📌 Regression I – analiza teoretyczna

🔹 1. Problem regresji

W zadaniu mamy klasyczny problem regresji:

$$y = g(x) + \varepsilon$$

gdzie:

- $g(x)$ — nieznana funkcja regresji,
- ε — szum: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

🔹 2. Generowanie danych

- Funkcja testowa (benchmarkowa):

  $$g(x) = 4.26 \left(e^{-x} - 4e^{-2x} + 3e^{-3x}\right)$$

- Argumenty x_i są generowane z rozkładu jednostajnego na $[0, 4]$.
- Szum: $\varepsilon_i \sim \mathcal{N}(0, 0.1^2)$.

🔹 3. Metody estymacji

📌 (a) Nadaraya-Watson kernel regression

- Nadaraya-Watson to **nieparametryczny** estymator gęstości warunkowej:

$$\hat{g}_{NW}(x_0) = \frac{\sum_{i=1}^{n} K\left(\frac{x_0 - x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_0 - x_i}{h}\right)}$$

- $K(\cdot)$ — jądro (np. jądro Gaussa):

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-0.5u^2}$$

- h — szerokość pasma (bandwidth) — decyduje o gładkości estymatora.

📌 (b) Smoothing Splines

- Alternatywne podejście — rozwiązujemy problem optymalizacji:

$$\min_{f \in C^2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int [f''(t)]^2 dt$$

- Parametr λ kontroluje kompromis między gładkością a dopasowaniem:
  - Małe λ → ścisłe dopasowanie do danych.
  - Duże λ → bardziej gładka krzywa.

🔹 4. Dobór hiperparametrów

📌 **Dla Nadaraya-Watson:**

- Szerokość pasma h można dobrać np. przez cross-validation:
  - Podziel dane na 5 części (5-fold CV).
  - Oblicz błąd średniokwadratowy (MSE) dla różnych h.
  - Wybierz h z najmniejszym MSE.

📌 **Dla Smoothing Splines:**

- Najczęściej dobór λ przez np. RidgeCV lub Leave-One-Out CV.

🔹 5. Ewaluacja: Mean Squared Error (MSE)

Dla testowego zbioru:

$$MSE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} [g(x_i) - \hat{g}(x_i)]^2$$

gdzie:

- $g(x_i)$ — prawdziwa wartość funkcji regresji.
- $\hat{g}(x_i)$ — estymowana wartość funkcji regresji.

🔹 6. Eksperymenty

✅ Porównaj:

- Krzywe dopasowane przez Nadaraya-Watson i Smoothing Splines.
- Zbadaj wpływ rozmiaru próbki n na MSE:
  - Wygeneruj dane testowe niezależnie od danych treningowych.
  - Dla różnych n (np. 25, 50, 100, 200, 400, 800) oblicz MSE.

✅ Zrób wykres:

- Oś x: rozmiar próbki n (skala logarytmiczna).
- Oś y: MSE (skala logarytmiczna).