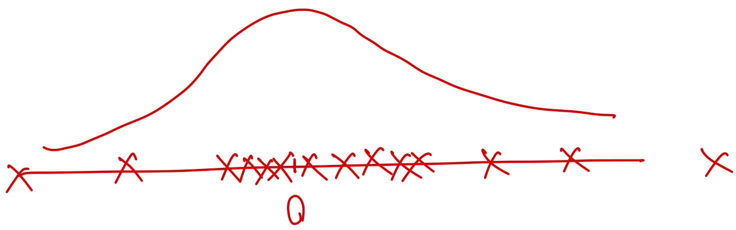
Problem 1: Ordinary Least Squares in Python

Assume the input has one dimension x and the target function is $f(x) = (x - 0.1)^3 - 5(x - 0.5)^2 + 10x + 5\sin 5x + 10$. Using ordinary least squares solution

(a) Find $a \in \mathbb{R}$ such that the hyperplane $\hat{y} = ax$ fits the data the best when x is distributed from a Gaussian with mean 0 and variance 1, $x \sim \mathcal{N}(0, 1)$.

(b) Find $a, b \in \mathbb{R}$ such that the hyperplane $\hat{y} = ax + b$ fits the data the best when x is distributed from a Gaussian with mean 0 and variance 1, $x \sim \mathcal{N}(0, 1)$.

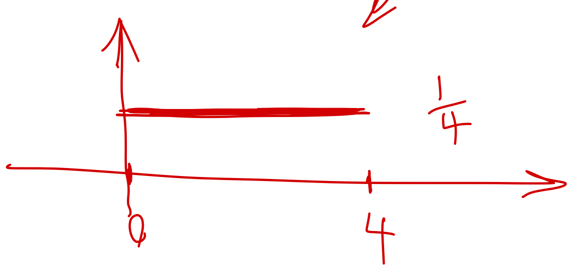


$$\frac{1}{n}\sum (ax^i + b - f(x^i))^2$$

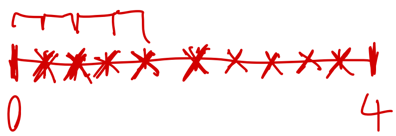
$$\frac{1}{n}\sum g(x^i) = E[g(X)] = \int_{-\infty}^{+\infty} g(x)\, f_X(x)\, dx$$

$n \to \infty$ $\quad x^i \sim U \quad X \sim U$

$$= \int_{-\infty}^{+\infty} \underbrace{(ax+b-y)^2}_{g(x)} f_X(x)\, dx$$



$$f_X(x) = \begin{cases} \frac{1}{4} & 0 \le x \le 4 \\ 0 & \text{oth} \end{cases}$$



$\sum \left(\frac{4i}{n}\right)^2 \qquad \sum \left(\frac{4i}{n}\right)^3$

$0 \qquad 4$

$\frac{4\times 1}{n}, \frac{4\times 2}{n}, \ldots, \frac{4\times n}{n}$

$\frac{4}{n} \ldots, 4$

Problem 2: Ordinary Least Squares

We will use least squares to find the best line $\hat{y} = ax+b$ that fits a non-linear function, namely $f(x) = x^2 - 3x^4 + 2$. For this, assume that you are given a set of n training points $\{(x^i, y^i)\}_{i=1}^n = \{((\frac{4i}{n}), (4i/n)^2 - 3(4i/n)^4 + 2)\}$. Find a line that fits the training data the best when $n \to \infty$.

$\left(\frac{4}{n}, \left(\frac{4}{n}\right)^2 - 3\left(\frac{4}{n}\right)^4 + 2\right)$

$\left(\frac{4\times 2}{n} \ldots\right)$

Solution. Writing ordinary least squares we have

$$\min_{a,b} \sum_{i=1}^{n} (ax^i + b - y^i)^2 = \min_{a,b} \sum_{i=1}^{n} \left(a(4i/n) + b - ((4i/n)^2 - 3(4i/n)^4 + 2)\right)^2. \tag{1}$$

In the case that $n \to \infty$, we can work with integral instead of summation. In this case, we know that the training samples come from a uniform distribution on $[0, 4]$, i.e., $x^i \sim U_{[0,4]}$ since there is an equal chance for any $x \in [0, 4]$ to be drawn. We know that for $X \sim U_{[0,4]}$, we have $f_X(x) = 1/4$ if $x \in [0, 4]$ and $f_X(x) = 0$ if $x \notin [0, 4]$.

Rewriting the OLS we have

$$\min_{a,b} \int_{x=0}^{x=4} (ax + b - f(x))^2 f_X(x)\, dx = \min_{a,b} \int_{x=0}^{x=4} (ax + b - (x^2 - 3x^4 + 2))^2 f_X(x)\, dx = \min_{a,b} g(a, b).$$

We can then expand the expression inside the integral and calculate the integral

$$\begin{aligned}
g(a, b) = &\int_{x=0}^{x=4} \frac{1}{4}\big[9x^8 - 6x^6 + 6ax^5 + (6b - 11)x^4 - 2ax^3 \\
&\qquad + (a^2 - 2b + 4)x^2 + (2ab - 4a)x + (b^2 - 4b + 4)\big]\, dx \\
=&\frac{1}{4}\big[x^9 - \frac{6}{7}x^7 + ax^6 + \frac{(6b - 11)}{5}x^5 - \frac{a}{2}x^4 + \frac{(a^2 - 2b + 4)}{3}x^3 + (ab - 2a)x^2 + (b^2 - 4b + 4)x\big]\Big|_0^4 \\
=& \ 61487.27 + 984a + \frac{4388}{15}b + 4ab + \frac{16}{3}a^2 + b^2.
\end{aligned}$$

To obtain the values of a and b that minimize the function $g(a, b)$, we can compute the partial derivatives of $g(a, b)$ with respect to a and b and find the pair (a, b) that result in zero derivatives.

$$\begin{aligned}
\frac{\partial g(a, b)}{\partial a} &= \frac{32}{3}a + 4b + 984 = 0 \\
\frac{\partial g(a, b)}{\partial b} &= 2b + 4a + \frac{4388}{15}. = 0
\end{aligned} \tag{2}$$

Setting the derivatives equal to 0 in equation (2), we need to solve the following system of linear equations

$$\begin{aligned}
\frac{32}{3}a + 4b &= -984 \\
2b + 4a &= -\frac{4388}{15}.
\end{aligned}$$

Solving the system of linear equation we conclude that $a = \frac{-748}{5}$ and $b = \frac{2294}{15}$ are the solutions of OLS and $\hat{y} = -149.6x + 152.93$ is the best line that fits $f(x) = x^2 - 3x^4 + 2$ using OLS.

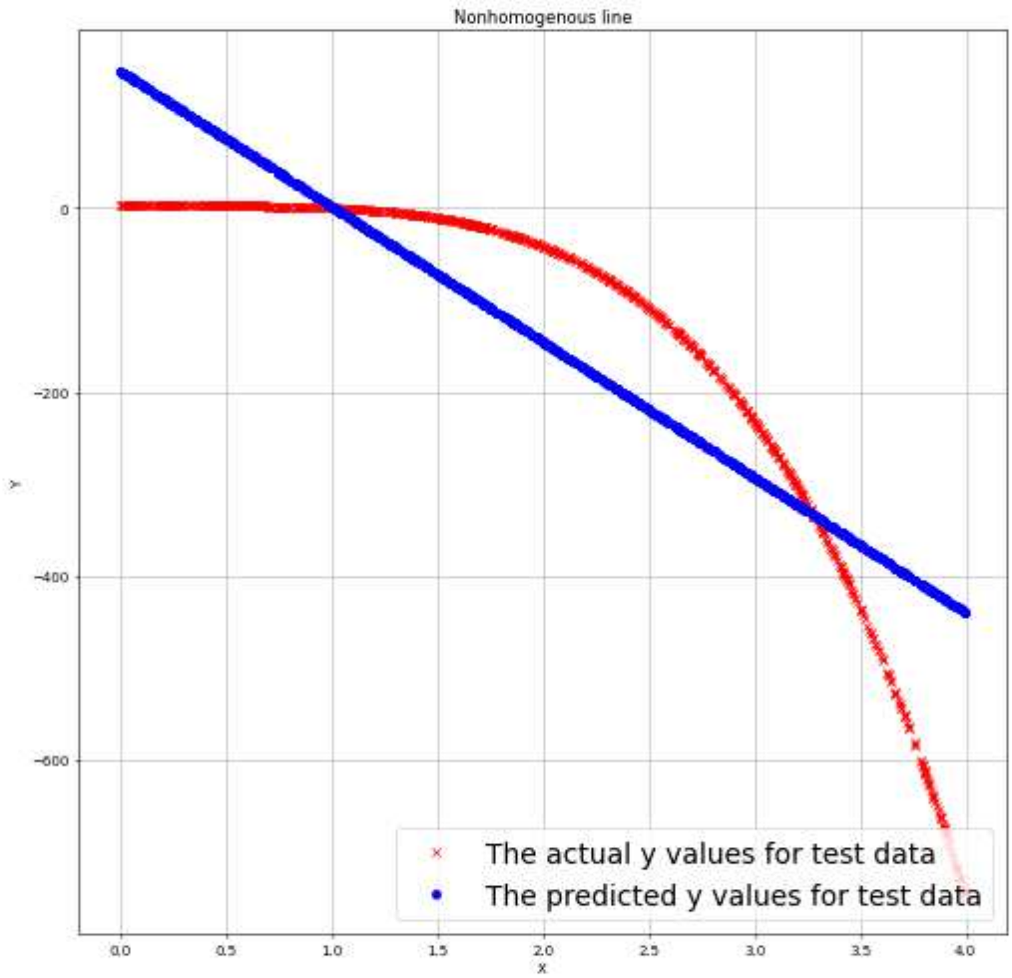


Figure 1: Optimal values of W found by simulating OLS for $n = 1000$ training samples. Optimal $W = [-147.18, 148.56]$.

$n \to \infty$