# Wet assignment

Alexander Shender 328626114
Eliran Cohen 204187801

# Part 0.

## Question 1.

For a single measurement we have:

$$y_i = f(x_i) + \omega_i \ ; \ \ i = 1, \dots, m$$

Where:

$$f(x_i) = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_n x_i^n$$

Inserting $f(x_i)$ into $y_i$:

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_n x_i^n + \omega_i \ ; \ \ i = 1, \dots, m$$

In a vector multiplication form:

$$y_i = [1 \quad x_i \quad \dots \quad x_i^n][a_0 \quad a_1 \quad \dots \quad a_n]^T + \omega_i$$

And for the $m$ different measurements we can write in a vector form:

$$\boldsymbol{y}_{\boldsymbol{m}\times\boldsymbol{1}} = \underbrace{\begin{bmatrix} 1 & x_1 & \dots & x_1^n \\ 1 & x_2 & \dots & x_2^n \\ \dots & \dots & \dots & \dots \\ 1 & x_m & \dots & x_m^n \end{bmatrix}}_{\boldsymbol{X}} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_n \end{bmatrix}}_{\boldsymbol{a}}$$

Where $\boldsymbol{y}_{\boldsymbol{m}\times\boldsymbol{1}}$ contains the measurements (which include the noise).

By aiming to minimize the following problem:

$$(P) \min_{a} \left\{ h(\boldsymbol{a}) = \frac{1}{2m} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2 \right\}$$

We try to find the coefficients $\boldsymbol{a}$ which can 'describe' the measurements in a best way.

## Question 2.

Convexity:

- The problem is defined over the whole range $R^n$ , without constraints. This range is a convex set
- The objective function is a squared Euclidean distance, which is a smooth convex function.

To find the minimum analytically, we need to make derivative w.r.t $\boldsymbol{a}$.

$$\frac{\partial h(a)}{\partial a} = \frac{\partial}{\partial a}\left(\frac{1}{2m}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2\right) = \frac{1}{2m}\left(-2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})\right) = \boldsymbol{0}$$

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}) = \boldsymbol{0}$$

$$\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{a} = \boldsymbol{0}$$

$$\boldsymbol{X}^T\boldsymbol{y} = \underbrace{\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}}_{\boldsymbol{n}\times\boldsymbol{n}}\boldsymbol{a}$$

$\boldsymbol{X}$ is of rank $n + 1$, thus $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is a full rank matrix. So the inverse exists, and we can write:

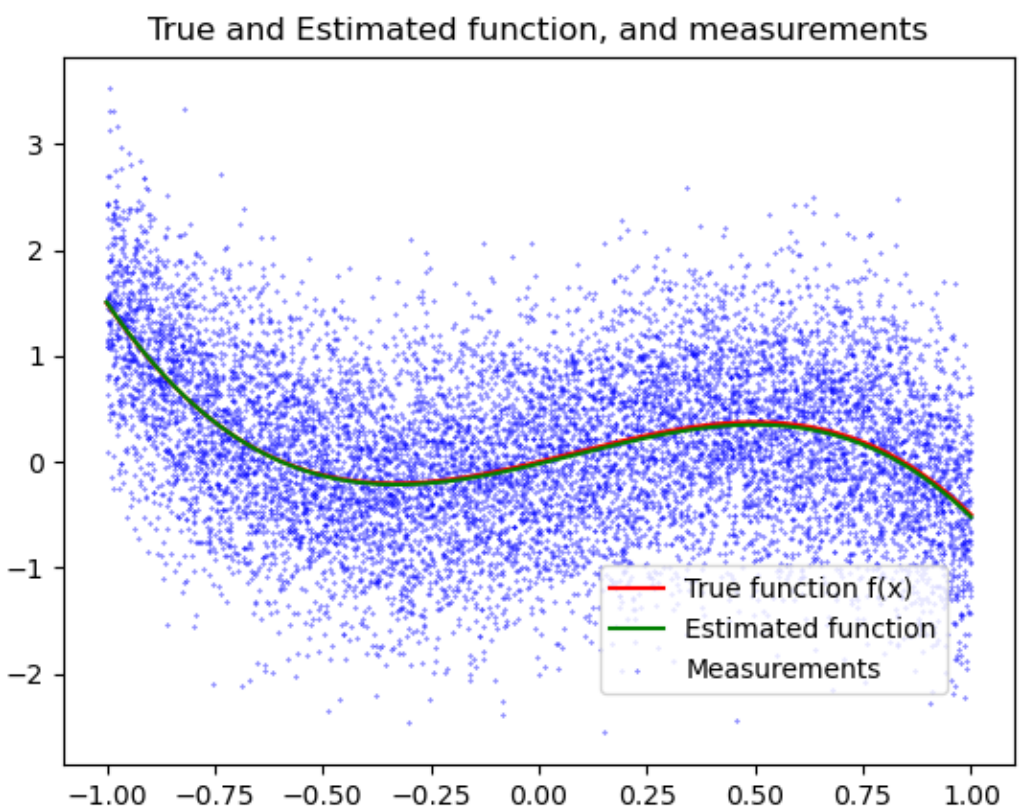
$$\boldsymbol{a} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-\boldsymbol{1}}\boldsymbol{X}^T\boldsymbol{y}$$

Which is the analytical solution.

# Part 1.

## Question 3+4.

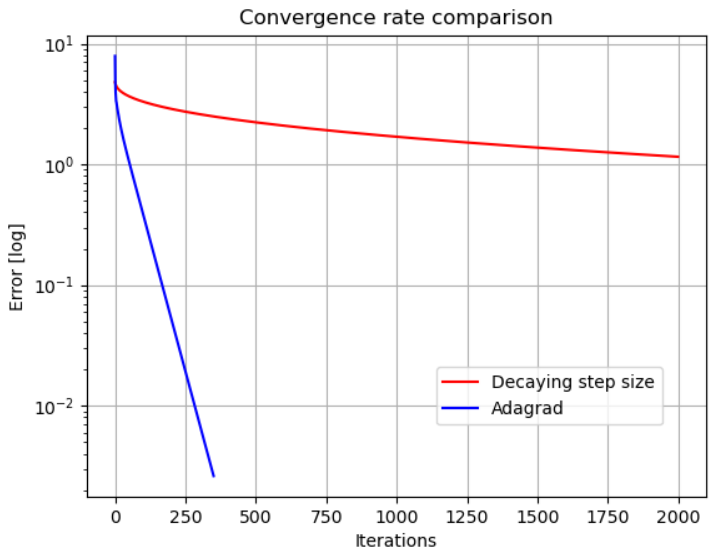
The function with the desired parameters was created, and the corresponding noise was added. The function was written which has estimated the $\boldsymbol{a}$ vector, according to the analytical solution derived in Question 2. The graph is the following:




We can indeed see that the Estimated function is not completely equal to the True original function. But this function minimizes the objective function that was defined.

## Question 5+6.

The function was written that  performs the Projected Gradient Descent algorithm with 2 possible step size calculations as described in the HW assignment. The graphs obtained are the following:




The additional operation of taking absolute value was added to the Error function.

We can indeed see that the AdaGrad step size converges much faster and gives a much smaller error over smaller number of iterations. The reason is that it takes into account all the previous gradients, and adapts the step size accordingly.

## Question 7.

We shall find the smoothness parameter L of the objective function.

From the definition of smoothness:

$$\|\nabla h(c) - \nabla h(b)\| \leq L \, \|\mathrm{c} - \mathrm{b}\|$$

$$\frac{\|\nabla h(c) - \nabla h(b)\|}{\|\mathrm{c} - \mathrm{b}\|} \leq L$$

Reminder:

$$h(\boldsymbol{a}) = \frac{1}{2m} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2$$

$$\nabla h(\boldsymbol{a}) = -\frac{1}{m} \boldsymbol{X}^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})$$

Inserting:

$$\frac{\left\| -\frac{1}{m} \boldsymbol{X}^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c}) + \frac{1}{m} \boldsymbol{X}^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \right\|}{\|\mathrm{c} - \mathrm{b}\|} \leq L$$

$$\frac{\left\| -\frac{1}{m} \left( \boldsymbol{X}^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{c}) - \boldsymbol{X}^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \right) \right\|}{\|\mathrm{c} - \mathrm{b}\|} \leq L$$

$$\frac{\left\| -\frac{1}{m} \left( \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y} - \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X}\boldsymbol{c} - \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y} + \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X}\boldsymbol{b} \right) \right\|}{\|\mathrm{c} - \mathrm{b}\|} \leq L$$

$$\frac{\left\| -\frac{1}{m} \left( -\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X}\boldsymbol{c} + \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X}\boldsymbol{b} \right) \right\|}{\|\mathrm{c} - \mathrm{b}\|} \leq L$$

$$\frac{\left\| -\frac{1}{m} \left( -\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X} (\boldsymbol{c} - \boldsymbol{b}) \right) \right\|}{\|\mathrm{c} - \mathrm{b}\|} \leq L$$

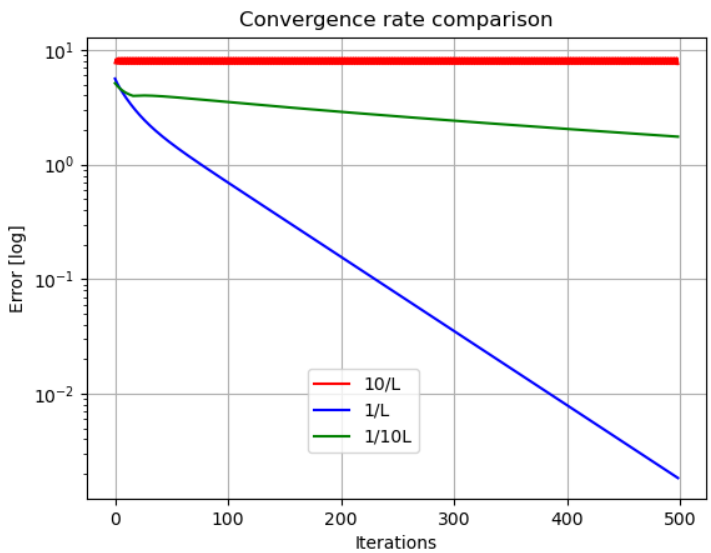
$$\frac{\left\| \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X} \right\|}{m} \frac{\|\boldsymbol{c} - \boldsymbol{b}\|}{\|\mathrm{c} - \mathrm{b}\|} \leq L$$

$$\frac{\left\| \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X} \right\|}{m} \leq L$$

**To be continued…**

## Question 8.

The following graph was obtained:



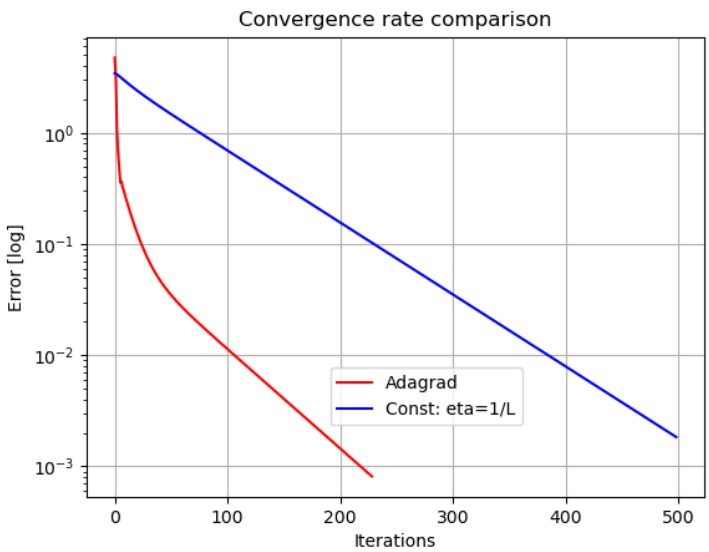

We can see that for the constant step size of 1/L we have the best convergence, indeed as was proved in the lectures and tutorials.

For step size of 10/L the algorithm “jumps over” the minimum point. The ‘update’ step is moving in the direction of the negative gradient, but the step is too big.

For 1/10L the algorithm would also converge, but it will take longer, as we may observe.

## Question 9.

The graph obtained is the following:




**FILL EXPLANATION**