EEC 351 – Coding Assignment#1

Department of Electrical Engineering
Indian Institute of Technology Roorkee

Use of LLMs or AI tools for solving this assignment is strictly prohibited.

Instructions

This assignment explores bias, variance, and generalization in a controlled two-point regression learning setup. The boiler code is given to you for both questions. The names of the functions, the input arguments to the function, and the return datatype and variables must remain the same as given in the boiler code.

Important Note

The solutions you submit will be checked against multiple hidden test cases. **Do not hard-code answers**. Your code must work for **all valid inputs** (for any UID, N, s, k, A, etc.) according to the assignment specifications.

Boilerplate Code

You can find the Boilerplate codes for both Part 1 and Part 2 on course webpage.

Make sure to download and use the correct files corresponding to each part of the assignment. All filenames and function signatures must remain unchanged.

Submission Guidelines

1. Submit your solutions using the following Google Form link:
   `Google Form for Submission`
2. Upload your solutions as two **separate files**:
   - Part 1 → `EnrollmentNo_1.py`
   - Part 2 → `EnrollmentNo_2.py`
3. Example: If your Enrollment No is `2101234`, then filenames must be: `2101234_1.py` and `2101234_2.py`.
4. Any submission not following the naming convention may be rejected.
5. Late submission will result in **Zero** marks.

Personalization

1. Compute your personal seed s from your University Enrollment No:

$$s = \Big(\text{sum of ASCII codes of your UID}\Big) \bmod 10{,}000.$$

2. From s, compute parameters:

$$k = 1 + (s \bmod 7) \qquad (k \in \{1, \ldots, 7\}),$$

and

$$A = 1 + \Big(\big\lfloor s/7 \big\rfloor \bmod 3\Big) \times 0.5 \qquad (A \in \{1, 1.5, 2\}).$$

3. The input space is $X = [-A, A]$ and the target function is

$$f(x) = \sigma\Big((A + k)x\Big) + A \cdot \ln(|x| + 1) \cdot x + k \cdot \sin(\pi x) \cdot x^2,$$

where we denote the sigmoid by $\sigma(t) = \dfrac{1}{1 + e^{-t}}$

4. Training set $D = \{(x_1, y_1), (x_2, y_2)\}$ is obtained by sampling x_1, x_2 independently from $\text{Unif}([-A, A])$ and setting $y_i = f(x_i)$. If $x_1 = x_2$ (numerical tie within tolerance 10^{-12}), perturb x_2 by 10^{-6}.

5. The learning algorithm: choose the hypothesis in the given family that minimizes the in-sample mean squared error (two-point least squares). If multiple minimizers exist, pick the one with smallest ℓ_2 norm.

Part 1 (Moderate) — Affine Hypotheses $h(x) = ax + b$

(i) **Fit affine hypothesis** $h^*(x) = a^*x + b^*$**:** Derive the normal equations for a^*, b^* minimizing mean square loss. Write equations explicitly in terms of A, k. Evaluate numerically for your s for some random sampled data point.

(ii) **Expected learned hypothesis** $\bar{g}(x)$**:** Let $g_D(x) = \hat{a}_D x + \hat{b}_D$. Derive expressions and report values for $\mathbb{E}_D[\hat{a}_D], \mathbb{E}_D[\hat{b}_D]$ for randomly sampled dataset $N = 100$. (Hint: Use the normal equation derived in above question.)

(iii) **Bias-Variance decomposition:** Compute Bias–Variance for E_{out} as mean square error on same dataset $N = 100$.

(iv) Plots: $f(x)$ vs $\bar{g}(x)$ with variance shading.

Part 2 (Hard) — Quadratic hypotheses (batch setting)

Using a quadratic hypothesis to approximate a convex function:

$$h(x) = c_0 + c_1 x + c_2 x^2.$$

Due to constraints, you can only query $f(x)$ in batches of exactly 3 samples at a time. You are allowed to take 10 such batches (total $M = 30$ samples).
Tasks:

(i) **Batch Learning:**
For each batch (3 samples), fit $h_k(x)$ by least squares using the 3 points. Then compute the final averaged hypothesis:

$$\bar{h}(x) = \frac{1}{10} \sum_{k=1}^{10} h_k(x).$$

Implement the batch learning to return the averaged coefficients $\bar{c}_0, \bar{c}_1, \bar{c}_2$. Your model will be tested on a different function on only 30 samples, so make it optimized such that it gives best results in minimum samples.

(ii) **Hoeffding Sample Bound:**
Using Hoeffding's inequality, the sample complexity is given by

$$\Pr\left(\left|E(h) - \hat{E}(h)\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2M\epsilon^2}{B^2}\right),$$

where M is the number of samples, ϵ is the accuracy parameter, and B is an upper bound on the squared error $(f(x_i) - h(x_i))^2$. Required values of the variables are given in the boiler code.

Task: Compute the ratio

$$\frac{M_{\min}(B = 4)}{M_{\min}(B = 1)}.$$

and return the value of $M_{\min}(B = 1)$.