HW 1 – 4803, Spring 2023
Each problem is worth 10 points

1 Part I – theoretical problems

1. Find the least-squares solution $\vec{x}^*$ of the linear system

$$A\vec{x} = \vec{b}$$

where

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \text{ and } \vec{b} = \begin{bmatrix} 0 \\ 0 \\ 6 \end{bmatrix}.$$

What is a geometric relationship between $A\vec{x}^*$ and $\vec{b}$? Draw a picture to illustrate this.

2. Suppose you wish to fit a function of the form

$$f(t) = c + p \sin t + q \cos t$$

to a given continuous function $g(t)$ on the closed interval from 0 to 2π. One approach is to first choose $n+1$ equally spaced points a_i, $i = 0, 1, \ldots, n$ between 0 and 2π ($a_i = i\frac{2\pi}{n}$, say). We can then fit a function

$$f_n(t) = c_n + p_n \sin t + q_n \cos t$$

to the data points $(a_i, g(a_i))$ for $i = 0, \ldots, n$. An example of $n = 8$ is given in Figure 1.

In this problem, we will examine what happens to the coefficients c_n, p_n, q_n of $f_n(t)$ as $n \to +\infty$.

(a) For a fixed n, write down the linear system for this fitting where

$$A_n \begin{bmatrix} c_n \\ p_n \\ q_n \end{bmatrix} = \vec{y}$$

where A_n is an $(n+1) \times 3$ matrix, and $\vec{y}$ is a vector of length $n+1$.

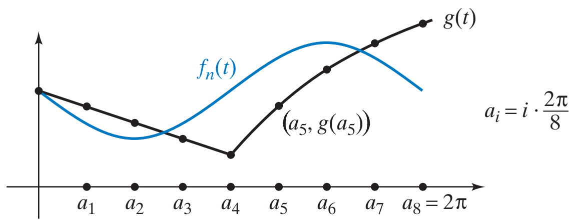

Figure 1:

(b) Find

$$\lim_{n \to +\infty} \frac{2\pi}{n} A_n^T A_n \text{ and } \lim_{n \to +\infty} \frac{2\pi}{n} A_n^T \vec{y}.$$

Hint: Interpret the entries of the matrix $\frac{2\pi}{n} A_n^T A_n$ and the components of the vector $\frac{2\pi}{n} A_n^T \vec{y}$ as Riemann sums. Then the limits are the corresponding Riemann integrals. Evaluate as many integrals as you can. Note that

$$\lim_{n \to +\infty} \frac{2\pi}{n} A_n^T A_n$$

is a diagonal matrix.

(c) Find

$$\begin{bmatrix} c_\infty \\ p_\infty \\ q_\infty \end{bmatrix} \equiv \lim_{n \to +\infty} \begin{bmatrix} c_n \\ p_n \\ q_n \end{bmatrix}.$$

The resulting vector $\begin{bmatrix} c_\infty \\ p_\infty \\ q_\infty \end{bmatrix}$ gives the fitting function

$$f_\infty(t) = c_\infty + p_\infty \sin t + q_\infty \cos t.$$

(d) For a given continuous function $g(t)$, is $f_\infty(t)$ necessarily equal to $g(t)$ on the interval $t \in [0, 2\pi]$? For what kind of $g(t)$ will we have $f_\infty(t) = g(t)$, $t \in [0, 2\pi]$.

3. From the Book "An Introduction to Statistical Learning"– 2.4 Exercise 7 (the Bayes decision boundary is defined in Pages 37-38).

4-7. From the Book "An Introduction to Statistical Learning"– 3.7 Exercises 3,4,5,6

2 Part II – programming

Programming Problem 1 This question involves the use of simple linear regression on the Auto data set, which can be downloaded at `https://www.statlearning.com/resources-second-edition`.

1. Read the Auto data into matlab, python, or R (or other programming language of your choosing).
2. Use least squares to perform a simple linear regression with "mpg"as the response and "horsepower"as the predictor.
3. Plot the response and the predictor, and display the least squares regression line.
4. Does the regression line fit the data well? What is the RSS?

Programming Problem 2 This question involves the use of multiple linear regression on the Auto data set.

1. Compute the matrix of correlations. (1) Produce a colorplot of the correlation matrix; (2) Display the matrix. You will need to exclude the "name"variable since it is qualitative and not useful here. Which variables are highly correlated? Here we say a pair of variables are highly correlated if their pairwise correlation is above 0.8 (in absolute value).
2. Use least squares to perform a multiple linear regression with "mpg"as the response and all other variables except "name "as the predictors. What are the coefficients and the RSS?
3. Compare the result of multiple linear regression with simple linear regression above. Which one is better?

Programming Problem 3 From the Book "An Introduction to Statistical Learning"– 3.7 Exercise 13 parts (a-i) If you use matlab or Python (or something else), the functions are different from the ones the book uses (which are for R).