Mathematical Foundations

BITS Pilani
Pilani Campus

MFDS Team

innovate achieve lead

BITS Pilani

Pilani Campus

Mathematical Foundations

Webinar 3

Agenda

1. Maxima and Minima
2. Gradient of a Vector-Valued Function
3. Gradient Descent
4. Gradient Descent Examples
5. Previous Year Problems

1. Maxima and Minima

- For a given function $f(x, y)$,
- If $f_{xx} < 0$ and $f_{xx}f_{yy} - f^2_{xy} > 0$ at (a,b), then f has a local maximum value at (a,b).
- If $f_{xx} > 0$ and $f_{xx}f_{yy} - f^2_{xy} > 0$ at (a,b), then f has a local minimum value at (a,b).
- If $f_{xx}f_{yy} - f^2_{xy} < 0$ at (a,b), then f has a saddle point at (a,b).
- If $f_{xx}f_{yy} - f^2_{xy} = 0$ at (a,b), then the test is inconclusive and further investigation is required.

Hessian matrix

- The considerations of the previous slide show that determining whether there is a minimum or maximum at the point (*a, b*) boils down to looking at the following matrix and asking if it is positive-definite or not.

- This is the Hessian matrix is $\begin{bmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{bmatrix}$

$$H = \begin{vmatrix} r & s \\ s & t \end{vmatrix} = rt - s^2$$

Note : The discriminant $f_{xx}f_{yy} - f_{xy}^2$ ***is noting but the determinant of the Hessian matrix.***

Example

1. Find the local extreme values of the function $f(x, y) = xy - x^2 - y^2 - 2x - 2y + 4.$

Solution

The function is defined and differentiable for all x and y and its domain has no boundary points. The function therefore has extreme values only at the points where f_x and f_y are simultaneously zero. This leads to

$$f_x = y - 2x - 2 = 0, \qquad f_y = x - 2y - 2 = 0, \qquad \text{or} \qquad x = y = -2.$$

Therefore, the point $(-2, -2)$ is the only point where f may take on an extreme value. To see if it does so, we calculate

$$f_{xx} = -2, \qquad f_{yy} = -2, \qquad f_{xy} = 1.$$

The discriminant of f at $(a, b) = (-2, -2)$ is $f_{xx}f_{yy} - f_{xy}^2 = (-2)(-2) - (1)^2 = 4 - 1 = 3.$

The combination $f_{xx} < 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ tells us that f has a local maximum at $(-2, -2)$

The value of f at this point is $f(-2, -2) = 8$

Thus, the function has a local maximum value of 8 at the point $(-2, -2)$, and there are no other local extrema.

Example

2. Find the local extreme values of $f(x, y) = xy$.

Solution Since f is differentiable everywhere (Figure 14.43), it can assume extreme values only where

$$f_x = y = 0 \quad \text{and} \quad f_y = x = 0.$$

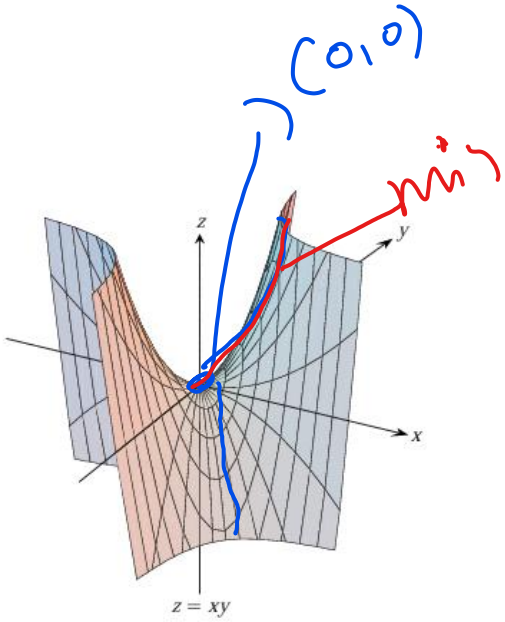
Thus, the origin is the only point where f might have an extreme value. To see what happens there, we calculate

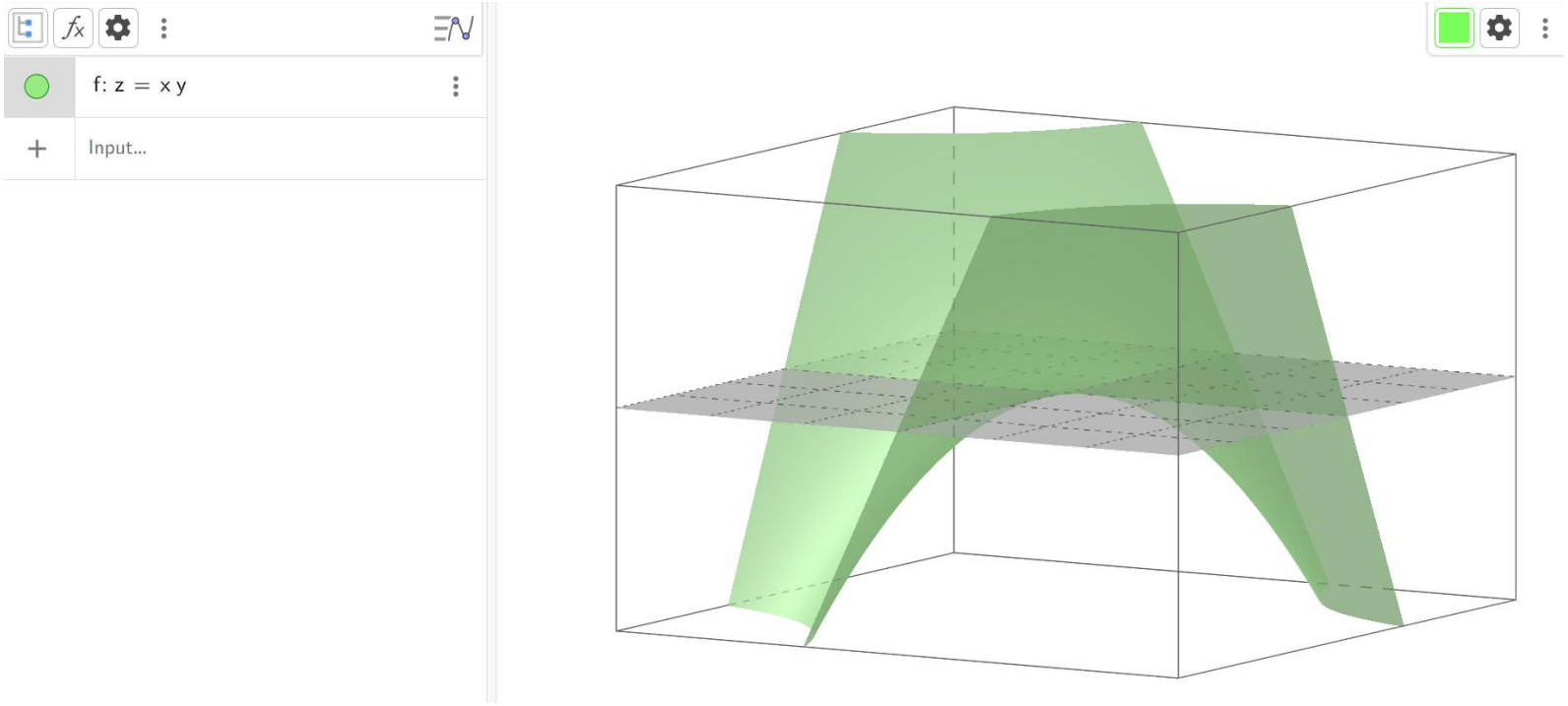
$$f_{xx} = 0, \quad f_{yy} = 0, \quad f_{xy} = 1.$$

The discriminant,

$$f_{xx}f_{yy} - f_{xy}^2 = -1,$$

is negative. Therefore, the function has a saddle point at (0, 0). We conclude that $f(x, y) = xy$ has no local extreme values.




f: z = x y
Input...

Maxima and Minima of a function of three variables using the Hessian.

Consider a function f (x, y, z) of three variables.

Step 1: Critical Points

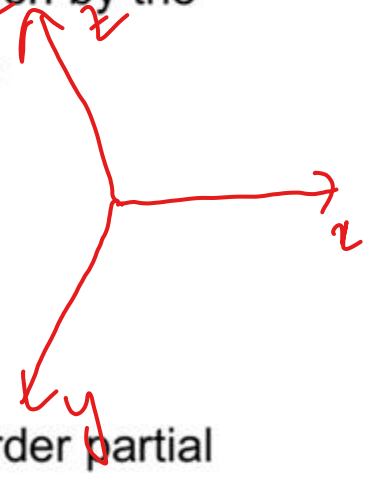
To find critical points, we need to solve the system of equations given by the gradient of f and is given by $\nabla f = \left[\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \quad \frac{\partial f}{\partial z}\right]^T = \left[0 \quad 0 \quad 0\right]^T$

The solutions to this system represent critical points

Step 2: Hessian Matrix

The Hessian matrix, denoted as H, is a square matrix of second-order partial derivatives of f and is given by

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{bmatrix}$$



$f_x = 0$

$f_y = 0$

$f_z = 0$

Step 3: Second Derivative Test

The determinant of the Hessian matrix, det(H), is used to determine the nature of critical points:

Let x_0 be a critical point of f.

i) If the Hessian matrix H is positive definite (i.e. all eigenvalues are strictly positive) then $\delta^T H(x_0)\delta > 0$. So $f(x) > f(x_0)$ and hence x_0 is a point of local minima.

ii) If the Hessian matrix H is negative definite (i.e. all eigenvalues are strictly negative) then $\delta^T H(x_0)\delta < 0$. So, $f(x) < f(x_0)$ and hence x_0 is a point of local maxima.

iii) If the Hessian matrix H has both positive and negative eigenvalues then x_0 is a saddle point.

iv) If the determinant of Hessian matrix H is zero then the test is inconclusive.

This method helps classify critical points as maxima, minima, or saddle points in functions of three variables

Example

3. Discuss the maximum and minimum value of the function $f(x,y,z)=x^2+y^2+z^2+x-2z-xy$.

Solution: $f(x,y,z)=x^2+y^2+z^2+x-2z-xy.$

$$\nabla_{x,y,z} f(x,y,z)=[f_x \quad f_y \quad f_z]=[2x+1-y \quad 2y-x \quad 2z-2]$$

$$f_x=2x+1-y\ ,\ f_y=2y-x,\quad f_z=2z-2$$

To find critical points, we need to solve the system of equations given by the gradient of f is equal to zero. i.e

$$\nabla f=\left[\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \quad \frac{\partial f}{\partial z}\right]^T=[0 \quad 0 \quad 0]^T$$

$$2x+1-y=0 \quad \Rightarrow y=2x+1 \qquad (1)$$

$$2y-x=0 \quad \Rightarrow x=2y \qquad (2)$$

$$2z-2=0 \quad \Rightarrow z=1$$

Substituting (2) in (1) gives y= -1/3, x=-2/3 and z=1

$$(x,y,z)=\left(-\frac{1}{3},-\frac{2}{3},1\right)$$

$f_x = 2x + 1 - y$, $f_y = 2y - x$, $f_z = 2z - 2$

$f_{xx} = 2$, $f_{yy} = 2$, $f_{zz} = 2$

$f_{xy} = -1$, $f_{yz} = 0$

The only critical point is (-2/3,-1/3,1) where the function will be either minima or maxima.

The Hessian matrix $H = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{bmatrix} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

$2 \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix}$

Further, The determinant of Hessian H is , $\det(H) = \begin{vmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 2 \end{vmatrix} = 6 > 0$

$= 2(4-1)$

$= 2 \times 3$

$= 6$

The determinants of all the leading principal minors of H are given by .

$M_1 = 2 > 0; \quad M_2 = \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} = 3 > 0; \quad M_2 = \begin{vmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 2 \end{vmatrix} = 6 > 0$

If det(H) > 0 and all the leading principal minors of H are positive, then the critical point is a local minimum. Thus we have a minimum at (-2/3,-1/3,1) and the minimum value is -4/3.

2. Gradient of a Vector-Valued Function

a) Gradient of a vector with respect to the input vector

For a function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ and a vector valued function $\boldsymbol{x} = [x_1, x_2, \ldots . x_n]^T \in \mathbb{R}^n$, the corresponding vector function is given as $\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} \in \mathbb{R}^m$. The gradient of

$f: \mathbb{R}^n \to \mathbb{R}^m$ with respect to $x \in \mathbb{R}^n$ is given by $\frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \left[\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} \cdots \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \right]$

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Example

4. Let $f = [\,3x_1^2x_2 \quad 2x_1 + x_2^8\,]^T$, then find the Gradient of f.

Solution:

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} \end{bmatrix}$$

$$= \begin{bmatrix} 6x_1x_2 & 3x_1^2 \\ 2 & 8x_2^7 \end{bmatrix}$$

b) Gradient of a scalar with respect to a matrix

If $f : \mathbb{R}^{M\times N} \to \mathbb{R}$ is a scalar function of a matrix $X \in \mathbb{R}^{M\times N}$, then the **gradient** of f with respect to X is defined as the $M \times N$ matrix given by

$$\nabla_X f = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \frac{\partial f}{\partial X_{12}} & \cdots & \frac{\partial f}{\partial X_{1N}} \\ \frac{\partial f}{\partial X_{21}} & \frac{\partial f}{\partial X_{22}} & \cdots & \frac{\partial f}{\partial X_{2N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{M1}} & \frac{\partial f}{\partial X_{M2}} & \cdots & \frac{\partial f}{\partial X_{MN}} \end{bmatrix} \in \mathbb{R}^{M\times N}.$$

Example

5. Let $f = 2x+4y+3z+4w$ and the matrix $A = \begin{bmatrix} x & y \\ z & w \end{bmatrix}$ the find the gradient of f w.r.t matrix A.

Solution: The gradient of f is $\frac{df}{dA} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} & \frac{\partial f}{\partial w} \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 3 & 4 \end{bmatrix}$

c) Proof one of the gradient identities involving trace.

Example

6. **Consider the function $f(\mathrm{x}) = \mathrm{Tr}(\mathrm{A}\,\mathrm{x})$, where $\mathrm{A} \in \mathrm{R}^{3\times 2}$ and $\mathrm{x} \in \mathrm{R}^{2\times 2}$. Determine the gradient of 'f' with respect to x, denoted as $\nabla_X f(\mathrm{x})$.**

Solution:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \text{ and } \mathrm{x} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}$$

$$A\mathrm{x} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}$$

$$A\mathrm{x} = \begin{pmatrix} a_{11}x_{11} + a_{12}x_{21} & a_{11}x_{12} + a_{12}x_{22} \\ a_{21}x_{11} + a_{22}x_{21} & a_{21}x_{12} + a_{22}x_{22} \end{pmatrix}$$

$$f(\mathrm{x}) = Tr(\mathrm{A}\,\mathrm{x})$$

$$f(\mathrm{x}) = (a_{11}x_{11} + a_{12}x_{21} + a_{21}x_{12} + a_{22}x_{22})$$

$$\nabla_X f(\mathrm{x}) = \begin{pmatrix} \frac{\partial f(\mathrm{x})}{\partial x_{11}} & \frac{\partial f(\mathrm{x})}{\partial x_{12}} \\ \frac{\partial f(\mathrm{x})}{\partial x_{21}} & \frac{\partial f(\mathrm{x})}{\partial x_{22}} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix} = A^T$$

$f = Tr(Ax)$ $\frac{df}{dx} = A^T$

innovate achieve lead

3. Gradient Descent

Definition: Gradient Descent is an optimization algorithm used to minimize a function by iteratively moving in the direction of steepest descent.

Purpose: The primary goal of Gradient Descent is to find the local minimum of a function. It's widely used in machine learning for optimizing models and finding optimal parameters.

How it Works:

- At each iteration, the algorithm computes the gradient of the objective function with respect to the parameters.
- It then updates the parameters in the opposite direction of the gradient to reduce the value of the function.
- This process continues until convergence is achieved or a stopping criterion is met.

Algorithm

Data : $x_0 \in \mathbb{R}^n$

Step 0: Set $i = 0$

Step 1: If $\nabla f(x_i) = 0$ then stop
else, compute search direction $h_i = -\nabla f(x_i)$.

Step 2: Compute the size $\lambda_i \in \arg\min_{\lambda \geq 0} f(x_i + \lambda \cdot h_i)$

Step 3: Set $x_{i+1} = x_i + \lambda_i h_i$ go to step 1.

Derivation of Step size (τ)

- Suppose we want to minimize the **quadratic function**: $f(x) = \frac{1}{2}x^T Qx - b^T x$
- Let the iterative formula for minimum point be $x_{k+1} = x_k - \tau_k \nabla f(x_k)$

Where τ_k be the step size

- Let $g(\tau) = f(x_{k+1}) = f(x_k - \tau_k \nabla f(x_k))$

$$g(\tau) = \frac{1}{2}\left[x_k - \tau_k \nabla f(x_k)\right]^T Q\left[x_k - \tau_k \nabla f(x_k)\right] - b^T\left[x_k - \tau_k \nabla f(x_k)\right]$$

$$g(\tau) = a\tau^2 + d\tau + c$$

$$a = \frac{1}{2}\left[\nabla f(x_k)\right]^T Q \nabla f(x_k)$$

$$d = \left[b^T - x^T Q\right]\nabla f(x_k) = -\left[\nabla f(x_k)\right]^T \nabla f(x_k)$$

Here $g(\tau)$ is quadratic and concave function

- Condition for extrema of $\frac{dg}{d\tau} = 0$ is

$$\frac{dg}{d\tau} = 2a\tau + d = 0 \Rightarrow \tau = -\frac{d}{2a}$$

$$\tau_k = \frac{\left[\nabla f(x_k)\right]^T \nabla f(x_k)}{\left[\nabla f(x_k)\right]^T Q \nabla f(x_k)}$$

$$Therefore$$

$$\tau = \frac{S^T S}{S^T Q S}$$

4. Gradient Descent Examples

Example:

7. Find the minimum of $f(x,y) = 3x^2+y^2$ by

a) Computing the gradient of f and τ

b) Solve with initial values $x_0 = 1$ and $y_0 = 3$

Solution: $f(x,y) = 3x^2 + y^2$ with $(x_0, y_o) = (1,3)$

$$f(1,3) = 3.1^2 + 3^2 = 12$$

Step1: We find $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$

$$s = \nabla f = \begin{bmatrix} 6x \\ 2y \end{bmatrix} \text{ and } H = \begin{bmatrix} 6 & 0 \\ 0 & 2 \end{bmatrix}$$

Step2: Compute $\tau_i = \dfrac{s^T s}{s^T H s}$ is the step size.

$$\tau_i = \frac{\begin{bmatrix} 6x_i & 2y_i \end{bmatrix} \begin{bmatrix} 6x_i \\ 2y_i \end{bmatrix}}{\begin{bmatrix} 6x_i & 2y_i \end{bmatrix} \begin{bmatrix} 6 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 6x_i \\ 2y_i \end{bmatrix}}$$

$$\tau_i = \frac{36x_i^2 + 4y_i^2}{216x_i^2 + 8y_i^2}$$

$x_0 = 3$

$y_0 = 1$

$$\alpha = \frac{(36)3^2 + 4.1^2}{216(3)^2 + 8.1^2}$$

Step:3 Iterate the minimum point

$$x_{i+1} = x_i - \tau_i \nabla f_i$$

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} = \begin{bmatrix} x_i \\ y_i \end{bmatrix} - \left(\frac{36x_i^2 + 4y_i^2}{216x_i^2 + 8y_i^2} \right) \begin{bmatrix} 6x_i \\ 2y_i \end{bmatrix}$$

1st Iteration

$i=0$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - \left(\frac{36x_0^2 + 4y_0^2}{216x_0^2 + 8y_0^2} \right) \begin{bmatrix} 6x_0 \\ 2y_0 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} - \left(\frac{36.1^2 + 4.3^2}{216.1^2 + 8.3^2} \right) \begin{bmatrix} 6.1 \\ 2.3 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} - \left(\frac{72}{288} \right) \begin{bmatrix} 6 \\ 6 \end{bmatrix}$$

$\frac{1}{\frac{288}{72}} = \frac{1}{4}$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 1.5 \end{bmatrix}$$

$$f(-0.5, 1.5) = 3(-0.5)^2 + (1.5)^2 = 3$$

2nd Iteration

$i=1$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} - \left(\frac{36x_1^2 + 4y_1^2}{216x_1^2 + 8y_1^2} \right) \begin{bmatrix} 6x_1 \\ 2y_1 \end{bmatrix}$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 1.5 \end{bmatrix} - \left(\frac{36.(-0.5)^2 + 4.(1.5)^2}{216.(-0.5)^2 + 8.(1.5)^2} \right) \begin{bmatrix} 6.(-0.5) \\ 2.(1.5) \end{bmatrix}$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 1.5 \end{bmatrix} - \left(\frac{18}{72} \right) \begin{bmatrix} -3 \\ 3 \end{bmatrix}$$

$\frac{1}{4} = 0.25$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.75 \end{bmatrix}$$

$$f(0.25, 0.75) = 3(0.25)^2 + (0.75)^2 = 0.75$$

3rd Iteration

i=2

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} - \left(\frac{36x_2^2 + 4y_2^2}{216x_2^2 + 8y_2^2} \right) \begin{bmatrix} 6x_2 \\ 2y_2 \end{bmatrix}$$

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.75 \end{bmatrix} - \left(\frac{36.(0.25)^2 + 4.(0.75)^2}{216.(0.25)^2 + 8.(0.75)^2} \right) \begin{bmatrix} 6.(0.25) \\ 2.(0.75) \end{bmatrix}$$

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.75 \end{bmatrix} - \left(\frac{4.5}{18} \right) \begin{bmatrix} 1.5 \\ 4.5 \end{bmatrix}$$

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} -0.125 \\ -0.375 \end{bmatrix}$$

$$f(-0.125, -0.375) = 3(-0.125)^2 + (-0.375)^2 = 0.1875$$

4th Iteration

i=3

$$\begin{bmatrix} x_4 \\ y_4 \end{bmatrix} = \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} - \left(\frac{36x_3^2 + 4y_3^2}{216x_3^2 + 8y_3^2} \right) \begin{bmatrix} 6x_3 \\ 2y_3 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} -0.125 \\ 0.375 \end{bmatrix} - \left(\frac{36.(-0.125)^2 + 4.(0.375)^2}{216.(-0.125)^2 + 8.(0.375)^2} \right) \begin{bmatrix} 6.(-0.125) \\ 2.(0.375) \end{bmatrix}$$

$$\begin{bmatrix} x_4 \\ y_4 \end{bmatrix} = \begin{bmatrix} -0.125 \\ 0.375 \end{bmatrix} - \left(\frac{1}{4} \right) \begin{bmatrix} -0.75 \\ 0.75 \end{bmatrix}$$

$$\begin{bmatrix} x_4 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0.0625 \\ 0.1875 \end{bmatrix}$$

$$f(0.0625, 0.1875) = 3(0.0625)^2 + (0.1875)^2 = 0.046875$$

At the end of 10th iteration, we get $\begin{bmatrix} x_{10} \\ y_{10} \end{bmatrix} = \begin{bmatrix} -0.001953 \\ 0.005859 \end{bmatrix}$ and the minimum value is

$$f(-0.001953, 0.005859) = 3(-0.001953)^2 + (0.005859)^2 = 0.000046$$

Gradient Descent with Exponential Learning Rate Decay

Example

8. Consider the one-dimensional cost function $J(w) = (w-4)^2$

The gradient of the cost function is given by $\frac{dJ}{dw} = 2(w-4)$

You are asked to minimize $J(w)$ using **Gradient Descent with an exponentially decaying learning rate**.

The learning rate at iteration t is defined as $\alpha_t = \alpha_n e^{-kt}$, where initial learning rate $\alpha_n = 0.5$, decay constant $k = 0.1$ and initial weight $w_0 = 0$. The gradient descent update rule is $w_{t+1} = w_t - \alpha_t \cdot \frac{dJ}{dw}$. Find

(a) Compute the learning rate α_t for t=0,1 and 2.

(b) Perform three iterations of gradient descent and compute w_1 ,w_2 ,and w_3.

Solution

(a)

Iteration t	Learning rate $\alpha_t = \alpha_n e^{-\mathrm{kt}}$
0	$\alpha_0 = 0.5\, e^{-0.1(0)} = \mathbf{0.5000}$
1	$\alpha_1 = 0.5e^{-0.1(1)} \approx \mathbf{0.4524}$
2	$\alpha_2 = 0.5e^{-0.1(2)} \approx \mathbf{0.4094}$

(b)

Iteration t	$\nabla J(w_t)$	$w_{t+1} = w_t - \alpha_t \cdot \frac{dJ}{dw}$
0	$\nabla J(w_0) = 2(0-4) = -8$	$w_1 = 0 - (0.5)(-8) = \mathbf{4.0}$
1	$\nabla J(w_1) = 2(4-4) = 0$	$w_2 = 4 - (0.4524)(0) = \mathbf{4.0}$
2	$\nabla J(w_2) = 0$	$w_3 = 4 - (0.4094)(0) = \mathbf{4.0}$

Gradient Descent with Inverse Decay Rate.

Example: The same previous problem is solved using Inverse Decay

9. Given a Cost function: $J(w) = (w-4)^2$ and Gradient is $\frac{dJ}{dw} = 2(w-4)$.

Inverse (time-based) learning rate decay: $\alpha_t = \frac{\alpha_n}{1+kt}$. Parameters: initial learning rate $\alpha_n = 0.5$, Decay constant $k = 0.1$ and initial weight $w_0 = 0$.

Solution

(a)

Iteration t	Learning rate $\alpha_t = \frac{\alpha_0}{1+kt}$
0	$\alpha_0 = \frac{0.5}{1+(0.1)(0)} = \mathbf{0.5000}$
1	$\alpha_1 = \frac{0.5}{1+(0.1)(1)} \approx \mathbf{0.4545}$
2	$\alpha_1 = \frac{0.5}{1+(0.1)(1)} \approx \mathbf{0.4167}$

(b)

Iteration t	$\nabla J(w_t)$	$w_{t+1} = w_t - \alpha_t \cdot \frac{dJ}{dw}$
0	$\nabla J(w_0) = 2(0-4) = -8$	$w_1 = 0-(0.5)(-8) = \mathbf{4.0}$
1	$\nabla J(w_1) = 2(4-4) = 0$	$w_2 = 4-(0.4545)(0) = \mathbf{4.0}$
2	$\nabla J(w_2) = 0$	$w_3 = 4-(0.4167)(0) = \mathbf{4.0}$

Determining Step Size in Gradient Descent Using Binary Search

In machine learning optimisation, choosing an appropriate **step size (learning rate)** is crucial for fast and stable convergence. One way to determine the step size is by performing a **binary search (line search)** over a range of possible learning rates.

Example:

10. Consider the cost function: $J(w) = (w - 3)^2$. The gradient of the cost function is $\nabla J(w) = 2(w - 3)$. You are currently at the parameter value $w = 0$
Assume that the learning rate α is known to lie in the interval $\alpha \in [0, 1]$.
The goal is to choose α such that the **updated cost** $J(w - \alpha \nabla J(w))$ is **minimized**.

(a) Write the expression for the updated weight w' in terms of α.
(b) Write the cost function $J(w')$ as a function of α.
(c) Use **binary search** to determine the optimal step size α up to **two decimal places**. Stop when the interval width is less than 0.05.
(d) Compute the updated weight using the optimal α.

Solution (a) Updated weight as a function of α

$$w' = w - \alpha \nabla J(w)$$

At $w = 0$,

$$\nabla J(0) = 2(0 - 3) = -6$$

$$w' = 0 - \alpha(-6) = 6\alpha$$

(b) Cost as a function of α $J(w') = (6\alpha - 3)^2$

(See next slide for why it is **convex quadratic function**)

This is a **convex quadratic function** in α, so binary search can be applied.

(c). Iteration 1:

Initial interval: $a = 0, b = 1$

Midpoint: $m = \frac{0+1}{2} = 0.5$

Derivative of J with respect to α: $J'(\alpha) = 72\alpha - 36$

Evaluate at midpoint: $J'(0.5) = 72 \times 0.5 - 36 = 36 - 36 = 0$

Since the derivative is exactly zero, the stationary point (minimum) is found immediately.

So optimal $\alpha = 0.5$

(d) Updated Weight Using Optimal Step Size $w' = 6 \times 0.5 = 3.$

$$J(\alpha) = 36\alpha^2 - 36\alpha + 9$$

The coefficient of α^2 is $36 > 0$, so the parabola opens upwards.

For any quadratic function $f(x) = ax^2 + bx + c$:

- If $a > 0$, it is **convex** (also called strictly convex if $a > 0$).
- If $a < 0$, it is concave.
- If $a = 0$, it is linear (both convex and concave).

Since the leading coefficient $36 > 0$, $J(\alpha)$ is **strictly convex** in α.

A convex function has the property that any local minimum is also the global minimum, and the function has at most one minimum.

Determining Step Size in Gradient Descent Using Golden-Section Search

Solve the same previous problem by the Golden-Section search to answer the following:

(a) Write the expression for the updated weight w' in terms of α.

(b) Write the cost function $J(w')$ as a function of α.

(c) Use **binary search** to determine the optimal step size α up to **two decimal places**. Stop when the interval width is less than 0.05.

(d) Compute the updated weight using the optimal α.

5. Previous Year Problems

Q1: Consider the function $f(x) = ax^3 + bx^2 + cx + d$ where $a > 0$.

(a) Find the condition on a, b, c such that the function has two distinct critical points. Calculate the critical points in terms of a, b, c. Identify the nature of each critical point (i.e. maxima or minima). [3 Marks]

(b) Define a zone of attraction around each local minimum to be the region around it such that if gradient descent starts at any point in the region, it would end up at the given local minimum. Find the zone of attraction for each local minimum, if any, of the critical points. Justify your answer mathematically. [3 Marks]

$f(x) = ax^3 + bx^2 + cx + d$

Q1a) Solution :

To find the critical points we take $\frac{df}{dx} = 3ax^2 + 2bx + c = 0$, to obtain two roots

$$x_1 = \frac{-b - \sqrt{b^2 - 3ac}}{3a} \quad \text{and} \quad x_2 = \frac{-b + \sqrt{b^2 - 3ac}}{3a}.$$

In order for the two roots to be real and distinct, the quantity under the square-root sign needs to be strictly positive, i.e.

$$b^2 - 3ac > 0.$$

$ax^2 + bx + c = 0$

$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

We see that the second derivative $\frac{d^2 f}{dx^2} = 6ax + 2b.$

The second derivative is negative for the critical point x_1 and positive for x_2 which means x_1 is a maxima and x_2 is a minima.

Q1b) Solution :

As identified in part (a), there is a single local minimum $x_2 = \dfrac{-b + \sqrt{b^2 - 3ac}}{3a}$.

We can rewrite $\dfrac{df}{dx} = 3a\left(x + \dfrac{b}{3a}\right)^2 + \left(c - \dfrac{b^2}{3a}\right)$.

Solving for $\frac{df}{dx} = 0$

we see that $\frac{df}{dx} < 0$ when $x > \frac{-b-\sqrt{b^2-3ac}}{3a}$ and $x < \frac{-b+\sqrt{b^2-3ac}}{3a}$.

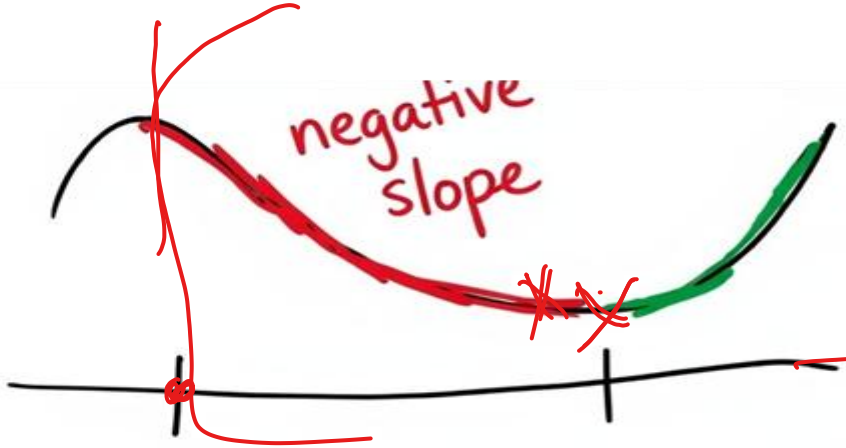
When $x > \frac{-b+\sqrt{b^2-3ac}}{3a}$, $\frac{df}{dx} > 0$.

Gradient descent will take any point on the left of the local minimum x_2 but greater than $\frac{-b-\sqrt{b^2-3ac}}{3a}$ to x_2 in a sufficiently large number of steps with a suitable step-size.

Similarly since the derivative is positive on the right of x_2, gradient descent will take any point on the right of x_2 to x_2 in a sufficiently large number of steps with a suitable step-size.

Thus the zone of attraction for the local minimum $x_2 = \frac{-b+\sqrt{b^2-3ac}}{3a}$ is $\dfrac{-b - \sqrt{b^2 - 3ac}}{3a} < x < \infty$.

1) a)



$x_1 = \dfrac{-b - \sqrt{b^2 - 3ac}}{3a}$ max

$x_2 = \dfrac{-b + \sqrt{b^2 - 3ac}}{3a}$ min

$x_1 < x < -\infty$

Q2:

Consider a quadratic function $f(x, y) = x^2 + \beta y^2$ where $\beta \in \mathbb{R}$ is an unknown constant. Also assume that $\beta > 0$. Consider the problem of minimizing this function using gradient descent algorithm:

(i) Derive a closed form expression (involving β) for the optimal step-size α for the first iteration of gradient descent if the initial point is

$$\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

(ii) If it is given to you that the optimal step-size $\alpha = 0.5$, derive the value of constant β using the formula derived in (i).

$f(x,y) = x^2 + \beta y^2$

Q2)a) Solution:

The gradient of $f(x,y)$ at (x_0,y_0) is as follows $\nabla f(x_0,y_0) = \begin{bmatrix} 2x_0 \\ 2\beta y_0 \end{bmatrix} = \begin{bmatrix} 2 \\ 2\beta \end{bmatrix}$.

$f_x = 2x$

$f_y = 2\beta y$

The optimal step size is obtained as the solution of the following one-dimensional optimization problem $\text{argmin}_\alpha f\left(\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - \alpha \nabla f(x_0,y_0)\right)$.

$f\left((x_0,y_0) - \alpha \nabla f(x_0,y_0)\right)$

After substituting $x_0 = 1$ and $y_0 = 1$, this is equivalent to solving $\text{argmin}_\alpha f(1-2\alpha,\ 1-2\alpha\beta)$.

$f\left((1,1) - \alpha(2,2\beta)\right)$

In other words, we need to minimise the function $g(\alpha) = (1-2\alpha)^2 + \beta(1-2\alpha\beta)^2$.

To find the minimum of $g(\alpha)$ with respect to α, we compute the derivative $g'(\alpha)$ and set it to zero:

$$g'(\alpha) = -4(1-2\alpha) + 2\beta(1-2\alpha\beta)(-2\beta) = 0.$$

$f(1-2\alpha,\ 1-2\alpha\beta)$

Expanding: $g'(\alpha) = -4 + 8\alpha - 4\beta^2 + 8\alpha\beta^3 = 8\alpha(1+\beta^3) - 4(1+\beta^2) = 0.$

$2\alpha(1+\beta^3) - (1+\beta^2) = 0$

Solving for α: $8\alpha(1+\beta^3) = 4(1+\beta^2) \Rightarrow \alpha = \dfrac{1+\beta^2}{2(1+\beta^3)}$.

Hence the closed form expression for the optimal step-size α is $\alpha = \dfrac{1+\beta^2}{2(1+\beta^3)}$.

$\alpha = \dfrac{1+\beta^2}{2(1+\beta^3)}$

$\alpha = \frac{1}{2}$

Q2)b) Solution:

If it is given that $\alpha = 0.5$, substituting in the previous expression we get the equation

$$0.5 = \frac{1+\beta^2}{2+2\beta^3}.$$

Rearranging we get

$$1+\beta^2 = 0.5(2+2\beta^3) = 1+\beta^3.$$

$$\beta^2 = \beta^3 \quad \Rightarrow \quad \beta^3 - \beta^2 = 0 \quad \Rightarrow \quad \beta^2(\beta-1) = 0.$$

Hence potential values of β are 0 or 1.

It is given in the question that $\beta > 0$ and (implicitly) $\beta \neq 0$ since it is an unknown positive constant.

Hence the final answer is $\beta = 1$.

Q3:

A data science intern arrived at a loss function given by

$$f(x) = 3x^4 - 20x^3 + 36x^2 + 10.$$

i) Help the intern to find the stationary points and classify and hence the global minima. [3 marks]

ii) Suggest the intern whether $x = 0.5$ or $x = 3.5$ is a better initial condition to find global minima using simple gradient descent method with reasons. [1 mark]

Q3)i) Solution:

Given $f(x) = 3x^4 - 20x^3 + 36x^2 + 10$. So, we have

$$f'(x) = 12x^3 - 60x^2 + 72x = 0$$

$$\Rightarrow \quad 12x(x-2)(x-3) = 0 \quad \Rightarrow \quad x = 0,\ 2,\ 3.$$

$$f''(x) = 36x^2 - 120x + 72$$

$$\Rightarrow \quad f''(0) = 72 > 0 \quad \Rightarrow \quad \text{Point of local minima}$$

$$\Rightarrow \quad f''(2) = -24 < 0 \quad \Rightarrow \quad \text{Point of local maxima}$$

$$\Rightarrow \quad f''(3) = 36 > 0 \quad \Rightarrow \quad \text{Point of local minima}$$

Now evaluate the function values:

$f(0) = 10, \quad f(3) = 3(81) - 20(27) + 36(9) + 10 = 243 - 540 + 324 + 10 = 37.$

Since $f(0) = 10 < f(3) = 37$, and there is a local maximum at $x = 2$, the global minimum is at $x = 0$ with $f(0) = 10$.

Therefore 10 is the answer (the value of the global minimum).

Q3)ii) Solution:

Clearly, 0.5 is closer to 0 (global minima) and 3.5 is closer to 3 (local minima). Starting near the global minimum helps gradient descent converge to it faster and avoids getting stuck in the local minimum at x=3 .
So, x = 0.5 is a better initial condition.

THANK YOU