

**Course Name: Optimization Theory and Algorithms**  
**Professor Name: Dr. Uday K. Khankhoje**  
**Department Name: Electrical Engineering**  
**Institute Name: Indian Institute of Technology Madras**  
**Week - 02**  
**Lecture - 11**

**Summary of Background Material - Calculus 3**

Convexity

We discussed convexity in relation to three different aspects: the convexity of points (convex combination of points), sets, and functions. The question raised was: can convexity be checked using a double derivative test?

Optimization - Review of Calculus.

$$C_1 \|x\|_b \leq \|x\|_a \leq C_2 \|x\|_b$$

To clarify, we refer to the convexity of a function. When we visualize a function, it may appear concave yet is described as convex. For a function to be convex, any convex combination of the function values must maintain this property. The double derivative test can confirm this. The double derivative captures changes in the first derivative, indicating convexity when  $f''(x) \geq 0$ . This means that if the double derivative exists and is non-negative, the function is convex.

In the case of multivariable functions, the equivalent of the second derivative is known as the Hessian matrix. We will delve into this concept further in our discussions on second-order methods, often referred to as Newton methods.

## Convex Combination

For convex functions, the parameter  $\alpha$  must lie between 0 and 1. We previously proved continuity for the  $p$ -norm with  $p = 2$ .

Optimization Theory and Algorithms

NPTEL

Derivative of a function

$$f'(x) = \lim_{y \rightarrow 0} \frac{f(x+y) - f(x)}{y} \quad x, y \in \mathbb{R}$$
$$\lim_{y \rightarrow 0} \frac{f(x+y) - f(x) - y f'(x)}{y} = 0$$

OPTIMIZATION THEORY AND ALGORITHMS

If we consider different norms, we need to address the concept of equivalence of norms, which states that if we obtain a result for one norm, we can derive bounds for another norm through specific constants.

## Gradient Descent

Next, a question arose regarding the use of gradient descent in machine learning, particularly concerning loss functions and backpropagation.

Gradient descent is favored because it is a simple yet effective algorithm that scales well with problem complexity.

## Continuity

We also explored the difference between uniformly continuous and Lipschitz continuous functions. A Lipschitz continuous function restricts the rate of change of the function, indicating that the derivative is bounded. This prevents sharp transitions in the function's behavior. While uniformly continuous functions also exhibit controlled change, not all uniformly continuous functions are Lipschitz continuous. However, it is established that every Lipschitz continuous function is uniformly continuous.

## Derivatives in Multivariable Calculus

We then transitioned to discussing derivatives of multivariate functions. The definition of the derivative can be daunting.

The image shows a whiteboard with handwritten mathematical notes. At the top right is the NPTEL logo. The main text on the board is:

$$\lim_{y \rightarrow 0} \left[ \frac{f(x+y) - f(x) - y^T \nabla f(x)}{\|y\|} \right] = 0$$

Go to  $x, y \in \mathbb{R}^n$   
 $f: \mathbb{R}^n \rightarrow \mathbb{R}$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \quad n \times 1 \text{ Col vector.}$$

Define  $\nabla f$ :  $\lim_{\|y\| \rightarrow 0} \frac{f(x+y) - f(x) - \nabla f^T y}{\|y\|}$

At the bottom left, the text "OPTIMIZATION THEORY AND ALGORITHMS" is written in bold. On the right side, there is a small inset video of a man speaking.

Starting from high school calculus, we recall that the derivative is defined as a limit. If  $f$  is a function, we express this as:

$$\lim_{y \rightarrow 0} \frac{f(x+y) - f(x)}{y}$$

This limit exists if the function is differentiable.

In the context of  $\mathbb{R}^n$ , the gradient is denoted by  $\nabla f$ , which consists of partial derivatives:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

### Fréchet Derivative

If the gradient exists for all  $x$  in the domain, the function is said to be differentiable. When the gradient is continuous, the function is continuously differentiable.

In terms of functions  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ , this represents multiple outputs from multiple inputs. For instance, optimizing various factors such as time, fuel, and exertion leads to a more complex objective function. The Jacobian matrix captures the relationship of the gradients in this case:

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix},$$



which has dimensions  $m \times n$ .

Define  $v$  s.t.  $\lim_{\|y\| \rightarrow 0} \left[ \frac{J(x)y - (f(x+y) - f(x))}{\|y\|} \right] = 0$

If  $\nabla f$  exists s.t.  $\uparrow$  holds the  $\nabla f$  is called the Frechet derivative of  $f$ .

$\nabla f \rightarrow \nabla_x f$

$\hookrightarrow$  If  $\nabla f$  exists  $\forall x \in \text{dom} f \rightarrow f$  is differentiable.

**OPTIMIZATION THEORY AND ALGORITHMS**

2/3

$$\hookrightarrow f: \mathbb{R}^n \rightarrow \mathbb{R}^{(m)}$$

$$f: \mathbb{R}^n \rightarrow \mathbb{R}$$

$\nabla f \rightarrow$  Jacobian:

$$J \in \mathbb{R}^{m \times n}$$

$$J_{ij} = \frac{\partial f_i}{\partial x_j}$$

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \dots & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}_{m \times n}$$



## Chain Rule

Lastly, we discussed the chain rule. For functions  $x: \mathbb{R}^n$  and  $y: \mathbb{R}$ , the derivative can be expressed as:

s

$\hookrightarrow$  Chain rule

$x(t)$  &  $y(x)$

$$\frac{dy}{dt} = \frac{d}{dt} y(x(t))$$

Scalar case.

$$= \frac{dy}{dx} \cdot \frac{dx}{dt}$$

$\hookrightarrow x \in \mathbb{R}^n$

$$h(t) = f(x(t))$$

$$\nabla h(t) = \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial t} + \dots + \frac{\partial f}{\partial x_n} \frac{\partial x_n}{\partial t}$$

$$= (\nabla f)^T \Delta x$$



$$\frac{dy}{dt} = \frac{dy}{dx} \cdot \frac{dx}{dt}$$

In the multivariable context, for  $h(t) = f(x(t))$ , the derivative generalizes to:

$$\frac{dh}{dt} = \nabla f^T \cdot \frac{dx}{dt},$$

where  $\nabla f$  is the gradient vector.