Question 3

Consider the following function:

$f(x_1, x_2, x_3) = -2x_1^2 + 4x_1 + x_2x_1 - x_2 + x_3x_2 - x_3^2$

a) Compute ∇f

Solution

Recall, for a function $f: \mathbb{R}^n \to \mathbb{R}$, $x \mapsto f(x)$, $x \in \mathbb{R}^n$ of n variables $x_1, \dots, x_n$ we define the Partial derivatives as:

$$\frac{\partial f}{\partial x_1} = \lim_{h \to 0} \frac{f(x_1 + h, x_2, \dots, x_n) - f(x)}{h}$$

$$\vdots$$

$$\frac{\partial f}{\partial x_n} = \lim_{h \to 0} \frac{f(x_1, \dots, x_{n-1}, x_n + h) - f(x)}{h}$$

and collect them in the row vector

$$\nabla f = \text{grad } f = \frac{df}{dx} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} & \frac{\partial f(x)}{\partial x_2} & \dots & \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1 \times n} \quad \dots (A)$$

where n is the number of variables and 1 is the dimension of the image/range/codomain of f. Here we defined the column vector $x = [x_1, \dots, x_n]^T \in \mathbb{R}^n$.

The row vector in (A) is called the gradient of f or the jacobian

* Therefore, given $f(x_1, x_2, x_3) = -2x_1^2 + 4x_1 + x_2x_1 - x_2 + x_3x_2 - x_3^2$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \frac{\partial f}{\partial x_3} \end{bmatrix} \in \mathbb{R}^{1 \times 3}$$

- $\frac{\partial f}{\partial x_1} = -4x_1 + 4 + x_2$
- $\frac{\partial f}{\partial x_2} = x_1 - 1 + x_3$
- $\frac{\partial f}{\partial x_3} = x_2 - 2x_3$

$\therefore \nabla f = \begin{bmatrix} -4x_1 + 4 + x_2, & x_1 - 1 + x_3, & x_2 - 2x_3 \end{bmatrix}$

b) Find the Solutions for $\nabla f(x) = 0$

Solution

Recall $\nabla f = [-4x_1 + 4 + x_2,\ x_1 - 1 + x_3,\ x_2 - 2x_3]$

At $\nabla f = 0$;

$[-4x_1 + 4 + x_2,\ x_1 - 1 + x_3,\ x_2 - 2x_3] = [0, 0, 0]$

i.e $-4x_1 + 4 + x_2 = 0 \implies 4x_1 - x_2 = 4 \quad \ldots (1)$

$x_1 - 1 + x_3 = 0 \implies x_1 + x_3 = 1 \quad \ldots (2)$

$x_2 - 2x_3 = 0 \implies x_2 = 2x_3 \quad \ldots (3)$

Recall from (2) : $x_1 = 1 - x_3 \quad \ldots (4)$

Sub for (4) and (3) in eqn ... (1)

$\implies 4(1 - x_3) - 2x_3 = 4$

i.e $6x_3 = 0 \implies x_3 = 0,$

Sub for x_3 in (4): $x_1 = 1 - 0 = 1$

Sub for x_3 in (3): $x_2 = 2x_3 = 2(0) = 0$

Therefore; $(x_1, x_2, x_3) = (1, 0, 0)$.

1c

Starting at a value $\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$, and consider learning rates $\{1, \frac{1}{2}\}$

Execute two steps of the gradient descent algorithm for each learning rate separately

Solution

* Gradient descent algorithm :

repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

(for $j = 1$ and $j = 0$)

}

- Alpha is called the Learning rate - A tuning parameter in the optimization process. It decides the length of the steps.

Hence, we have:

$$x_i := x_i - \alpha \frac{\partial}{\partial x_i} f(x_1, x_2, x_3) \quad \text{for } i = 1, 2, 3.$$

where α = learning rate

$\Rightarrow$ Starting from Point $[1, 2, 0]$, updating the values of x_1, x_2 and x_3 according to the gradient rule.

☐ STEP 1:
Considering learning rate (α) OF 1 [Recall the learning rates are $\{1, \frac{1}{2}\}$]

* $x_1 = x_1 - \alpha \frac{\partial f}{\partial x_1} = 1 - 1[-4(1) + 4 + 2] = -1$
* $x_2 = x_2 - \alpha \frac{\partial f}{\partial x_2} = 2 - 1[1 - 1 + 0] = 2$
* $x_3 = x_3 - \alpha \frac{\partial f}{\partial x_3} = 0 - 1[2 - 2(0)] = -2$

The new Point is $[-1, 2, -2]$

☐ STEP 2

* $x_1 = x_1 - \alpha \frac{\partial f}{\partial x_1} = -1 - 1[-4(-1) + 4 + 2] = -11$
* $x_2 = x_2 - \alpha \frac{\partial f}{\partial x_2} = 2 - 1[-1 + 1 + (-2)] = 6$
* $x_3 = x_3 - \alpha \frac{\partial f}{\partial x_3} = -2 - 1[2 - 2(-1)] = -8$

The final Position of learning rate 1 is: $(-11, 6, -8)$.

Using learning rate $\frac{1}{2}$ [starting Point $[1, 2, 0]$]

STEP 2:

* $x_1 = x_1 - \alpha \frac{\partial f}{\partial x_1} = 1 - \frac{1}{2}[2] = 0$
* $x_2 = x_2 - \alpha \frac{\partial f}{\partial x_2} = 2 - \frac{1}{2}[0] = 2$
* $x_3 = x_3 - \alpha \frac{\partial f}{\partial x_3} = 0 - \frac{1}{2}[0] = 0$

New Point is $[0, 2, 0]$

☐ STEP 2

* $x_1 = x_1 - \alpha \frac{\partial f}{\partial x_1} = 0 - \frac{1}{2}[6] = -3$
* $x_2 = x_2 - \alpha \frac{\partial f}{\partial x_2} = 2 - \frac{1}{2}[-1] = 5/2$
* $x_3 = x_3 - \alpha \frac{\partial f}{\partial x_3} = 0 - \frac{1}{2}[2] = -1$

The final Position of learning rate 2 is: $(-3, 5/2, -1)$.

ii) Discuss the Outcome and Compare the Output with from (b) whether they are distances from points in (b) or getting closer?

Solution

Recall, the learning rate is a hyperparameter that controls the step size of each iteration while moving toward a minimum of a loss function.

A smaller learning rate may lead to slower convergence but can avoid overshooting the minimum, while a larger learning rate may lead to faster convergence but may overshoot the minimum.

* Now let's consider the given final positions of the gradient descent algorithm with two different learning rates:

Learning rate 1 : $[-11, 6, -8]$

Learning rate 2 : $[-3, 5/2, -1]$

- For the learning rate 1, the final position is quite far from the point $[1, 0, 0]$ indicating that the algorithm overshot the minimum. It may have taken larger steps at each iteration, causing it to miss the minimum.

$$d = \sqrt{(1-(-11))^2 + (0-6)^2 + (0-(-8))^2} = 15.62$$ [distance between two points $[1,0,0]$ and $[-11, 6, -8]$]

- For the learning rate $1/2$, the final position is closer to the point $[1, 0, 0]$. This result suggests that the algorithm may have taken smaller steps at each iteration, leading to slower convergence but a more accurate result.

$$d = \sqrt{(1-(-3))^2 + (0-5/2)^2 + (0-(-1))^2} = 4.82$$ [distance between two points $[1,0,0]$ and $[-3, 5/2, -1]$]

- Overall, we can say the algorithm with a learning rate of $1/2$ is better than the algorithm with a learning rate of 1, as it gets closer to the desired point.

- Note: Choosing an appropriate learning rate and a good starting point are crucial for the success of the gradient descent algorithm.