IIT Hyderabad **Scribed By: Divij (2021101001)**
- **Haran Raajesh (2021101005)**

Optimization Methods(CS1.404) **Lecture #1**
Instructor: Dr. Naresh Manwani 18^{th} **Jan 2025**

Outline. *The scribe addresses second-order directional derivatives, the Taylor Series Approximation, and the truncated Taylor Series for* 1^{st} *and* 2^{nd} *orders. It also explores optimization problems involving global and local minima, presenting relevant theorems and conditions, including the Weierstrass Theorem and the First-Order Necessary Condition (FONC).*

1 Second order directional directive

The Hessian matrix of a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$ is a square matrix of second-order partial derivatives, capturing the local curvature of f. For a twice differentiable function f defined on $S \subseteq \mathbb{R}^n$, the Hessian matrix $H_f(\mathbf{x})$ at $\mathbf{x}$ is given by:

$$H_f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

Statement: For a set $S \subseteq \mathbb{R}^n$, let $f : S \to \mathbb{R}$ be a function that is twice differentiable and continuous. For any $\mathbf{x} \in S$ and $\mathbf{d} \in \mathbb{R}^n$, the second-order directional derivative of f at $\mathbf{x}$ in the direction $\mathbf{d}$ is given by:

$$D_{\mathbf{d}}^2 f(\mathbf{x}) = \mathbf{d}^\top H_f(\mathbf{x})\mathbf{d},$$

where $H_f(\mathbf{x})$ is the Hessian matrix of f at $\mathbf{x}$.

Proof

1. **Definition of the Second-Order Directional Derivative:**

$$D_{\mathbf{d}}^2 f(\mathbf{x}) = \frac{\partial^2}{\partial t^2} f(\mathbf{x} + t\mathbf{d})\Big|_{t=0}.$$

2. **Taylor Series Expansion:** Expanding $f(\mathbf{x} + t\mathbf{d})$ around $t = 0$ using the Taylor series:

$$f(\mathbf{x} + t\mathbf{d}) = f(\mathbf{x}) + t\nabla f(\mathbf{x})^\top \mathbf{d} + \frac{t^2}{2}\mathbf{d}^\top H_f(\mathbf{x})\mathbf{d} + o(t^2),$$

where:

- $\nabla f(\mathbf{x})$ is the gradient of f,
- $H_f(\mathbf{x})$ is the Hessian matrix of f,

- $o(t^2)$ represents higher-order terms that vanish faster than t^2 as $t \to 0$.

3. **Differentiating Twice with Respect to t:**

   (a) First derivative:

   $$\frac{\partial}{\partial t} f(\mathbf{x} + t\mathbf{d}) = \nabla f(\mathbf{x})^\top \mathbf{d} + t\mathbf{d}^\top H_f(\mathbf{x})\mathbf{d} + \frac{o(t^2)}{\partial t}.$$

   (b) Second derivative:

   $$\frac{\partial^2}{\partial t^2} f(\mathbf{x} + t\mathbf{d}) = \mathbf{d}^\top H_f(\mathbf{x})\mathbf{d} + \frac{o(t^2)}{\partial t^2}.$$

4. **Evaluate at $t = 0$:** At $t = 0$, $o(t^2)$ vanishes, so:

   $$\frac{\partial^2}{\partial t^2} f(\mathbf{x} + t\mathbf{d})\Big|_{t=0} = \mathbf{d}^\top H_f(\mathbf{x})\mathbf{d}.$$

5. **Conclusion:** By the definition of the second-order directional derivative:

   $$D^2_{\mathbf{d}} f(\mathbf{x}) = \mathbf{d}^\top H_f(\mathbf{x})\mathbf{d}.$$

2 Taylor Series Approximation

The **Taylor series** provides an approximation of a function $f : \mathbb{R}^n \to \mathbb{R}$ around a point $\mathbf{x}_0$ using its derivatives at that point. For f that is m-differentiable at $\mathbf{x}_0$, the series gives an expression for how f behaves locally, incorporating derivatives of f up to the m-th order.
The Taylor series expansion of f around $\mathbf{x}_0$ is:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top H_f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \cdots + R_m(\mathbf{x}),$$

where:

- $\nabla f(\mathbf{x}_0)$ is the gradient of f at $\mathbf{x}_0$,
- $H_f(\mathbf{x}_0)$ is the Hessian matrix of f at $\mathbf{x}_0$,
- Higher-order terms involve higher derivatives of f,
- $R_m(\mathbf{x})$ is the remainder term after m-th order.

3 Truncated Taylor Series

3.1 First order

The first-order truncated Taylor series expansion for a function $f : \mathbb{R}^n \to \mathbb{R}$, differentiable at $\mathbf{x}_0$, is:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0).$$

This provides a linear approximation of f around $\mathbf{x}_0$, capturing the local behavior of f in terms of its value and gradient.

3.1.1 Proof Using the Mean-Value Theorem

Statement of the Mean-Value Theorem:
If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, and $\mathbf{x}$ and $\mathbf{x}_0$ are two points in $\mathbb{R}^n$, there exists a point $\mathbf{c}$ on the line segment joining $\mathbf{x}_0$ and $\mathbf{x}$ such that:

$$f(\mathbf{x}) - f(\mathbf{x}_0) = \nabla f(\mathbf{c})^\top (\mathbf{x} - \mathbf{x}_0),$$

where $\mathbf{c} = \mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)$ for some $t \in (0, 1)$.
Proof of the First-Order Taylor Approximation:

1. The goal is to approximate $f(\mathbf{x})$ near $\mathbf{x}_0$ using:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0).$$

2. Applying the mean-value theorem for f, we have:

$$f(\mathbf{x}) - f(\mathbf{x}_0) = \nabla f(\mathbf{c})^\top (\mathbf{x} - \mathbf{x}_0),$$

where $\mathbf{c}$ lies on the line segment between $\mathbf{x}_0$ and $\mathbf{x}$.

3. Since $\mathbf{x}$ is close to $\mathbf{x}_0$, the gradient $\nabla f(\mathbf{c})$ is close to $\nabla f(\mathbf{x}_0$, because f is differentiable and ∇f is continuous near $\mathbf{x}_0$. Thus:

$$f(\mathbf{x}) - f(\mathbf{x}_0) \approx \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0).$$

4. Rearranging terms, we conclude:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0).$$

3.2 Second order

The second-order truncated Taylor series for a function $f : \mathbb{R}^n \to \mathbb{R}$, twice differentiable at $\mathbf{x}_0$, is:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top H_f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

This provides a quadratic approximation of f around $\mathbf{x}_0$, incorporating the local curvature of f through the Hessian matrix $H_f(\mathbf{x}_0)$.

3.2.1 Proof Using the Mean-Value Theorem

Mean-Value Theorem for Second Derivatives:
For $f : \mathbb{R}^n \to \mathbb{R}$, twice differentiable, there exists a point $\mathbf{x}_1$ on the line segment joining $\mathbf{x}_0$ and $\mathbf{x}$ (i.e., $\mathbf{x}_1 = \mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)$ for some $t \in (0, 1)$) such that:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top H_f(\mathbf{x}_1)(\mathbf{x} - \mathbf{x}_0),$$

where $H_f(\mathbf{x}_1)$ is the Hessian matrix of f evaluated at $\mathbf{x}_1$, capturing the curvature of f along the segment.
Proof:

1. From the mean-value theorem, we write:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top H_f(\mathbf{x}_1)(\mathbf{x} - \mathbf{x}_0),$$

   where $\mathbf{x}_1$ lies on the line segment between $\mathbf{x}_0$ and $\mathbf{x}$.

2. If $\mathbf{x}$ is close to $\mathbf{x}_0$, the Hessian $H_f(\mathbf{x}_1) \approx H_f(\mathbf{x}_0)$ because f is twice differentiable, and H_f is continuous near $\mathbf{x}_0$.

3. Substituting $H_f(\mathbf{x}_0)$ in place of $H_f(\mathbf{x}_1)$, we get:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top H_f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

4 Global Minima optimization Problem

In this section, we consider the optimization problem requiring the minimization of a function. This function $f : \mathbb{R}^n \to \mathbb{R}$ is called the *objective function* or *cost function.* Let $S \subseteq \mathbb{R}^n$ and $f : S \to \mathbb{R}$ then the optimization problem in consideration becomes

$$\min_{\overline{\mathbf{x}} \in S} f(\overline{\mathbf{x}})$$

We say that $\overline{\mathbf{x}}^*$ is a global minima of f over set S if

$$f(\overline{\mathbf{x}}^*) \leq f(\overline{\mathbf{x}}) \quad \forall\, \overline{\mathbf{x}} \in S$$

There are also optimization problems that require maximization of the *objective function.* These problems, however, can be represented in the above form because maximizing $\mathbf{f}$ is equivalent to minimizing $-\mathbf{f}$. Therefore, we can confine our attention to minimization problems without loss of generality.

4.1 Conditions for existence of Global Minima

The following conditions on the set **S** and function **f** are sufficient to ensure that the global minima of **f** exists over **S**:

1. **f should be continuous in S**
   Let $S = [-1, 1]$ and $f : S \to \mathbb{R}$ such that

$$f(x) = \begin{cases} x & \text{if } x \in (-1, 1) \\ 0 & \text{if } x \in \{-1, 1\} \end{cases}$$

   The graph for function f is given in Figure 1. Here,

$$f(-1) = 0 \tag{1}$$

$$\begin{aligned} f(-1+\delta) &= -1 + \delta \\ \lim_{\delta \to 0} f(-1+\delta) &= -1 \end{aligned} \tag{2}$$

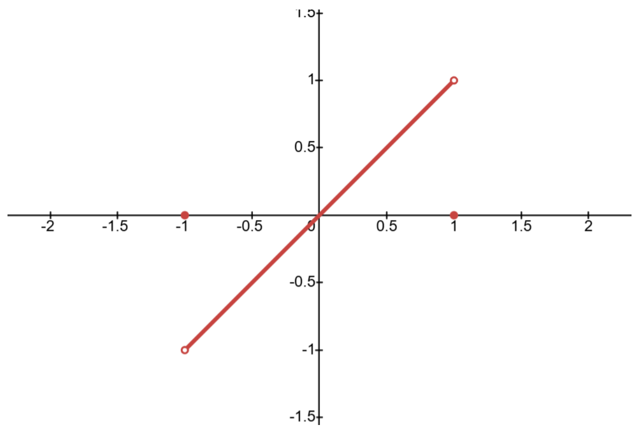

Figure 1: Discontinuous Function

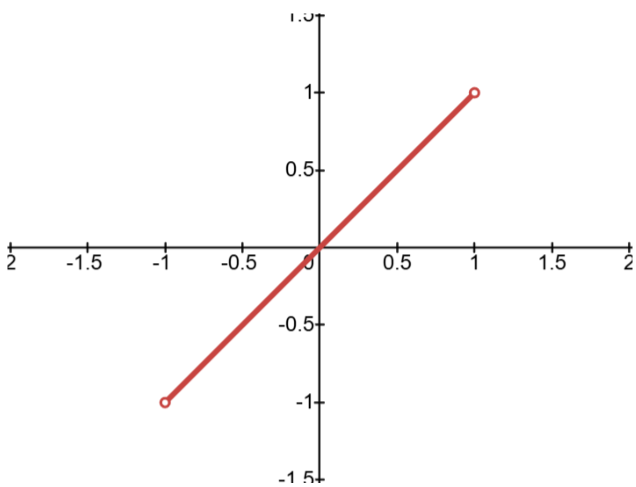

Figure 2: Open set

From (1) and (2), we observe that f is not continuous at -1. Similarly, it can be shown that f is not continuous at 1 as well.

$$f(1) = 0 \tag{3}$$

$$\begin{aligned} f(1-\delta) &= -1-\delta \\ \lim_{\delta \to 0} f(1-\delta) &= 1 \end{aligned} \tag{4}$$

Hence, f cannot attain global minima at -1, and therefore, it is not possible to define a global minima for f.

2. S **should be a closed set**
Let $S = (-1, 1)$ and $f : S \to \mathbb{R}$ such that

$$f(x) = x \quad \forall\, x \in S$$

The graph for function f is given in Figure 2. For this function, it is possible to investigate the infimum in this case but not the minimum/minima. This is because the minimum should

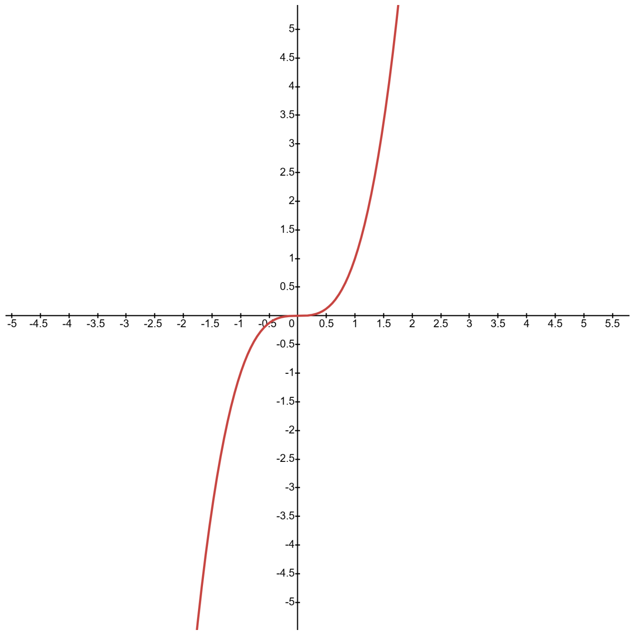

Figure 3: Unbounded Set

always be attained, while the infimum isn't necessarily. f attains it's minima at -1, but $-1 \notin S$. This is because S is an open set, and so the boundary points {-1, 1} are not included. Therefore, f never attains the value -1 and hence, does not have a global minima.

3. **S should be a bounded set**
   Let $S = \mathbb{R}$ and $f : S \to \mathbb{R}$ such that

$$f = x^3 \quad \forall\, x \in \mathbb{R}$$

   The graph for function f is given in Figure 3. For this function, global minima of f cannot be defined as the minimum value achieved by f over $\mathbb{R}$ is $-\infty$ which is undefined.

It is possible to express conditions 2 and 3 as one condition that **S must be a compact set**.

It is possible for a discontinuous f^n to attain minimum value at some point, such as

$$f(x) = \begin{cases} 0 & \text{if } x \in (-1, 1) \\ x & \text{if } x \in \{-1, 1\} \end{cases}$$

Similar examples can be given for functions defined over an open set and functions defined over an unbounded set. For example,

$$f(x) = |x| \quad \forall\, x \in (-1, 1) \quad \text{and,}$$

$$f(x) = x^2 \quad \forall\, x \in \mathbb{R}$$

It's easily observable that these conditions are not necessary for the existence of a minimum value for a function f over set S. However, these conditions are still included since the goal is to define only the *sufficient conditions* for attaining global minima. Hence, the *necessity* of a condition does not play much role in defining them.

4.2 Weierstrass Theorem

Let $S \subseteq \mathbb{R}^n$ be a non-empty compact set and $f : S \to \mathbb{R}$ be a continuous function over set S. Then f attains a maxima and a minima on S, that is, $\exists\, \overline{X}_{min}, \overline{X}_{max} \in S$ such that

$$f(\overline{X}_{min}) \leq f(\overline{X}) \leq f(\overline{X}_{max}) \quad \forall\, \overline{X} \in S$$

Therefore, *Weierstrass Theorem* defines the existence of global minima based on the conditions discussed in Section 4.1. However, it's apparent that there are flaws with this theory since these conditions are only sufficient but not strictly necessary for the existence of global minima. This is why this theorem is not used widely these days for real-world applications.

5 Local Minima optimization problem

5.1 Local Minima

The previous section discussed optimization problems regarding the minimization of X over the entire set S. The solution to that problem was the global minima of f over the set S, however, in this section we discuss the local minima of f instead. Let $S \subseteq \mathbb{R}^n$ and $f : S \to \mathbb{R}$ and consider the optimization problem

$$\min_{\overline{\mathbf{x}} \in S} f(\overline{\mathbf{x}})$$

Then, we say that $\overline{X}^*$ is a local minima of f over S if $\exists\, \delta > 0$ such that

$$f(\overline{X})^* \leq f(\overline{X}) \quad \forall\, \overline{X} \in B(\overline{X}^*, S)$$

Here, $B(\overline{X}^*, S)$ represents the ball around $\overline{X}^*$ in set S. However, this definition is incomplete and does not hold true when the ball of $\overline{X}^*$ doesn't lie completely inside set S. For example, consider a simple function f such that

$$f(x) = x \quad \forall\, x \in S = [-1, 1]$$

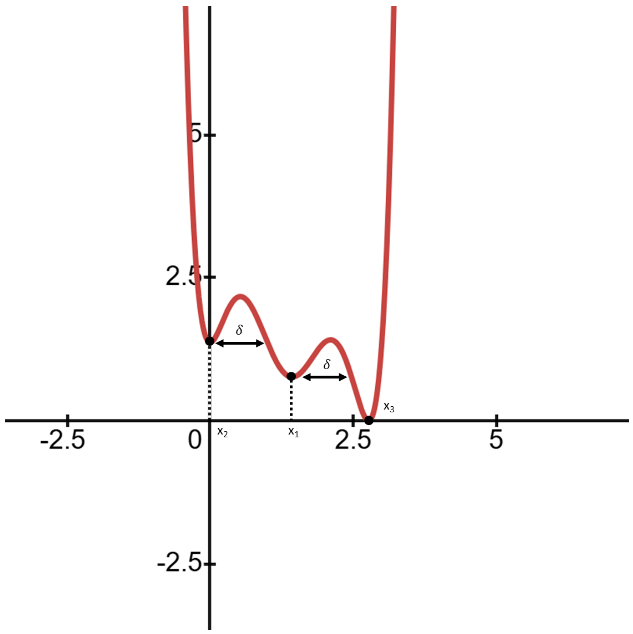

Figure 4: Local Minima

Here, the left neighborhood of $\overline{X}^*$ or simply, -1 does not exist within the set S. Hence, f is undefined in the left neighborhood and hence, for part of the ball of -1. Therefore, it's better to re-define the earlier condition as

$$f(\overline{X})^* \leq f(\overline{X}) \quad \forall\, \overline{X} \in B(\overline{X}^*, S) \cap S$$

> Due to the equality condition present in the condition, every point on a constant function is a local minima for that function.

Example
In Figure 4, x_1, x_2 and, x_3 are all local minima for the function f over $\mathbb{R}$. This can be seen as:

- $f(x_3)$ is the minimum value achieved by the function over $\mathbb{R}$. Hence, it is minimum in both it's left and right neighborhood.
- $f(x_2)$ is the minimum value achieved by the function compared to the function on the left of x_2 hence, it is clearly the minimum value in it's left neighborhood and even for the right neighborhood, based on graphical inference, we can say that

$$f(x_2) \leq f(x) \quad \forall\, x \in (x_2 - \delta, x_2 + \delta).$$

- We can show similar calculations for x_1 as we have done for x_2.

5.2 Strict Local Minima

Let $S \subseteq \mathbb{R}^n$ and $f : S \to \mathbb{R}$, then $\overline{X}^* \in S$ is said to be a strict local minima of f over S if $\exists\, \delta > 0$ such that

$$f(\overline{X}^*) < f(\overline{X}) \quad \forall\, \overline{X} \in B(\overline{X}^*, \delta) \cap S \text{ and } \overline{X} \neq \overline{X}^*$$

The condition $\overline{X} \neq \overline{X}^*$ is a compulsory condition since LHS and RHS would be equal when $\overline{X} = \overline{X}^*$ and hence the condition of *strictly less than* will not hold.

> It is possible to define functions such that they have local minima while the global minima of the function does not exist. For example,
>
> $$f(x) = (x+1)x(x-1) \quad \forall\, x \in \mathbb{R}$$
>
> Here, the function f attains it's local minima somewhere in $(0, 1)$ while the global minima of the function is attained at $-\infty$ which is undefined.

6 First order necessary condition

Till now in previous sections, we have focused our attention on sufficient conditions on set S and f rather than the necessary conditions. However, in this section our main focus would be on the *necessary and sufficient* conditions for local minima. The difference between the two conditions can be understood in the following way:

- **Necessary Conditions**
  These are the conditions which are satisfied by every local minima.
- **Sufficient Conditions**
  These are the conditions which guarantee a local minima.

In order to achieve our goal, we rely on the First Order Necessary Condition (FONC).

Let $S \subseteq \mathbb{R}^n$, $f \in C^1(S)$ and $\overline{X}^*$ be a local minima for f over S. Then, according to the **First Order Necessary Condition (FONC)**, for any feasible direction $\overline{d}$ at $\overline{X}^*$, the following holds true:

$$\nabla f(\overline{X}^*)^T \overline{d} \geq 0$$

> Note that a feasible direction $\overline{d}$ at $\overline{X}^*$ is such that the following holds true for some $\alpha_0 > 0$:
>
> $$\overline{X}^* + \alpha\overline{d} \;\in S \quad \forall\, \alpha \in (0, \alpha_0]$$

Proof
Let $\overline{X}^*$ be the local minima of f and let $\overline{d}$ be a feasible direction at $\overline{X}^*$. Now, we define a new function $\phi(\alpha)$ such that for $\alpha > 0$

$$\begin{aligned} \phi(\alpha) &= f(\overline{X}^* + \alpha\overline{d}) \\ \phi(0) &= f(\overline{X}^*) \\ \phi'(\alpha) &= \nabla f(\overline{x}^* + \alpha\overline{d})^T \overline{d} \\ \phi'(0) &= \nabla f(\overline{X}^*)^T \overline{d} \end{aligned} \tag{5}$$

Expanding $\phi(\alpha)$ around $\alpha = 0$ using first order Taylor series approximation,

$$\phi(\alpha) = \phi(0) + \alpha\phi'(0) + o(|\alpha|) \tag{6}$$

Substituting values from (5) into (6) gives

$$f(\overline{X}^* + \alpha\overline{d}) = f(\overline{X}^*) + \alpha\nabla f(\overline{X}^*)^T\overline{d} + o(|\alpha|) \tag{7}$$

Further, since $\overline{X}^*$ is a local minima and $\overline{d}$ is a feasible direction, then $\exists\ \alpha_0 \in \mathbb{R}^+$ such that

$$f(\overline{X}^* + \alpha\overline{d}) \geq f(\overline{X}^*) \quad \forall\ \alpha \in (0, \alpha_0] \tag{8}$$

Then after substituting the value of $f(\overline{X}^* + \alpha\overline{d})$ from (7) in (8), the following holds $\forall\ \alpha \in (0, \alpha_0]$

$$\begin{gathered} f(\overline{X}^*) + \alpha\nabla f(\overline{X}^*)^T\overline{d} + o(|\alpha|) \geq f(\overline{X}^*) \\ \alpha\nabla f(\overline{X}^*)^T\overline{d} + o(|\alpha|) \geq 0 \end{gathered} \tag{9}$$

Dividing (9) with α and applying $\lim \alpha \to 0$,

$$\nabla f(\overline{X}^*)^T\overline{d} + \lim_{\alpha\to 0} \frac{o(|\alpha|)}{\alpha} \geq 0 \tag{10}$$

The second term in (10) will be 0 on applying $\lim \alpha \to 0$ as $o(|\alpha|)$ (numerator) decays to 0 faster than α itself (denominator). Applying this to (10) gives

$$\nabla f(\overline{X}^*)^T\overline{d} \geq 0 \qquad \square$$

References

[1] Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB by Amir Beck

[2] An Introduction to Optimization by Edwin K. P. Chong and Stanislaw H. Żak