Homework 4 (20 Points)

Due: 2023/11/01

Submission Guidlines

1. You need to write your homework in *Rmd/* ipynb/Rnotebook.
2. Please submit two files: one script file and one generated *pdf file.

In this assignment, you will explore spline regression with the secret data, you are **not** allowed to use built-in spline regression modules. Please implement the code by yourself.

1. For the dataset, please import "hw4_sample.txt". Each line contains exactly 2 numbers representing the X and Y coordinates of points generated by adding noise to a secret function.

   Implement the linear least square curve fitting (No regularization terms)

   Provide **plots** showing the points and the fitted curve for the dataset using the following basis functions.Note: You need to make three figures for the three different spline basis functions:

   - Linear splines: $B(x) = (1, x, [x+2]_+, [x]_+, [x-2]_+)$
   - Quadratic splines: $B(x) = \left(1, x, x^2, x^3, [x+2]_+^2, [x]_+^2, [x-2]_+^2\right)$
   - Cubic splines: $B(x) = \left(1, x, x^2, x^3, [x+2]_+^3, [x]_+^3, [x-2]_+^3\right)$

   The notation $[z]_+$ is a shorthand for $\max(0, z)$.

2. For the cubic splines regression model with $B(x) = \left(1, x, x^2, x^3, [x+2]_+^3, [x]_+^3, [x-2]_+^3\right)$, we introduce the regularization term, that is adding $\lambda\Sigma_{j=1}^3\beta_j^2$, that is we only regularize the coefficient on cubic splines: $[x+2]_+^3, [x]_+^3, [x-2]_+^3$. To find the best λ, you first **randomly** split your data into training set (80% of the data) and validation set (20% of the data).Then, try some different values of λs and visualize the traing error and validation error as a function of λ.

   For the λ you like most, Provide the **plot** showing the points and the fitted curve.

3. We now consider **natural splines** with k evenly spaced knots such that $r_1 = -4$ and $r_k = 4$ (i.e. $k = 1$ means we place one knot at x=0). You can ignore regularization for simplicity

   Use the same training set and validation set in Q2 to determine the best k for the dataset.

   Provide one plot showing the average of the training and validation MSE obtained during cross-validation as a function of $k = 1 \ldots 8$. Provide another plot showing the best fitted natrual spline and the dataset.