Question 0

0.1 We might choose as follows:

- $\phi^{(1)}$ should be a trigonometric transform, as it would capture the periodic changes in the time series.
- $\phi^{(3)}$ should be a linear transform, a more complex transformation would over-fit to the data.
- $\phi^{(4)}$ should be an RBF transformation, as this is a good all-round function to approximate a function we do not have much information about.

0.2 Here we consider what happens to testing and training errors when the unnecessary variable $x^{(2)}$ is included in the analysis. Denote $\mathbf{w}_{LS}$ as the least square estimate fitted without $x^{(2)}$ and $\mathbf{w}'_{LS}$ the estimate fitted including $x^{(2)}$. Similarly, denote $\boldsymbol{\phi}(\mathrm{X})$ as the matrix with columns $\boldsymbol{\phi}^{(i)}(x^{(i)})$ for $i = 1, 3, 4$ and $\boldsymbol{\phi}'(\mathrm{X}')$ the equivalent for the model including $x^{(2)}$.

Training Error: We can see that the objective function for $\mathbf{w}_{LS}$ is the same as that of $\mathbf{w}'_{LS}$, but with the constraint $w_2 = 0$, hence

$$\begin{aligned} E(D_0, \mathbf{w}_{LS}) &= \min_{w\,:\,w_2=0} \sum_{i \in D_0} [y_i - f(\mathbf{x}_i; \mathbf{w}_{LS})]^2 \\ &\geq \min_{w} \sum_{i \in D_0} \left[y_i - f(\mathbf{x}_i; \mathbf{w}'_{LS})\right]^2 \\ &= E(D_0, \mathbf{w}'_{LS}) \end{aligned}$$

and so the training error decreases.

Testing Error: The testing error would increase as the f_{LS} would capture a pattern which is not present in the distribution of the response, hence we would expect a larger discrepancy in the response and fitted values of the testing set.

General Fit: The combination of the training error decreasing and the testing error increasing suggests a lack of generalisability, and thus over-fitting.

Question 1

1.1 We already know for $\mathbf{y} = (y_1, \cdots, y_n)$ and $\mathbf{f}(\mathrm{X}) = (f_1(\mathbf{x}_1), \cdots, f_n(\mathbf{x}))$

$$\begin{aligned} p(\mathbf{y}|\mathbf{f}(\mathrm{X}), \sigma^2) &= \mathcal{N}_y(\mathbf{f}(\mathrm{X}), \sigma^2\mathbf{I}) \\ p(\mathbf{f}(\mathrm{X})|\sigma) &= \mathcal{N}_{\mathrm{f}}(\mathbf{0}, \mathbf{K}) \end{aligned}$$

Identity (2.115) in Bishop (2006) [1] tells us $p(y|\sigma, \mathbf{K})$ is a normal distribution $\mathcal{N}(\boldsymbol{\mu}_{\mathrm{y}}, \Sigma_{\mathrm{y}})$ where

$$\begin{aligned} \boldsymbol{\mu}_{\mathrm{y}} &= \mathbf{I}\boldsymbol{\mu}_{\mathrm{f}} = \mathbf{0} \\ \Sigma_{\mathrm{y}} &= \sigma^2\mathbf{I} + \mathbf{I}^T\mathbf{K}\mathbf{I} = \sigma^2\mathbf{I} + \mathbf{K} \end{aligned}$$

1.2 If we partition $\mathbf{y} = [y_1, \mathbf{y}']$ and denote $\mathbf{\Lambda} = \Sigma_{\mathbf{y}}^{-1}$. Identity (2.96) in Bishop (2006) [1] tells us $p(y_1|\mathbf{y}', \sigma, \mathbf{K}) = \mathcal{N}_{y_1}(\mu_{1|2:n}, \sigma^2_{1|2:n})$ where

$$\begin{aligned}\mu_{1|2:n} &= -\frac{1}{\Lambda_{1,1}}\mathbf{\Lambda}_{1,2:n}\mathbf{y}' \\ \sigma^2_{1|2:n} &= \frac{1}{\Lambda_{1,1}}\end{aligned}$$

1.3 We define $\boldsymbol{\phi}$ to be the feature transform generated from $\mathbf{K}$. If we train our model on $(y_i, \mathbf{x}_i)_{i=2:n}$, $K_{22} = \mathbf{\Phi}^T\mathbf{\Phi}$ and $\mathbf{K}_{21} = \boldsymbol{\phi}(\mathbf{x}_1)^T\mathbf{\Phi} = \mathbf{K}_{12}^T$ we can write the prediction of y_1 as

$$\begin{aligned}\hat{y}_1 &= \left\langle \mathbf{K}_{21}, \mathbf{K}_{22}^{-1}(\mathbf{y}')^T \right\rangle \\ &= \mathbf{K}_{12}\mathbf{K}_{22}^{-1}(\mathbf{y}')^T\end{aligned}$$

1.4 In contrast to least square estimation, here we have not assumed our data is independently distributed. This means our methods here are more applicable to real world data, as this assumption is often not satisfied.

Question 2

2.1 We can write the expression for $\text{var}[f(\mathbf{x}_i; \mathbf{w}_{LS})|\mathbf{x}_i]$ as

$$\begin{aligned}\text{var}[f(\mathbf{x}_i; \mathbf{w}_{LS})|\mathbf{x}_i] &= \boldsymbol{\phi}(\mathbf{x}_i)^T\text{cov}[(\boldsymbol{\phi}\boldsymbol{\phi}^T)^{-1}\boldsymbol{\phi}\mathbf{y}^T]\boldsymbol{\phi}(\mathbf{x}_i) \\ &= \boldsymbol{\phi}(\mathbf{x}_i)^T(\boldsymbol{\phi}\boldsymbol{\phi}^T)^{-1}\boldsymbol{\phi}\text{cov}[\mathbf{y}]\boldsymbol{\phi}^T(\boldsymbol{\phi}\boldsymbol{\phi}^T)^{-1}\boldsymbol{\phi}(\mathbf{x}_i) \\ &= \sigma^2\boldsymbol{\phi}(\mathbf{x}_i)^T(\boldsymbol{\phi}\boldsymbol{\phi}^T)^{-1}\boldsymbol{\phi}\boldsymbol{\phi}^T(\boldsymbol{\phi}\boldsymbol{\phi}^T)^{-1}\boldsymbol{\phi}(\mathbf{x}_i) \\ &= \sigma^2\boldsymbol{\phi}(\mathbf{x}_i)^T(\boldsymbol{\phi}\boldsymbol{\phi}^T)^{-1}\boldsymbol{\phi}(\mathbf{x}_i)\end{aligned}$$

Then we get

$$\begin{aligned}\frac{1}{n}\sum_{i\in D}\text{var}[f(\mathbf{x}_i; \mathbf{w}_{LS})|\mathbf{x}_i] &= \frac{\sigma^2}{n}\sum_{i\in D}\text{tr}(\boldsymbol{\phi}(\mathbf{x}_i)^T(\boldsymbol{\phi}\boldsymbol{\phi}^T)^{-1}\boldsymbol{\phi}(\mathbf{x}_i)) \\ &= \frac{\sigma^2}{n}\sum_{i\in D}\text{tr}\big((\boldsymbol{\phi}\boldsymbol{\phi}^T)^{-1}\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^T\big) \\ &= \frac{\sigma^2}{n}\text{tr}\left((\boldsymbol{\phi}\boldsymbol{\phi}^T)^{-1}\sum_{i\in D}\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^T\right) \\ &= \frac{\sigma^2}{n}\text{tr}\big((\boldsymbol{\phi}\boldsymbol{\phi}^T)^{-1}\boldsymbol{\phi}\boldsymbol{\phi}^T\big) \\ &= \frac{\sigma^2}{n}\text{tr}(\mathbf{I}_b) = \frac{\sigma^2 b}{n}\end{aligned}$$

and so we can see the average variance increases with b.

2.2 We can rewrite the in-sample error of f_{LS} as

$$\begin{aligned}\frac{1}{n}\sum_{i\in D}\mathbb{E}\big[(y_i - f_{LS}(\mathbf{x}_i))^2|\mathbf{x}_i\big] &= \frac{1}{n}\sum_{i\in D}\big[\text{var}[\varepsilon] + [\text{g}(\mathbf{x}_i) - \mathbb{E}[f_{LS}(\mathbf{x}_i)|\mathbf{x}_i]]^2 + \text{var}[f_{LS}|\mathbf{x}_i]\big] \\ &= \text{var}[\varepsilon] + \frac{1}{n}\sum_{i\in D}[\text{g}(\mathbf{x}_i) - \mathbb{E}[f_{LS}(\mathbf{x}_i)|\mathbf{x}_i]]^2 + \frac{1}{n}\sum_{i\in D}\text{var}[f_{LS}|\mathbf{x}_i]\end{aligned}$$

The first term is fixed and doesn't change with n. We have shown in the above question that the third term is equal to $\frac{\sigma^2 b}{n}$ and so decreases with larger n. Finally the least squares estimate f_{LS} is unbiased and so $\mathbb{E}[f_{LS}(\mathbf{x}_i)|\mathbf{x}_i] = g(\mathbf{x}_i)$, hence the second term is 0, no matter the value of n. Thus the in-sample error decreases when n increases.

2.3 Increasing n would improve the fit and reduce overfitting. This happens since as we gather more data, the distribution of our sample becomes more representative of the 'true distribution'.

The more b increases, the more flexible the prediction function becomes and adapts to our specific dataset. This makes the model less generalisable and overfit. That being said, a low values of b has the potential to underfit the prediction function, so it does not capture important features of the model.

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Berlin, Heidelberg: Springer-Verlag, 2006.