**IIIT Hyderabad**
**-**

**Scribed By: Abhinav Marri (2021112015)**
**Shreya Bollimuntha (2021112014)**

**Optimization Methods(CS1.404)**
**Instructor: Dr. Naresh Manwani**

**Lecture 12**
$20^{th}$ **Feb 2024**

---

**Outline.** *The document covers the Arithmetic Mean - Geometric Mean (AM-GM) inequality, providing a proof using Jensen's inequality along with practical examples to illustrate its significance. Following this, the second-order characterization of convex functions is discussed, highlighting the role of the Hessian matrix in determining convexity. Several examples demonstrate different scenarios where this property is utilized.*
*The later sections explore descent direction methods used in iterative optimization. The concept of descent direction, its mathematical formulation, and the descent property ensuring function value reduction in optimization algorithms are explained. Additionally, the document outlines the schematic steps for descent direction methods, the challenges in selecting descent directions and step sizes, and practical considerations for algorithm convergence.*

# 1 AM-GM Inequality

The Arithmetic Mean (AM)- Geometric Mean (GM) inequality is as follows:

$$\frac{1}{k}\sum_{i=1}^{k} x_i \geq \left(\prod_{i=1}^{k} x_i\right)^{\frac{1}{k}}$$

**Proof:** Let us consider the function:

$$f(x) = -\log(x)$$

which is ***convex*** since its second derivative,

$$f''(x) = \frac{1}{x^2},$$

is non-negative for all $x > 0$.

By Jensen's inequality, for any convex function $f(x)$,

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leq \sum_{i=1}^{k} \lambda_i f(x_i),$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are non-negative weights such that $\sum_{i=1}^{k} \lambda_i = 1$. Applying this to $f(x) = -\log(x)$, we get:

$$-\log\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leq \sum_{i=1}^{k} \lambda_i (-\log x_i)$$

Rearranging,

$$\log\left(\sum_{i=1}^{k}\lambda_i x_i\right) \geq \sum_{i=1}^{k}\lambda_i \log x_i$$

Choosing equal weights $\lambda_i = \frac{1}{k}$ for all $i$, we obtain

$$\log\left(\frac{1}{k}\sum_{i=1}^{k} x_i\right) \geq \frac{1}{k}\sum_{i=1}^{k}\log x_i$$

Exponentiating both sides,

$$\frac{1}{k}\sum_{i=1}^{k} x_i \geq e^{\frac{1}{k}\sum_{i=1}^{k}\log x_i}$$

The right-hand side can be simplified to,

$$e^{\frac{1}{k}\sum_{i=1}^{k}\log x_i} = \left(\prod_{i=1}^{k} x_i\right)^{\frac{1}{k}},$$

which is the geometric mean of variables $x_1, x_2, \ldots x_k$, we obtain the AM-GM inequality:

$$\frac{1}{k}\sum_{i=1}^{k} x_i \geq \left(\prod_{i=1}^{k} x_i\right)^{\frac{1}{k}}.$$

□

## Conclusion

Thus, we have proven that the arithmetic mean is always greater than or equal to the geometric mean:

$$\frac{x_1 + x_2 + \cdots + x_k}{k} \geq \sqrt[k]{x_1 x_2 \ldots x_k}$$

**Example 1.1**
Consider three positive numbers:

$$x_1 = 4, \quad x_2 = 1, \quad x_3 = 9$$

The arithmetic mean of the three numbers is

$$\text{AM} = \frac{x_1 + x_2 + x_3}{3} = \frac{4 + 1 + 9}{3} = \frac{14}{3} \approx 4.67.$$

The geometric mean of the three numbers is given as

$$\text{GM} = \sqrt[3]{x_1 x_2 x_3} = \sqrt[3]{4 \times 1 \times 9} = \sqrt[3]{36} \approx 3.30.$$

Since $4.67 \geq 3.30$, we observe:

$$\text{AM} \geq \text{GM}$$

This confirms that the arithmetic mean is always greater than or equal to the geometric mean.

## 2 Second Order Characterization of Convex Functions

**Theorem:** Let $S \subseteq \mathbb{R}^n$ be an open convex set. Let $f \in C^2(S)$. Then $f$ is convex if and only if $\nabla^2 f(\overline{x})$ is positive semidefinite for all $\overline{x} \in S$.

*Note:* This characterization can only be used if the function has second-order derivatives which is continuous. Thus, functions like $|x|$ cannot be used. Further, in the above theorem, $C^2(S)$ represent the set of all second-order partial derivatives.

**Proof:**
**(*Sufficiency*)** Assume $f$ is a convex function. Let $\overline{x} \in S$ and $\overline{d} \in \mathbb{R}^n$ be some direction. Since $S$ is an open set, there exists $\epsilon > 0$ such that

$$\overline{x} + \lambda \overline{d} \in S, \quad \forall \lambda \in (0, \epsilon).$$

Using the gradient property of convex functions, we can write

$$f(\overline{x} + \lambda \overline{d}) \geq f(\overline{x}) + \lambda \nabla f(\overline{x})^T \overline{d}. \tag{1}$$

Now, using the second-order Taylor series approximation, we get

$$f(\overline{x} + \lambda \overline{d}) = f(\overline{x}) + \lambda \nabla f(\overline{x})^T \overline{d} + \frac{\lambda^2}{2} \overline{d}^T \nabla^2 f(\overline{x}) \overline{d} + o(\lambda^2 \|\overline{d}\|^2). \tag{2}$$

Using (1) in (2), we obtain

$$\frac{\lambda^2}{2} \overline{d}^T \nabla^2 f(\overline{x}) \overline{d} + o(\lambda^2 \|\overline{d}\|^2) \geq 0.$$

Dividing by $\frac{\lambda^2}{2}$ and taking the limit as $\lambda \to 0$, we get

$$\overline{d}^T \nabla^2 f(\overline{x}) \overline{d} \geq 0.$$

Since this holds for all $\overline{x} \in S$ and for all $\overline{d} \in \mathbb{R}^n$, it follows that $\nabla^2 f(\overline{x})$ is positive semidefinite.

**(*Necessity*)** Suppose $\nabla^2 f(\overline{x})$ is positive semidefinite for all $\overline{x} \in S$. Let $\overline{x}_1, \overline{x}_2 \in S$. Using the truncated second-order Taylor series expansion, we write

$$f(\overline{x}_2) = f(\overline{x}_1) + \nabla f(\overline{x}_1)^T (\overline{x}_2 - \overline{x}_1) + \frac{1}{2} (\overline{x}_2 - \overline{x}_1)^T \nabla^2 f(\overline{z}) (\overline{x}_2 - \overline{x}_1), \tag{3}$$

where $\overline{z}$ lies on the line segment joining $\overline{x}_1$ and $\overline{x}_2$.
Since $\nabla^2 f(\overline{z})$ is positive semidefinite, we obtain

$$(\overline{x}_2 - \overline{x}_1)^T \nabla^2 f(\overline{z}) (\overline{x}_2 - \overline{x}_1) \geq 0. \tag{4}$$

Using (4) in (3), we get

$$f(\overline{x}_2) \geq f(\overline{x}_1) + \nabla f(\overline{x}_1)^T (\overline{x}_2 - \overline{x}_1).$$

Since this holds for all $\overline{x}_1, \overline{x}_2 \in S$, it follows that $f$ is a convex function. □

***Note:*** There exist some functions, where it is tough to prove convexity using conventional methods. This is where we can resort to using the second-order convexity theorem.

**Example 2.1**
**Function:** $f(x_1, x_2) = \frac{x_1^2}{x_2}$ for $x_2 > 0$

Compute the Hessian:

$$\nabla^2 f(x_1, x_2) = \begin{bmatrix} \frac{2}{x_2} & -\frac{2x_1}{x_2^2} \\ -\frac{2x_1}{x_2^2} & \frac{2x_1^2}{x_2^3} \end{bmatrix}$$

To check for positive semi-definiteness, compute trace and determinant:

$$\text{trace}(\nabla^2 f(x_1, x_2)) = \frac{2}{x_2} + \frac{2x_1^2}{x_2^3} \geq 0 \qquad \because x_2 > 0$$

$$\det(\nabla^2 f(x_1, x_2)) = \frac{4x_1^2}{x_2^4} - \frac{4x_1^2}{x_2^4} = 0$$

Since both determinant and trace of $\nabla^2 f(x_1, x_2)$ are non-negative, $\nabla^2 f(x_1, x_2)$ is positive semi definite. Thus, $f(x_1, x_2)$ is convex.

**Example 2.2**
**Function:** $f(x) = x \log x$ for $x > 0$

First derivative,

$$f'(x) = \log x + 1$$

Second derivative:

$$f''(x) = \frac{1}{x} > 0 \quad \text{for } x > 0$$

Thus, $f(x)$ is convex.

**Example 2.3** (Convexity of the log-sum-exp function)
**Function:** $f(\mathbf{x}) = \ln\left(e^{x_1} + e^{x_2} + \cdots + e^{x_n}\right)$
The *log-sum-exp* function is defined over the entire space $\mathbb{R}^n$. We will prove its convexity using the Hessian test. The partial derivatives of $f$ are given by

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \frac{e^{x_i}}{\sum_{k=1}^n e^{x_k}}, \quad i = 1, 2, \ldots, n,$$

and therefore

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \begin{cases} -\frac{e^{x_i} e^{x_j}}{\left(\sum_{k=1}^n e^{x_k}\right)^2}, & i \neq j, \\ -\frac{e^{x_i} e^{x_j}}{\left(\sum_{k=1}^n e^{x_k}\right)^2} + \frac{e^{x_i}}{\sum_{k=1}^n e^{x_k}}, & i = j. \end{cases}$$

We can thus write the Hessian matrix as

$$\nabla^2 f(\mathbf{x}) = \text{diag}(\mathbf{w}) - \mathbf{w}\mathbf{w}^T,$$

where $w_i = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}$. In particular, $\mathbf{w} \in \Delta_n$. To prove the positive semidefiniteness of $\nabla^2 f(\mathbf{x})$, take $0 \neq \mathbf{v} \in \mathbb{R}^n$ and consider the expression

$$\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} = \sum_{i=1}^n w_i v_i^2 - (\mathbf{v}^T \mathbf{w})^2.$$

The latter expression is nonnegative since employing the Cauchy-Schwarz inequality on the vectors $\mathbf{s}, \mathbf{t}$ defined by

$$s_i = \sqrt{w_i} v_i, \quad t_i = \sqrt{w_i}, \quad i = 1, 2, \ldots, n,$$

yields

$$(\mathbf{v}^T \mathbf{w})^2 = (\mathbf{s}^T \mathbf{t})^2 \leq \|\mathbf{s}\|^2 \|\mathbf{t}\|^2 = \left( \sum_{i=1}^n w_i v_i^2 \right) \left( \sum_{i=1}^n w_i \right).$$

Since $\mathbf{w} \in \Delta_n$, we have $\sum_{i=1}^n w_i = 1$, and thus

$$\sum_{i=1}^n w_i v_i^2 - (\mathbf{v}^T \mathbf{w})^2 \geq 0.$$

Since the latter inequality is valid for any $\mathbf{v} \in \mathbb{R}^n$, it follows that $\nabla^2 f(\mathbf{x})$ is indeed positive semidefinite. Thus, the function is convex. □

***Alternate Solution :***
Alternatively, to show that

$$\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} = \sum_{i=1}^n w_i v_i^2 - (\mathbf{v}^T \mathbf{w})^2 \geq 0,$$

we recognize that the right-hand side resembles the variance formula:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Here, we define a random variable $X$ that takes values $v_i$ with probability $w_i$, meaning its expectation is given by:

$$\mathbb{E}[X] = \sum_{i=1}^n w_i v_i.$$

Similarly, the expected value of $X^2$ is:

$$\mathbb{E}[X^2] = \sum_{i=1}^n w_i v_i^2.$$

Thus, using the variance identity:

$$\sum_{i=1}^n w_i v_i^2 - \left( \sum_{i=1}^n w_i v_i \right)^2 = \mathrm{Var}(X).$$

Since variance is always nonnegative, i.e., $\text{Var}(X) \geq 0$, it follows that:

$$\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} \geq 0.$$

This proves that the Hessian $\nabla^2 f(\mathbf{x})$ is positive semidefinite.

**Example 2.4**
**Function:** $f(x_1, x_2) = x_1^3 - 3x_1 x_2^2$

Compute the Hessian:

$$\nabla^2 f(x_1, x_2) = \begin{bmatrix} 6x_1 & -6x_2 \\ -6x_2 & -6x_1 \end{bmatrix}$$

To check for positive semi-definiteness, compute trace and determinant:

$$\text{trace}(\nabla^2 f(x_1, x_2)) = 6x_1 - 6x_1 = 0$$

$$\det(\nabla^2 f(x_1, x_2)) = (6x_1)(-6x_1) - (-6x_2)(-6x_2) = -36x_1^2 - 36x_2^2 = -36(x_1^2 + x_2^2)$$

Since the determinant is negative for any $(x_1, x_2) \neq (0, 0)$, the Hessian is not positive semi-definite. Thus, $f(x_1, x_2)$ is not convex.

# 3 Descent Direction Methods

In optimization, we consider the problem of unconstrained minimization of a function $f(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^n$ and $f$ is continuously differentiable over $\mathbb{R}^n$. A natural approach to finding the optimal points involves solving the equation $\nabla f(\mathbf{x}) = \mathbf{0}$ to determine the stationary points. However, in many cases, solving this equation analytically is challenging. Even when solutions exist, there may be infinitely many, and identifying the one that corresponds to a local minimum can be as difficult as solving the original optimization problem itself.

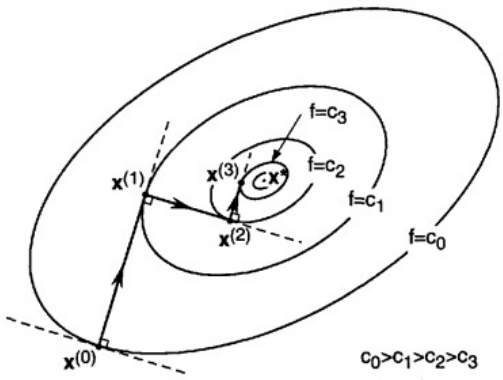


Figure 1: Caption

To overcome these difficulties, rather than seeking stationary points analytically, we employ iterative algorithms that progressively refine the solution. These algorithms follow an update rule of the

form:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \mathbf{d}_k, \quad k = 0, 1, 2, \ldots,$$

where $\mathbf{d}_k$ represents the descent direction, and $t_k$ is the step size that determines how far to move along this direction. The choice of $\mathbf{d}_k$ and $t_k$ plays a crucial role in ensuring convergence to a local minimum efficiently.

### 3.1 Descent Direction

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function over $\mathbb{R}^n$. A vector $\mathbf{d} \in \mathbb{R}^n$, where $\mathbf{d} \neq \mathbf{0}$, is called a **descent direction** of $f$ at a point $\mathbf{x}$ if moving in this direction results in a decrease in the function value. Mathematically, this condition is satisfied when the directional derivative of $f$ at $\mathbf{x}$ along $\mathbf{d}$ is negative, i.e.,

$$\nabla f(\mathbf{x})^T \mathbf{d} < 0.$$

This ensures that taking small enough steps along $\mathbf{d}$ leads to a reduction in the function value, making it a viable direction for iterative optimization methods.

### 3.2 Descent Property

Let $f$ be a continuously differentiable function defined over an open set $S \subset \mathbb{R}^n$, and let $\mathbf{x} \in S$. Suppose that $\mathbf{d}$ is a descent direction of $f$ at $\mathbf{x}$, meaning that the directional derivative along $\mathbf{d}$ is negative. Then, there exists a positive constant $\epsilon > 0$ such that for any step size $\alpha$ in the interval $(0, \epsilon]$, the function value strictly decreases:

$$f(\mathbf{x} + \alpha \mathbf{d}) < f(\mathbf{x}).$$

**Proof.** Since $f'(x; \mathbf{d}) < 0$, it follows from the definition of the directional derivative that

$$\lim_{\alpha \to 0^+} \frac{f(x + \alpha \mathbf{d}) - f(x)}{\alpha} = f'(x; \mathbf{d}) < 0.$$

Therefore, there exists an $\varepsilon > 0$ such that

$$\frac{f(x + \alpha \mathbf{d}) - f(x)}{\alpha} < 0$$

for any $\alpha \in (0, \varepsilon]$, which readily implies the desired result. □

This property confirms that taking sufficiently small steps in a descent direction guarantees a reduction in function value, which is a fundamental principle in iterative optimization methods.

### 3.3 Schematic Descent Directions Method

**Initialization:** Choose an arbitrary starting point $\mathbf{x}_0 \in \mathbb{R}^n$.
**General Step:** For $k = 0, 1, 2, \ldots$, perform the following steps:

1. Select a descent direction $\mathbf{d}_k$ such that moving in that direction reduces the function value.
2. Determine a step size $t_k$ that ensures $f(\mathbf{x}_k + t_k \mathbf{d}_k) < f(\mathbf{x}_k)$.

3. Update the current point:
$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \mathbf{d}_k.$$
4. Stop if a predefined stopping condition is met; otherwise, repeat from Step (1).

## Challenges and Their Solutions

While the method provides a structured way to approach optimization, several challenges arise in practice:

1. **How to choose the initial point $\mathbf{x}_0$?** The starting point is often selected arbitrarily, but choosing a good initial point can improve convergence speed. Domain knowledge or heuristics can help in selecting $\mathbf{x}_0$.
2. **How to choose the descent direction $\mathbf{d}_k$?** The choice of $\mathbf{d}_k$ depends on the optimization method used. Common choices include:
   - **Gradient Descent:** $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$, i.e., the direction of steepest descent.
   - **Newton's Method:** Uses second-order derivatives (Hessian matrix) for a more accurate search direction.
3. **How to choose the step size $t_k$?** The step size determines how far to move in the chosen direction. There are several approaches:
   - **Fixed step size:** $t_k$ is constant but may not always be optimal.
   - **Backtracking line search:** Gradually reduces $t_k$ until a sufficient decrease condition is met.
   - **Exact line search:** Finds the best $t_k$ by minimizing $f(\mathbf{x}_k + t_k \mathbf{d}_k)$ exactly, but this can be computationally expensive.
4. **What should be the stopping condition?** Stopping criteria ensure the algorithm terminates when a satisfactory solution is found. Common conditions include:
   - $\|\nabla f(\mathbf{x}_k)\| < \epsilon$ (small gradient norm).
   - $|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)| < \delta$ (small function value change).
   - Maximum number of iterations reached.
5. **Does the algorithm converge? If yes, how fast does it converge? Does convergence depend on $\mathbf{x}_0$?** Convergence depends on the function $f(\mathbf{x})$, the descent direction, and the step size strategy:
   - **Convex functions:** The algorithm generally converges to the global minimum.
   - **Non-convex functions:** The algorithm may converge to a local minimum or a saddle point.
   - **Rate of convergence:** First-order methods like gradient descent have linear convergence, while Newton's method can achieve quadratic convergence.
   - **Effect of $\mathbf{x}_0$:** A good initial guess can lead to faster convergence, especially in non-convex problems.

## References

[1] Introduction to Non-Linear Optimization by Amir Beck

[2] An Introduction to Optimization by Chong and Zak

[3] Optimization Methods Lecture Slides by Dr. Naresh Manwani