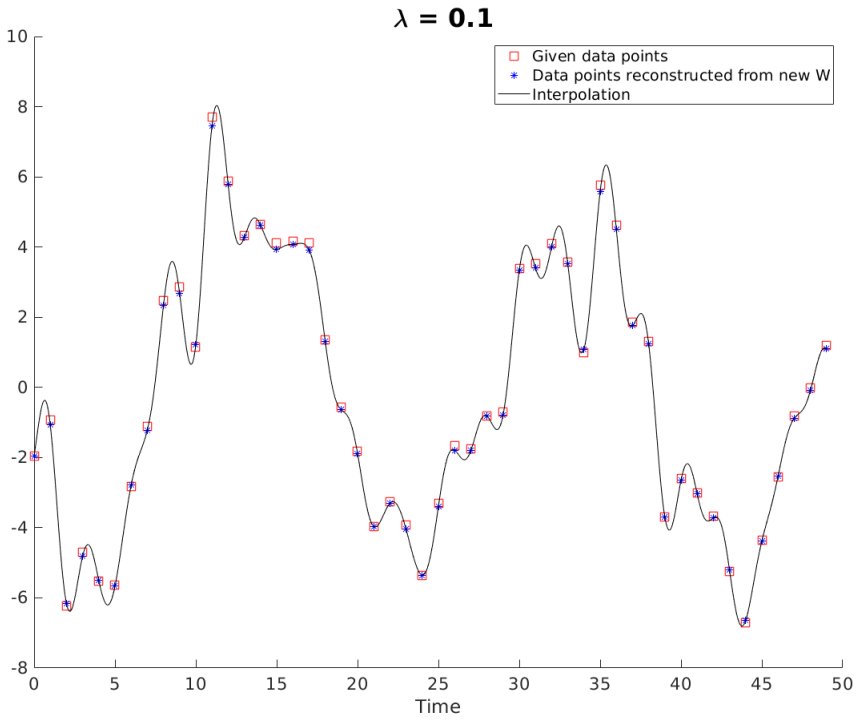

Figure 1: Reconstruction for $\lambda = 0.1$.

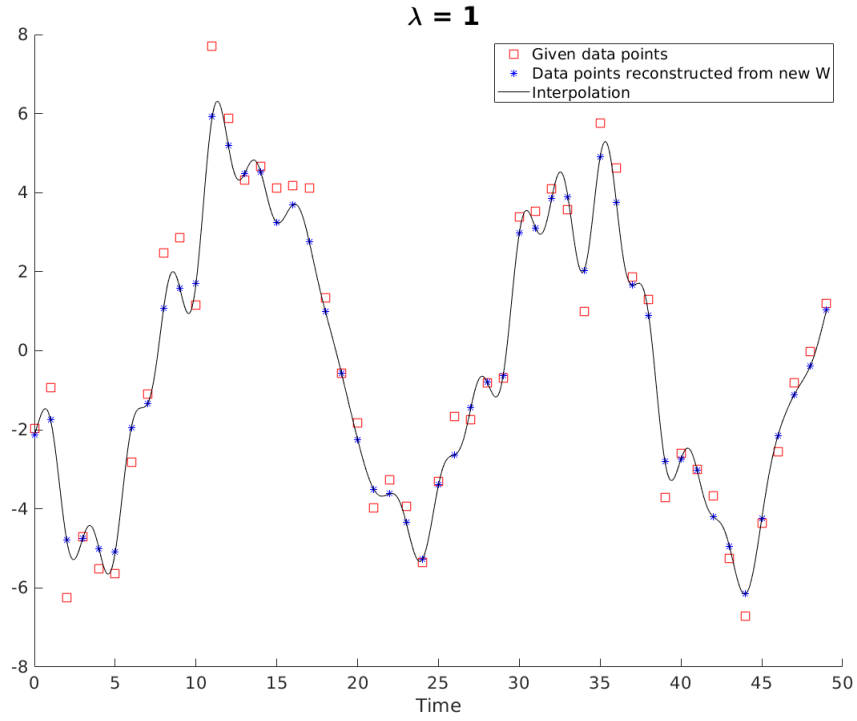

Figure 2: Reconstruction for $\lambda = 1$.

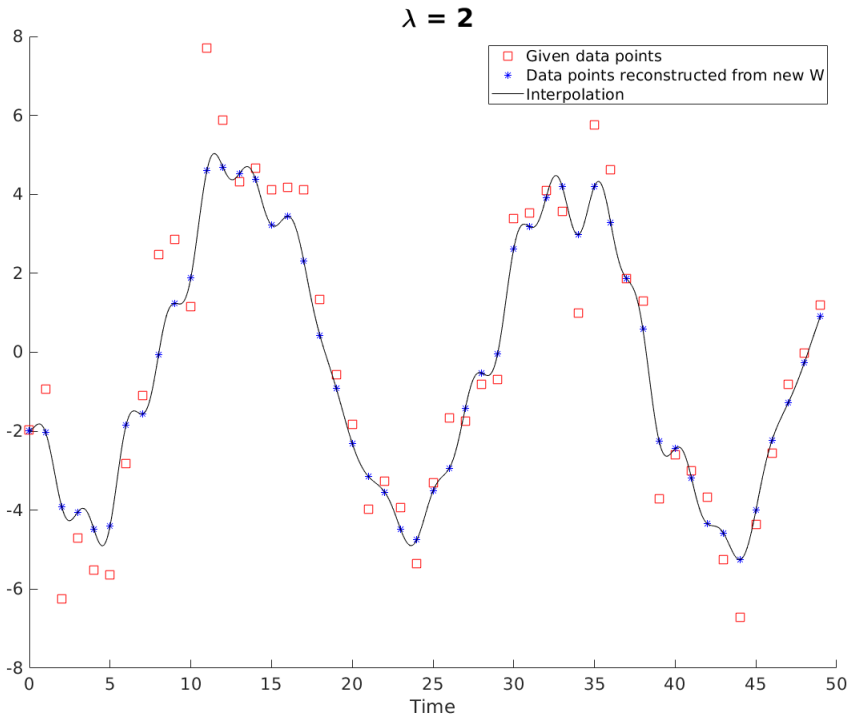

Figure 3: Reconstruction for $\lambda = 2$.

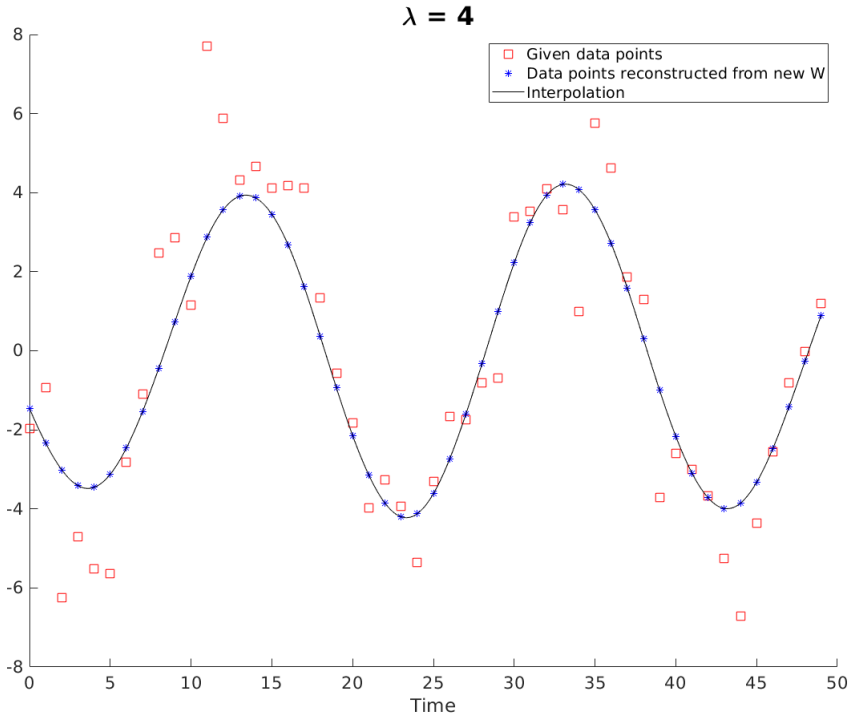

Figure 4: Reconstruction for $\lambda = 4$.

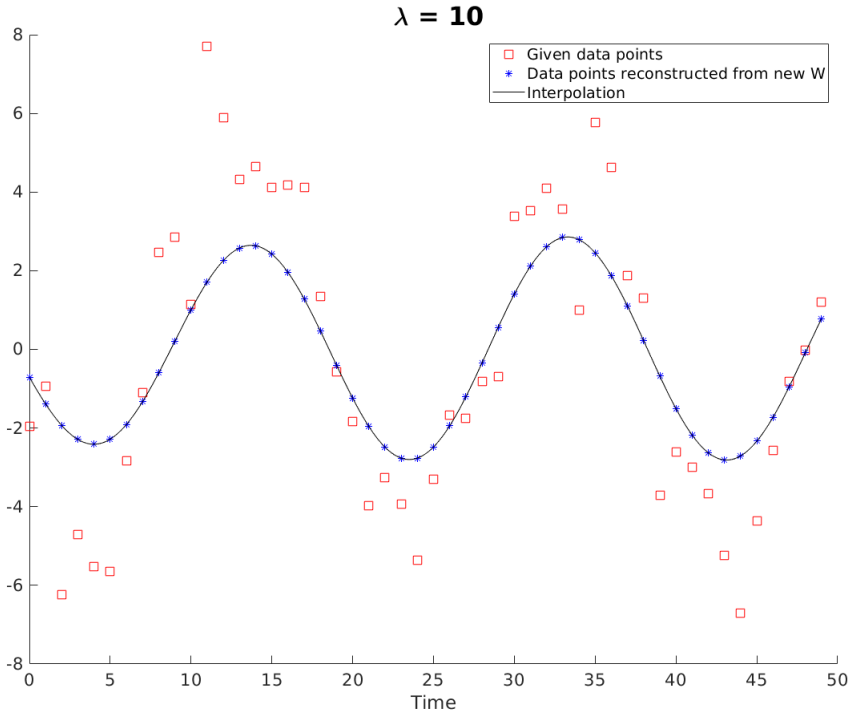

Figure 5: Reconstruction for $\lambda = 10$.

Table 1: Dependence of the number of non-zero coordinates on the choice of λ.

$\lambda =$	0.1	1	2	3	4	5	6	7	8	9	10
# non-zero coordinates =	269	90	46	16	9	7	9	7	8	7	6

As the λ value increases, fewer and fewer data points are reconstructed. This suggests that the model is moving towards underfitting. For $\lambda = 4$ and above, the model gives the interpolated curve to be sinusoidal with more data points as noise (not on the sinusoidal curve). For $\lambda = 10$, even more data points lie away from the interpolated curve. For $\lambda = 1$ and 2, the model seems to be able to achieve a balance between over and underfitting. This conclusion is drawn based on the similarity of interpolated curves in Figures 2 and 3 with the noise-free $f(n)$ curve provided in Figure 6 (this Figure was provided in the assignment description as a means to evaluate the ability of the model to reconstruct the data points). This exercise 4, shows that (i) Smaller $\lambda \implies$ higher variance i.e. model overfitting. (ii) Larger $\lambda \implies$ higher bias i.e. model underfitting. (iii) Therefore λ must lie between 1 and 2.

Subtask 2

It is seen that the number of actual non-zero coordinates required to model the data depends on the choice of λ. The higher the λ, the fewer the non-zero coordinates (i.e. fewer elements of $\hat{\mathbf{W}}$), see Table 1. The observations in Table 1 are justified as the coordinate minimizer needs to perform the minimization task presented in equation (1) (from exercise 1). As a result, a larger number

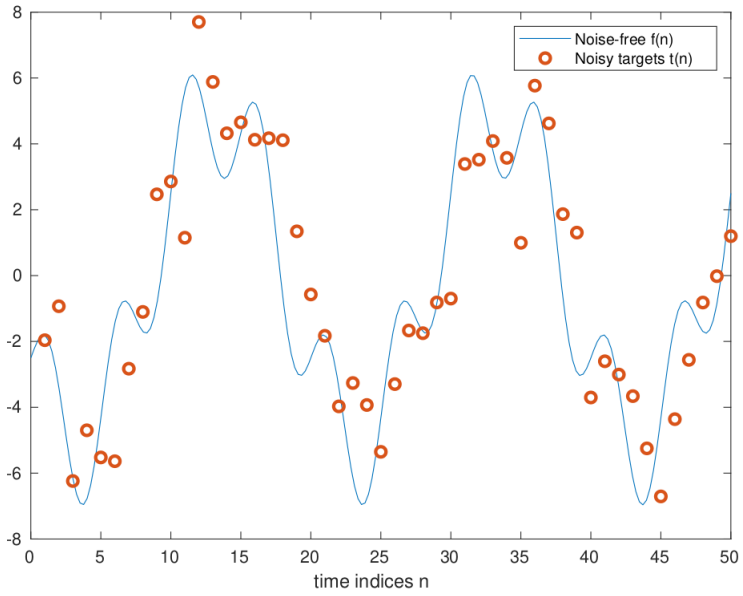

Figure 6: Reconstruction data provided in the assignment description.

of w_i are penalized and zeroed ($\lambda = 10$ gives only 6 non-zero w_is), thus causing underfitting. Likewise, overfitting is the result of many non-zero coordinates. Therefore, an optimal value of the hyperparameter λ is sought after.

Exercise 5: K-fold cross-validation scheme for LASSO solver from exercise 4.

The K-fold cross-validation scheme is implemented in the sketch "lasso_cv.m" following the algorithm given in the assignment description ("**Algorithm 1** K-fold cross-validation for LASSO"). In this task, using the K-fold validation scheme, the dependence of RMSE error of the validation dataset and of the estimation (training) dataset, on the λ-values is investigated. Then an optimal λ value is chosen based on the RMSE error on the validation test. Here the validation RMSE error is chosen because the model is required to perform better on the data that was not a part of the data (estimation dataset) on which it was trained. This way the generalization (or transferability) of the model can be improved.

Figure 7 shows the behavior of validation RMSE and estimation RMSE with varying λ-values. λ-values are chosen such that they are evenly spaced on a log-scale. $K = 10$ and $0.01 < \lambda < max_i(\left|\mathbf{x}_i^T\mathbf{t}\right|)$ are chosen for the K-fold CV scheme. At $\lambda = 0$, validation RMSE is large and estimation RMSE is very low. As the λ increases estimation RMSE keeps on increasing. However, validation RMSE reduces until an optimal value λ_{opt}, and it also steadily increases thereafter. Hence, it is seen that $\lambda = 1.8985$ produces the lowest validation RMSE and is chosen as the optimal value (λ_{opt}). The reconstructed data points using the $\lambda_{opt} = 1.8985$ is presented in Figure 8. This model ($\lambda_{opt} = 1.8985$) seems to provide good reconstruction as the reconstructed data points resemble to a good degree to those seen in Figure 6.

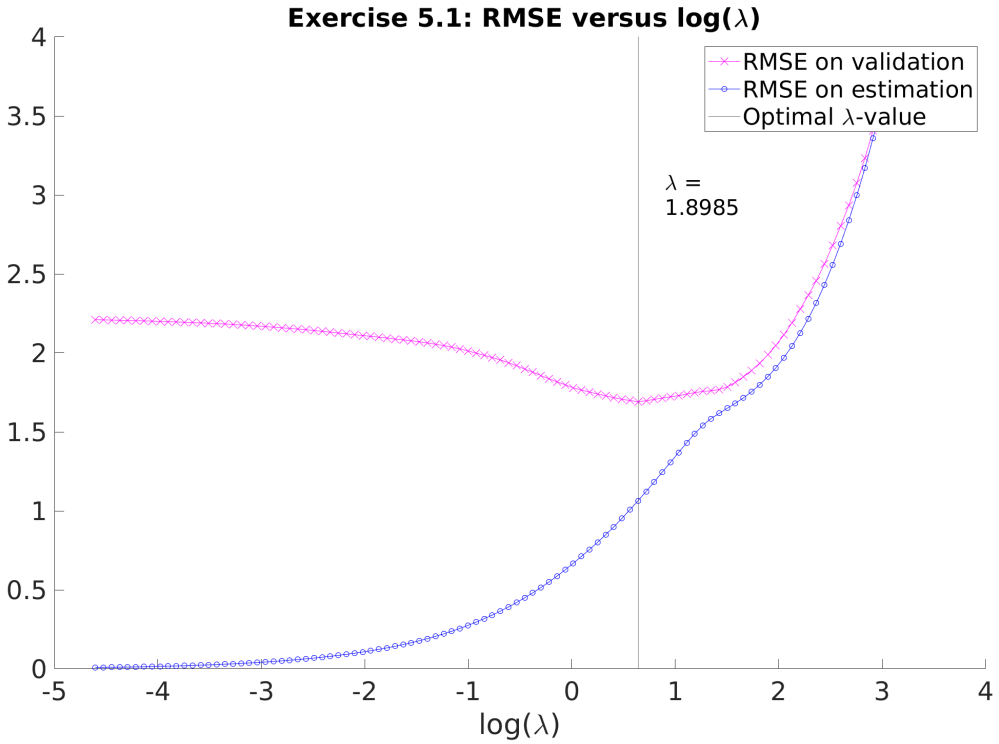

Figure 7: RMSE errors of validation and estimation datasets as a function of λ.

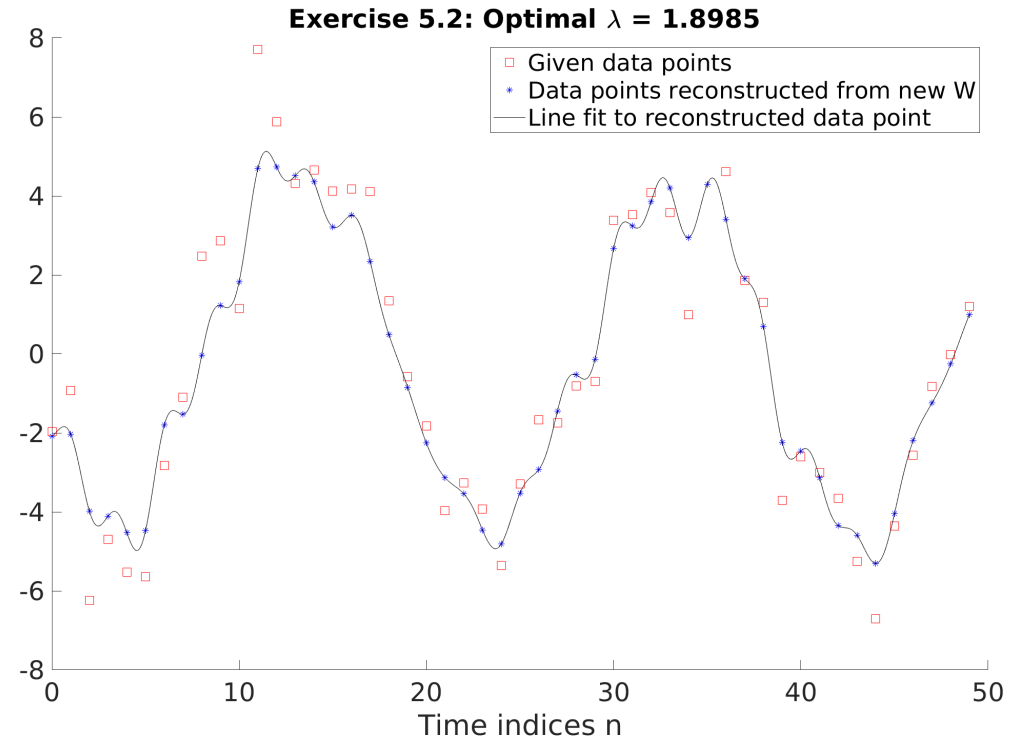

Figure 8: Reconstruction using $\lambda_{opt} = 1.8985$.

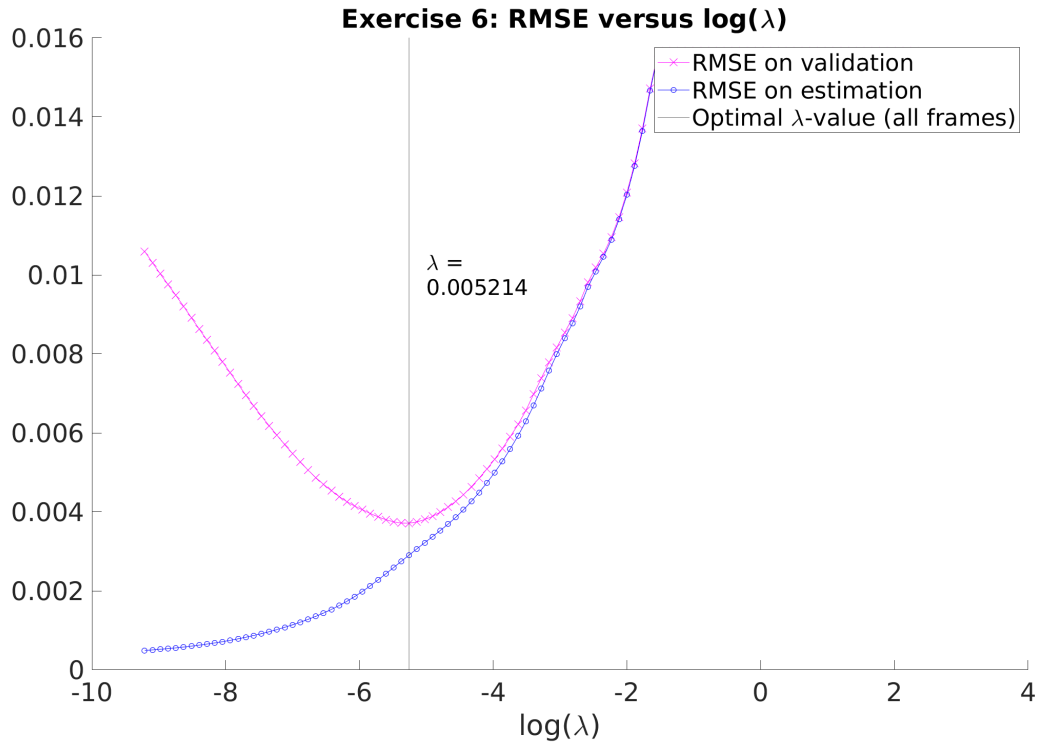

Figure 9: RMSE errors of validation and estimation datasets as a function of λ.

Exercise 6: Denoising of an audio excerpt

Find λ_{opt} for the denoising task.

In this exercise, a noisy audio excerpt is given which is divided into **Ttrain** dataset for estimation and validation, and **Ttest** dataset for testing purposes. The audio excerpt is 5 seconds in total. Again, similar to exercise 5, by using λ_{opt}, the noise from the audio excerpt can be minimized. For this purpose, a K-fold cross-validation scheme is implemented in the file "multiframe_lasso_cv.m". **Ttrain** was further divided into 55 frames with each frame being of length 352 (because Xaudio $\rightarrow \mathbf{X} \in \mathbb{R}^{352X2000}$ and Ttrain $\rightarrow \mathbf{X} \in \mathbb{R}^{19404}$). Again K = 10 is used for K-fold. For each of the 55 frames, λ_{opt} is computed and the average of all the 55 λ_{opt} values is taken as the optimal λ_{opt} for all the frames. The RMSE vs log(λ) curves shown in Figure 9 suggest that $\lambda_{opt} = 0.005214$ gives the least RMSE on validation. This model is saved as "model_from_exercise_6.mat" for the denoising task in exercise 7.

Exercise 7: $\lambda_{opt} = 0.005214$ is used to denoise the test data "Ttest".

The very low $\lambda_{opt} = 0.005214$ raised concern over the generalization of the model (from exercise 6). The concern was if the model had overfitting problem. However, for $\lambda_{opt} = 0.005214$, clearly less static (background signal) can be heard after denoising the excerpt. $\lambda_{opt} = 0.01$, 0.02, 0.04, and 0.1 are also tried. $\lambda_{opt} = 0.01$, 0.02, and 0.04 provided better noise-cancelling than $\lambda_{opt} = 0.005214$. However, I could not decide which among $\lambda_{opt} = 0.01$, 0.02, 0.04, and 0.1 gave the best noise-cancelling as they almost sounded similar. Upon trying $\lambda_{opt} = 0.1$, the noise-cancelling ability was reduced. The denoised audio files for these different λ_{opt} values are attached with the submission.

References

[1] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

1 Appendix

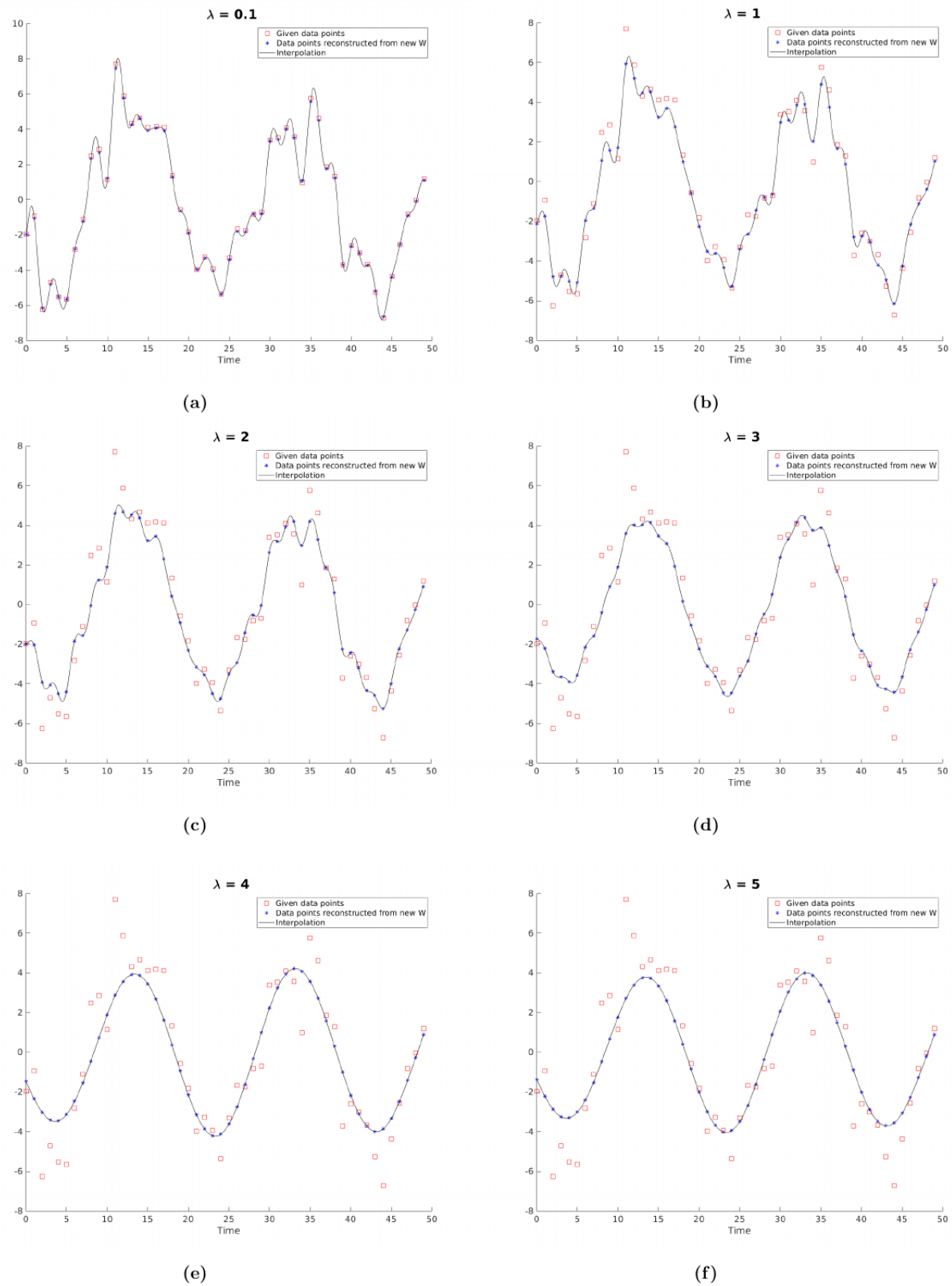

Figure 10: Exercise 4: Reconstruction for $\lambda =$ (a) 0.1, (b) 1, (c) 2, (d) 3, (e) 4 and (f) 5.

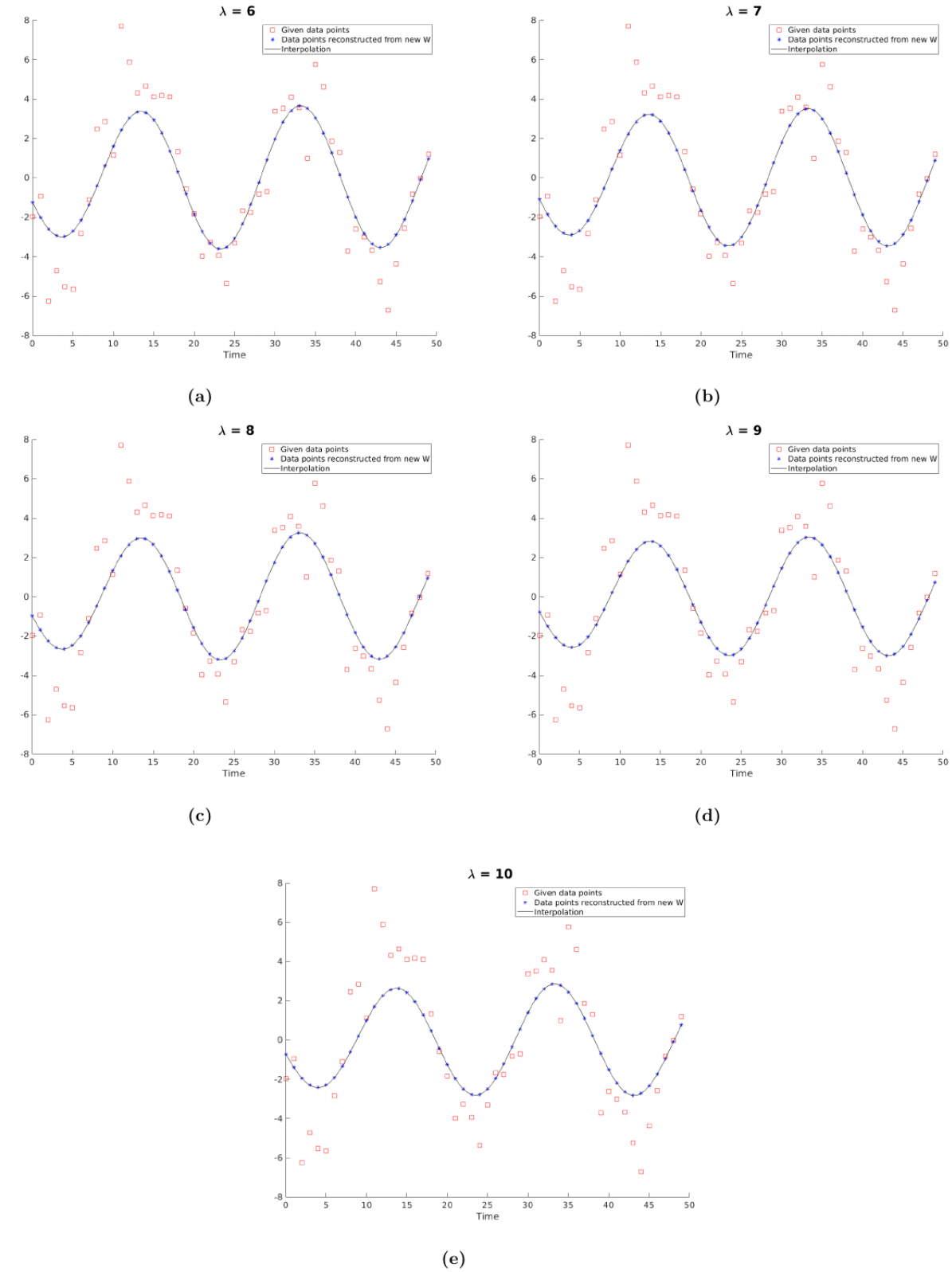

Figure 11: Exercise 4: Reconstruction for $\lambda =$ (a) 6, (b) 7, (c) 8, (d) 9 and (e) 10.