Homework 6

Due April 5 at 11 pm

Yves Greatti - yg390

1. (Gradient descent and ridge regression) In this problem we study the iterations of gradient descent applied to the ridge-regression cost function

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \left|\left| y - X^T \beta \right|\right|_2^2 + \frac{\lambda}{2} \left|\left| \beta \right|\right|_2^2 , \tag{1}$$

where $X \in \mathbb{R}^{p \times n}$ is a fixed feature matrix and $y \in \mathbb{R}^n$ is a response vector. (The factor of 1/2 is just there to make calculations a bit cleaner.)

(a) Derive a closed form expression for the value of the estimated coefficient $\beta^{(k)}$ at the kth iteration of gradient descent initialized at the origin in terms of the SVD of X when the step size is constant.

The gradient-descent updates are:

$$\begin{aligned}
\beta^{(k+1)} &= \beta^{(k)} - \ \alpha_k \nabla_\beta f(\beta^{(k)}) \\
&= \beta^{(k)} - \alpha_k (X X^T \beta^{(k)} - Xy + \lambda \beta^{(k)}) \\
&= ((1 - \lambda \alpha_k) I - \alpha_k X X^T) \beta^{(k)} + \alpha_k X y \\
&= \left((1 - \alpha\lambda) I - \alpha X X^T \right)^{k+1} \beta^{(0)} + \alpha \sum_{i=0}^{k} \left((1 - \alpha\lambda) I - \alpha X X^T \right)^i X y \\
&= \alpha \sum_{i=0}^{k} \left((1 - \alpha\lambda) I - \alpha X X^T \right)^i X y
\end{aligned}$$

since the step size $\alpha_k = \alpha$ is constant and $\beta^{(0)}$ is the zero vector (initialization at the origin). Let the svd of $X = USV^T$ then

$$\beta^{(k+1)} = \alpha \sum_{i=0}^{k} \left((1 - \alpha\lambda) I - \alpha U S^2 U^T \right)^i U S V^T y$$

Assuming $p \leq n$ and X is full rank, $UU^T = U^T U = I$ and we have:

$$\begin{aligned}
\beta^{(k+1)} &= \alpha \sum_{i=0}^{k} \left((1-\alpha\lambda)UU^T - \alpha US^2U^T\right)^i USV^T y \\
&= \alpha U \sum_{i=0}^{k} \left((1-\alpha\lambda)I - \alpha S^2\right)^i SV^T y \\
&= \alpha U \text{diag}_{j=1}^{p} \sum_{i=0}^{k} \left(1-\alpha(s_j^2+\lambda)\right)^i SV^T y \\
&= U \text{diag}_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^{k+1} s_j}{s_j^2+\lambda}\ V^T y
\end{aligned}$$

(b) Under what condition on the step size does gradient descent converge to the ridge-regression coefficient estimate as $k \to \infty$?
If step size α is small enough: $0 < \alpha < \frac{2}{\lambda + s_1^2} \leq \frac{2}{\lambda + s_j^2} \to |1 - \alpha(s_j^2+\lambda)| < 1$ then $\lim_{k\to\infty}(1-\alpha(s_j^2+\lambda))^k = 0, j = 1, \ldots, p$, gradient descent converges to:

$$\begin{aligned}
\lim_{k\to\infty} \beta^{(k)} &= U\ \text{diag}_{j=1}^{p}\left(\frac{s_j}{s_j^2+\lambda}\right) V^T y \\
&= U(S^2+\lambda I)^{-1} SV^T y
\end{aligned}$$

which are the ridge-regression coefficient estimates.

(c) Assume the following additive model for the data:

$$\tilde{y}_{\text{train}} := X^T \beta_{\text{true}} + \tilde{z}_{\text{train}}, \tag{2}$$

where $\tilde{z}_{\text{train}}$ is modeled as an n-dimensional iid Gaussian vector with zero mean and variance σ^2. What is the distribution of the estimated coefficient $\tilde{\beta}^{(k)}$ at the kth iteration of gradient descent initialized at the origin?
Using the expression of the estimated coefficients from part a, we now have:

$$\begin{aligned}
\tilde{\beta}^{(k)} &= U\text{diag}_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^k s_j}{s_j^2+\lambda}\ V^T(X^T\beta_{\text{true}} + \tilde{z}_{\text{train}}) \\
&= U\text{diag}_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^k s_j}{s_j^2+\lambda}\ V^T(VSU^T\beta_{\text{true}} + \tilde{z}_{\text{train}}) \\
&= U\text{diag}_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^k s_j^2}{s_j^2+\lambda}\ U^T\beta_{\text{true}} + U\text{diag}_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^k s_j}{s_j^2+\lambda}\ V^T\tilde{z}_{\text{train}}
\end{aligned}$$

Using theorem 8,6 from the notes on PCA, then the estimated coefficient $\tilde{\beta}^{(k)}$ at the kth iteration of gradient descent initialized at the origin is a Gaussian random vector with mean:

$$\beta_{\text{GD}} = \sum_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^k s_j^2}{s_j^2+\lambda} \langle u_j, \beta_{\text{true}} \rangle u_j$$

and covariance matrix

$$\Sigma_{\text{GD}} = \sigma^2 U \text{diag}_{j=1}^{p} \frac{(1-(1-\alpha(s_j^2+\lambda))^k)^2 s_j^2}{(s_j^2+\lambda)^2} U^T$$

(d) Complete the script *RR_GD_landscape.py* in order to verify your answer to the previous question. Report the figures generated by the script.

The figures for the mean and covariance matrix derived in part b, are reported below. We can see as k increases ($k = 300$), the estimated coefficients have a distribution similar to the ones reported for the ridge-regression estimated coefficients for the same value of $\lambda = 0.05$ in the PCA notes (figure 16). For small value of k, we have a bias: the estimated coefficients are shifting towards the initial starting point of gradient descent, the origin. Since $\alpha = .2 < \frac{2}{\lambda+s_1^2} = \frac{2}{0.05+1} \approx 1.9$ so as the number of steps k increases, the variance increases and is proportional to $\sigma^2 \frac{s_i^2}{(s_i^2+\lambda)^2}$. The eigenvalues are 1 and 0.1 and $\frac{1^2}{(1^2+0.05)^2} = 0.036 > \frac{0.1^2}{(0.1^2+0.05)^2} = 0.0177$ so the variance increases mostly in the direction corresponding to the larger singular values.

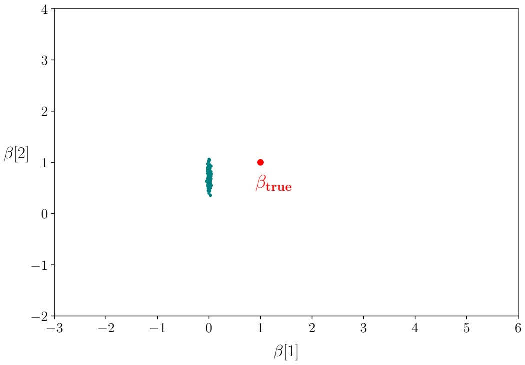

Figure 1: Scatterplot of the gradient-descent estimate corresponding to different noise realizations and iteration step $k = 3$.

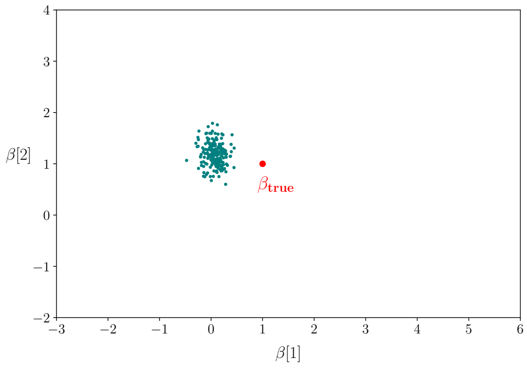

Figure 2: Scatterplot of the gradient-descent estimate corresponding to different noise realizations and iteration step $k = 50$.

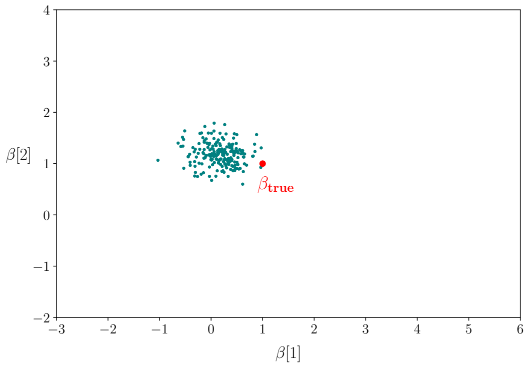

Figure 3: Scatterplot of the gradient-descent estimate corresponding to different noise realizations and iteration step $k = 500$.

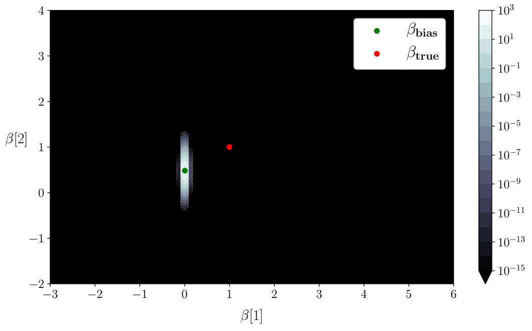

Figure 4: Heatmap of the distribution of the estimate following a Gaussian distribution with the mean and covariance matrix derived in part c, iteration step $k = 3$.

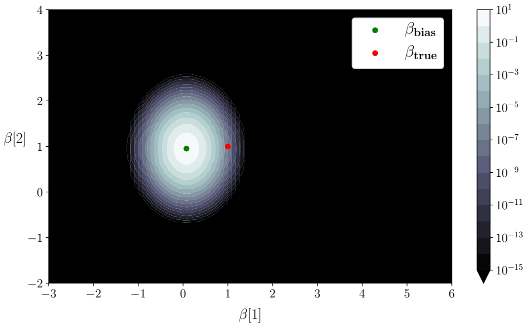

Figure 5: Heatmap of the distribution of the estimate following a Gaussian distribution with the mean and covariance matrix derived in part c, iteration step $k = 50$.

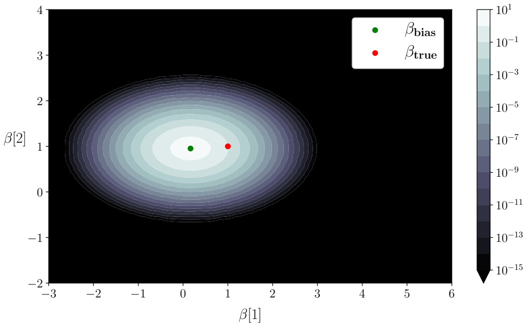

Figure 6: Heatmap of the distribution of the estimate following a Gaussian distribution with the mean and covariance matrix derived in part c, iteration step $k = 500$.

2. (Climate modeling) In this problem we model temperature trends using a linear regression model. The file `t_data.csv` contains the maximum temperature measured each month in Oxford from 1853-2014. We will use the first 150 years of data (the first $150 \cdot 12$ data points) as a training set, and the remaining 12 years as a test set.

   In order to fit the evolution of the temperature over the years, we fit the following model

$$y[t] = a + bt + c\cos(2\pi t/T) + d\sin(2\pi t/T) \tag{3}$$

   where $a, b, c, d \in \mathbb{R}$, $y[t]$ denotes the maximum temperature in Celsius during month t of the dataset (with t starting from 0 and ending at $162 \cdot 12 - 1$).

   (a) What is the number of parameters in your model and how many data points do you have to fit the model? Are you worried about overfitting?
   The number of parameters if 4 much less than the number of training samples $150 \cdot 12 = 1800$, so we have enough samples compared to the number of parameters and our linear regression model will not overfit as there are enough data points. Since we want to model the trend of the temperature over the years we have enough datapoints to detect the trend of the temperature. If we did not have enough points compared to the number of parameters our model will overfit, fitting mostly the noise.

   (b) Fit the model using least squares on the training set to find the coefficients for values of T equal to 1,2,...,20. Which of these models provides a better fit? Explain why this is the case. In the remaining question we will fix T to the value T^* that provides a better fit.

   When we plot the training errors we find out the model with the better fit is for $T = 12$. If we look at the different estimated coefficients, the cosine term is the larger compared to the sine term:

13.407	0	-7.62	-0.484

   The large absolute value on the cosine term implies that the phase of the cosine sinusoidal captures most of the cycle of the temperature oscillations. $T = 12$ corresponds also to the number of months in a year and is the period to which the temperatures fall in the same range from one year to the other for the corresponding month.

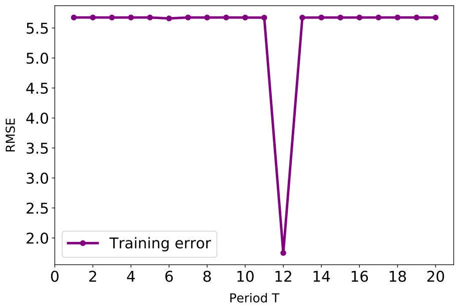

Figure 7: Training errors of the model using T equal to 1,2,...,20.

(c) Produce two plots comparing the actual maximum temperatures with the ones predicted by your model for $T := T^*$; one for the training set and one for the test set.

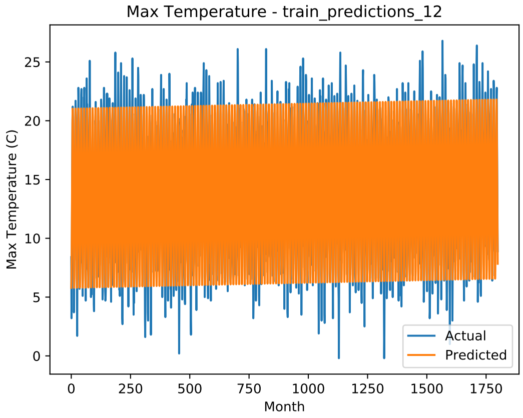

Figure 8: Actual and predicted max. temperatures from 1853-2014 for $T = 12$ on the training set.

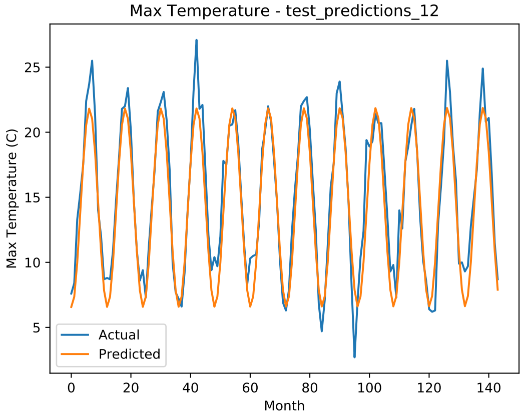

Figure 9: Actual and predicted max. temperatures from 1853-2014 for $T = 12$ on the test set.

(d) Fit the modified model

$$y[t] = a + bt + d\sin(2\pi t/T^*) \tag{4}$$

and plot the fit to the training data as in the previous question. Explain why it is better to also include a cosine term in the model.

Using the $T = 12$ the model estimate coefficients are about the same as the ones with the model in part b but without the cosine term:

13.394	0	-0.484

Not having the cosine term simplify the model but we are no longer capturing all the frequencies of the fundamental seasonal frequency of the temperatures. And this was confirmed by part b where, the estimate coefficient corresponding to the cosine term was much larger in comparison to the sine term. It is better to combine sine and cosine terms to span more frequencies. If we look at the prediction vs. actual, the predictions follow an almost flat trend since the oscillations corresponding to the sine term are very small being scaled by a small estimated coefficient.

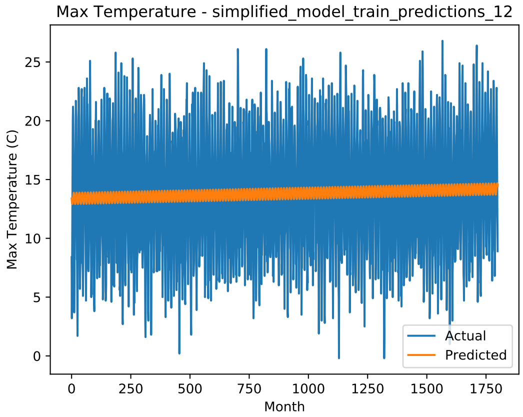

Figure 10: Actual and predicted max. temperatures from 1853-2014 for $T^* = 12$ on the training set, using the model $y[t] = a + bt + d\sin(2\pi t/T^*)$.

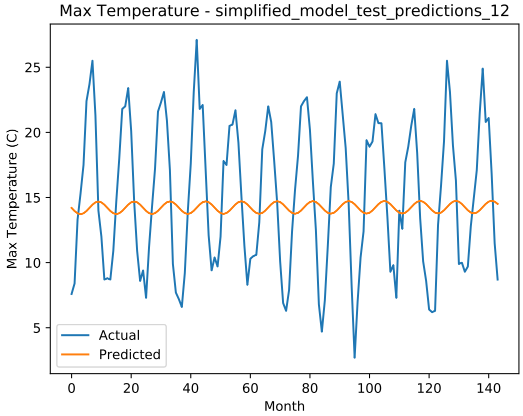

Figure 11: Actual and predicted max. temperatures from 1853-2014 for $T^* = 12$ on the test set, using the model $y[t] = a + bt + d\sin(2\pi t/T^*)$.

(e) Provide an intuitive interpretation of the coefficients a, b, c and d, and the corresponding features. According to your model, are temperatures rising in Oxford? By how much?

- a is the intercept or expected mean of the temperatures over the years

- b corresponds to the trend or drift of the temperatures
- c and d are related to the seasonal frequencies of the temperatures, they are included to capture the vertical or horizontal shift of these temperatures over time and their periodicity

If we look at the trend of the predicted temperatures on the training set (see fig. 8) or if we use a value of T which corresponds to a model less sensitive to the seasonality of the temperatures, like the trend exhibited by the model in part d (see fig. 10), we observe that the temperatures have been increasing in Oxford by 2-3 degrees for the last 100 years (exactly by 1.75 degree Celsius using the last 100 years of data).

3. (Sines and cosines) Let $x : [-1/2, 1/2) \to \mathbb{R}$ be a real-valued square-integrable function defined on the interval $[-1/2, 1/2)$, i.e. $x \in L_2[-1/2, 1/2)$. The Fourier series coefficients of x, are given by

$$\hat{x}[k] := \langle x, \phi_k \rangle = \int_{-1/2}^{1/2} x(t) \exp\left(-i 2\pi k t\right) \, \mathrm{d}t, \quad k \in \mathbb{Z}, \tag{5}$$

and the corresponding Fourier series of order k_c equals

$$\mathcal{F}_{k_c}\{x\}(t) = \sum_{k=-k_c}^{k_c} \hat{x}[k] \exp\left(i 2\pi k t\right). \tag{6}$$

As we will discuss in class, this is a representation of x in a basis of complex exponentials. In this problem we show that for real signals the Fourier series is equivalent to a representation in terms of cosine and sine functions.

(a) Prove that $\hat{x}[k] = \overline{\hat{x}[-k]}$ for all $k \in \mathbb{Z}$. [Hint: What is $\overline{e^{it}}$?]

$$\begin{aligned} \overline{\hat{x}[-k]} &= \overline{\int_{-1/2}^{1/2} x(t) \exp\left(i 2\pi k t\right) \, \mathrm{d}t} \\ &= \int_{-1/2}^{1/2} \overline{x(t) \exp\left(i 2\pi k t\right)} \, \mathrm{d}t \\ &= \int_{-1/2}^{1/2} x(t) \exp\left(-i 2\pi k t\right) \, \mathrm{d}t \\ &= \hat{x}[k] \end{aligned}$$

(b) Show that the Fourier series of x of order k_c can be written as

$$\mathcal{F}_{k_c}\{x\}(t) = a_0 + \sum_{k=1}^{k_c} a_k \cos(2\pi k t) + b_k \sin(2\pi k t),$$

for some $a_0, \ldots, a_k, b_1, \ldots, b_k \in \mathbb{R}$. [Hint: Group terms in $\mathcal{F}_{k_c}\{x\}(t)$ corresponding to $\pm k$ and use previous part. What is the real part of zw for $z, w \in \mathbb{C}$?]

$$\begin{aligned}
\mathcal{F}_{k_c}\{x\}(t) &= \sum_{k=-k_c}^{k_c} \hat{x}[k] \exp\left(i2\pi kt\right) \\
&= \hat{x}[0] + \sum_{k=-k_c}^{-1} \hat{x}[k] \exp\left(i2\pi kt\right) + \sum_{k=1}^{k_c} \hat{x}[k] \exp\left(i2\pi kt\right) \\
&= \hat{x}[0] + \sum_{k=k_c}^{1} \hat{x}[-k] \exp\left(-i2\pi kt\right) + \sum_{k=1}^{k_c} \hat{x}[k] \exp\left(i2\pi kt\right) \\
&= \hat{x}[0] + \sum_{k=1}^{k_c} (\hat{x}[-k] \exp\left(-i2\pi kt\right) + \hat{x}[k] \exp\left(i2\pi kt\right)) \\
&= \hat{x}[0] + \sum_{k=1}^{k_c} (\overline{\hat{x}[k]} \exp\left(-i2\pi kt\right) + \hat{x}[k] \exp\left(i2\pi kt\right)) \quad \text{using part a}
\end{aligned}$$

Given two complex numbers $z = a+ib$ and $w = c+id$, we have $zw = ac-bd+i(ad+bc)$ and $\overline{z}\ \overline{w} = ac - bd - i(ad + bc)$ giving that $zw + \overline{zw} = 2(ac - bd)$. Let $z_k = \hat{x}[k]$ and $w_k = \exp\left(i2\pi kt\right)$ thus

$$\begin{aligned}
\mathcal{F}_{k_c}\{x\}(t) &= \hat{x}[0] + \sum_{k=1}^{k_c} 2(\operatorname{Re}\left(\hat{x}[k]\right)\cos 2\pi kt - \operatorname{Im}\left(\hat{x}[k]\right)\sin 2\pi kt) \\
&= \hat{x}[0] + \sum_{k=1}^{k_c} (2\operatorname{Re}\left(\hat{x}[k]\right))\cos 2\pi kt + (-2\operatorname{Im}\left(\hat{x}[k]\right))\sin 2\pi kt \\
&= \hat{x}[0] + \sum_{k=1}^{k_c} a_k \cos(2\pi kt) + b_k \sin(2\pi kt)
\end{aligned}$$

(c) Give expressions for the coefficients a_k, b_k for $k \geq 1$ from the previous part as real integrals. Interpret them in terms of inner products.

From the definition

$$\begin{aligned}
\hat{x}[k] = \langle x, \phi_k \rangle &= \int_{-1/2}^{1/2} x(t) \exp\left(-i2\pi kt\right) \,\mathrm{d}t \text{ for } k \geq 1 \\
&= \int_{-1/2}^{1/2} x(t)(\cos(2\pi kt) + i\sin(2\pi kt)) \,\mathrm{d}t \\
&= \int_{-1/2}^{1/2} x(t)\cos(2\pi kt) \,\mathrm{d}t + i\int_{-1/2}^{1/2} x(t)\sin(2\pi kt) \,\mathrm{d}t \\
&= \operatorname{Re}\left(\hat{x}[k]\right) + i\operatorname{Im}\left(\hat{x}[k]\right)
\end{aligned}$$

thus

$$a_k = 2\int_{-1/2}^{1/2} x(t)\cos(2\pi kt)\,\mathrm{d}t = 2\left\langle x, \cos(2\pi kt)\right\rangle = 2\left\langle x, \mathrm{Re}\left(\phi_k\right)\right\rangle$$

and

$$b_k = 2\int_{-1/2}^{1/2} x(t)\sin(2\pi kt)\,\mathrm{d}t = 2\left\langle x, \sin(2\pi kt)\right\rangle = 2\left\langle x, \mathrm{Im}\left(\phi_k\right)\right\rangle$$

.

(d) Suppose $x(t) = \cos(2\pi(t+\phi))$ for some fixed $\phi \in \mathbb{R}$. What are the Fourier coefficients of x?

We can express the sinusoid $x(t)$ as:

$$\begin{aligned}\cos(2\pi(t+\phi) &= \frac{1}{2}[e^{i2\pi(t+\phi)} + e^{-i2\pi(t+\phi)}] \\ &= \frac{e^{-i2\pi\phi}}{2}e^{-i2\pi t} + \frac{e^{i2\pi\phi}}{2}e^{i2\pi t}\end{aligned}$$

So the Fourier coefficients of x are $\hat{x}[-1] = \frac{e^{-i2\pi\phi}}{2}$ and $\hat{x}[1] = \frac{e^{i2\pi\phi}}{2}$.

(e) Suppose that f is also even (i.e., $x(-t) = x(t)$). Prove that the Fourier coefficients are all real (i.e., that $\hat{x}[k] \in \mathbb{R}$ for all $k \in \mathbb{Z}$).

Using part a

$$\begin{aligned}\overline{\hat{x}[k]} &= \hat{x}[-k] \\ &= \int_{-1/2}^{1/2} x(t)\exp\left(i2\pi kt\right)\,\mathrm{d}t \\ &= \int_{1/2}^{-1/2} x(-u)\exp\left(-i2\pi ku\right)\left(-\,\mathrm{d}u\right) \text{ by change of variable } u = -t \\ &= \int_{-1/2}^{1/2} x(t)\exp\left(-i2\pi kt\right)\,\mathrm{d}t \quad \text{since x is even} \\ &= \hat{x}[k]\end{aligned}$$

$\overline{\hat{x}[k]} = \hat{x}[k]$, $\hat{x}[k] \in \mathbb{R}$ for all $k \in \mathbb{Z}$.