Notes: Optimization Methods | Week1&2

Runze Tian

September 16, 2025

1 Introduction to Optimization

1.1 Basic Concepts and Problem Formulation

Definition 1.1: Optimization

Optimization is the process of selecting a best element from a set of available alternatives, with regard to some criterion. The goal is to find a set of parameters or a model that minimizes or maximizes a certain objective function.

Optimization problems are central to fields like machine learning, operations research, and economics. We can formalize these problems into two primary categories.

Definition 1.2: Unconstrained Optimization

An unconstrained optimization problem is formulated as:

$$\min_{x \in \mathbb{R}^n} f(\boldsymbol{x})$$

Here, $\boldsymbol{x} \in \mathbb{R}^n$ is the **decision variable**, and $f : \mathbb{R}^n \to \mathbb{R}$ is the **objective function**. A solution $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$ is called a global minimizer. The value $f(\boldsymbol{x}^*)$ is the optimal value.

Definition 1.3: Constrained Optimization

A constrained optimization problem is formulated as:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & c_i(\boldsymbol{x}) = 0, \quad i \in \mathcal{E} \\ & c_i(\boldsymbol{x}) \geq 0, \quad i \in \mathcal{I} \end{aligned}$$

where $\mathcal{E}$ and $\mathcal{I}$ are index sets for the **equality constraints** and **inequality constraints**, respectively. The functions $c_i(\boldsymbol{x})$ are known as the constraint functions. Any maximization problem $\max f(\boldsymbol{x})$ can be converted to a minimization problem by considering $\min -f(\boldsymbol{x})$.

1.2 Classification of Optimization Problems

Optimization problems can be classified based on the nature of their variables and functions.

- **Continuous vs. Discrete Optimization**: If the decision variable $\boldsymbol{x}$ can take any real value (i.e., $\boldsymbol{x} \in \mathbb{R}^n$), the problem is continuous. If $\boldsymbol{x}$ is restricted to be an integer (i.e., $\boldsymbol{x} \in \mathbb{Z}^n$), the problem is discrete, often called combinatorial optimization, which is generally harder to solve.

- **Smooth vs. Non-smooth Optimization**: If the objective and constraint functions are continuously differentiable, the problem is smooth. Otherwise, it is a non-smooth problem.

- **Linear vs. Non-linear Programming**: If the objective function f and all constraint functions c_i are linear, the problem is a **Linear Program (LP)**. If the objective is quadratic and constraints are linear, it is a **Quadratic Program (QP)**. If any function is non-linear, it is a **Non-linear Program (NLP)**.

1.3 Convex Sets

The concept of convexity is fundamental to optimization theory because it allows us to make strong claims about the nature of optimal solutions.

Definition 1.4: Convex Set

A set $C \subseteq \mathbb{R}^n$ is said to be **convex** if for any two points $\boldsymbol{x}, \boldsymbol{y} \in C$ and any scalar $\theta \in [0,1]$, the line segment connecting them is also in C. That is,

$$\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y} \in C, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in C, \theta \in [0,1]$$

Geometrically, a set is convex if the line segment between any two of its points lies entirely within the set.

Proposition 1.1 (Properties of Convex Sets)**.** *Let C_1 and C_2 be convex sets in $\mathbb{R}^n$.*

1. *The intersection $C_1 \cap C_2$ is a convex set.*
2. *The Minkowski sum $C_1 + C_2 = \{\boldsymbol{x}_1 + \boldsymbol{x}_2 \mid \boldsymbol{x}_1 \in C_1, \boldsymbol{x}_2 \in C_2\}$ is a convex set.*

Proof. 1. **Proof of Intersection**: Let $\boldsymbol{x}, \boldsymbol{y} \in C_1 \cap C_2$. This implies $\boldsymbol{x}, \boldsymbol{y} \in C_1$ and $\boldsymbol{x}, \boldsymbol{y} \in C_2$. Since C_1 is convex, for any $\theta \in [0,1]$, $\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y} \in C_1$. Similarly, since C_2 is convex, $\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y} \in C_2$. Therefore, $\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y} \in C_1 \cap C_2$, proving the intersection is convex.
2. **Proof of Minkowski Sum**: Let $\boldsymbol{a}, \boldsymbol{b} \in C_1 + C_2$. Then by definition, $\boldsymbol{a} = \boldsymbol{x}_1 + \boldsymbol{x}_2$ and $\boldsymbol{b} = \boldsymbol{y}_1 + \boldsymbol{y}_2$ for some $\boldsymbol{x}_1, \boldsymbol{y}_1 \in C_1$ and $\boldsymbol{x}_2, \boldsymbol{y}_2 \in C_2$. For any $\theta \in [0,1]$, consider the point $\boldsymbol{z} = \theta\boldsymbol{a} + (1-\theta)\boldsymbol{b}$.

$$\boldsymbol{z} = \theta(\boldsymbol{x}_1 + \boldsymbol{x}_2) + (1-\theta)(\boldsymbol{y}_1 + \boldsymbol{y}_2) = \underbrace{(\theta\boldsymbol{x}_1 + (1-\theta)\boldsymbol{y}_1)}_{\in C_1} + \underbrace{(\theta\boldsymbol{x}_2 + (1-\theta)\boldsymbol{y}_2)}_{\in C_2}$$

Since C_1 and C_2 are convex, the two terms are in C_1 and C_2 respectively. Thus, $\boldsymbol{z} \in C_1 + C_2$, proving the sum is convex. □

1.4 Convex Functions

Definition 1.5: Convex Function

Let $C \subseteq \mathbb{R}^n$ be a convex set. A function $f : C \to \mathbb{R}$ is **convex** if for any $\boldsymbol{x}, \boldsymbol{y} \in C$ and any $\theta \in [0,1]$, the following inequality holds:

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y})$$

If the inequality is strict ($<$) for $\boldsymbol{x} \neq \boldsymbol{y}$ and $\theta \in (0,1)$, then f is **strictly convex**. Geometrically, the chord connecting any two points on the function's graph lies on or above the graph itself.

Theorem 1.1: First-Order Condition for Convexity

Let $f : C \to \mathbb{R}$ be a continuously differentiable function on an open convex set $C \subseteq \mathbb{R}^n$. Then f is convex if and only if for all $\boldsymbol{x}, \boldsymbol{y} \in C$:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$$

This means the first-order Taylor approximation of a convex function at any point provides a global underestimator of the function.

Proof. ($\Rightarrow$) Assume f is convex. By definition, for $\theta \in (0,1]$ and $\boldsymbol{x}, \boldsymbol{y} \in C$:

$$f(\theta\boldsymbol{y} + (1-\theta)\boldsymbol{x}) \leq \theta f(\boldsymbol{y}) + (1-\theta) f(\boldsymbol{x})$$

Rearranging gives:

$$f(\boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{x}) \leq \theta(f(\boldsymbol{y}) - f(\boldsymbol{x}))$$
$$\frac{f(\boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{x})}{\theta} \leq f(\boldsymbol{y}) - f(\boldsymbol{x})$$

Taking the limit as $\theta \to 0^+$, the left side becomes the definition of the directional derivative of f at $\boldsymbol{x}$ in the direction $\boldsymbol{y} - \boldsymbol{x}$, which is $\nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$. Thus, $\nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \leq f(\boldsymbol{y}) - f(\boldsymbol{x})$.
($\Leftarrow$) Assume $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in C$. Let $\boldsymbol{z} = \theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}$ for $\theta \in [0, 1]$. We apply the inequality twice:

$$f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^T(\boldsymbol{x} - \boldsymbol{z})$$
$$f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^T(\boldsymbol{y} - \boldsymbol{z})$$

Multiply the first inequality by θ and the second by $(1-\theta)$ and add them:

$$\begin{aligned}
\theta f(\boldsymbol{x}) + (1-\theta)f(\boldsymbol{y}) &\geq \theta f(\boldsymbol{z}) + (1-\theta)f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^T(\theta(\boldsymbol{x} - \boldsymbol{z}) + (1-\theta)(\boldsymbol{y} - \boldsymbol{z})) \\
&\geq f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^T(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y} - \boldsymbol{z}) \\
&\geq f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^T(\boldsymbol{z} - \boldsymbol{z}) = f(\boldsymbol{z})
\end{aligned}$$

This shows $f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \theta f(\boldsymbol{x}) + (1-\theta)f(\boldsymbol{y})$, so f is convex. □

Theorem 1.2: Second-Order Condition for Convexity

Let $f : C \to \mathbb{R}$ be a twice continuously differentiable function on an open convex set $C \subseteq \mathbb{R}^n$.

1. f is convex if and only if its Hessian matrix $\nabla^2 f(\boldsymbol{x})$ is positive semidefinite for all $\boldsymbol{x} \in C$.
2. If $\nabla^2 f(\boldsymbol{x})$ is positive definite for all $\boldsymbol{x} \in C$, then f is strictly convex.

Note: The converse of the second statement is not true. For example, $f(x) = x^4$ is strictly convex, but its second derivative at $x = 0$ is $f''(0) = 0$, which is not positive definite.

Proof. (Sketch of Part 1) ($\Rightarrow$) Assume f is convex. For any $\boldsymbol{x} \in C$ and direction $\boldsymbol{d}$, Taylor's theorem gives for small $t > 0$:

$$f(\boldsymbol{x} + t\boldsymbol{d}) = f(\boldsymbol{x}) + t\nabla f(\boldsymbol{x})^T\boldsymbol{d} + \frac{1}{2}t^2\boldsymbol{d}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{d} + o(t^2\|\boldsymbol{d}\|^2)$$

From the first-order condition, $f(\boldsymbol{x} + t\boldsymbol{d}) \geq f(\boldsymbol{x}) + t\nabla f(\boldsymbol{x})^T\boldsymbol{d}$. Combining these gives:

$$\frac{1}{2}t^2\boldsymbol{d}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{d} + o(t^2\|\boldsymbol{d}\|^2) \geq 0$$

Dividing by t^2 and letting $t \to 0$ yields $\boldsymbol{d}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{d} \geq 0$, which is the definition of positive semidefiniteness.
($\Leftarrow$) Assume $\nabla^2 f(\boldsymbol{x})$ is positive semidefinite. By Taylor's theorem with remainder:

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T\nabla^2 f(\boldsymbol{z})(\boldsymbol{y} - \boldsymbol{x})$$

for some $\boldsymbol{z}$ on the line segment between $\boldsymbol{x}$ and $\boldsymbol{y}$. Since $\nabla^2 f(\boldsymbol{z})$ is positive semidefinite, the last term is non-negative. Therefore, $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$, which implies f is convex by Theorem **??**. □

2 Fundamentals of Unconstrained Optimization

2.1 Optimality Conditions

Optimality conditions are mathematical statements that characterize solutions. They are essential for verifying if a point is a solution and for designing algorithms.

Definition 2.1: Local and Global Minima

Let $f : \mathbb{R}^n \to \mathbb{R}$.

- A point $\boldsymbol{x}^*$ is a **local minimizer** if there exists an $\epsilon > 0$ such that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x}$ with $\|\boldsymbol{x} - \boldsymbol{x}^*\| < \epsilon$.
- A point $\boldsymbol{x}^*$ is a **global minimizer** if $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$.
- If the inequalities are strict ($<$) for $\boldsymbol{x} \neq \boldsymbol{x}^*$, the minimizer is called **strict**.

For convex functions, any local minimizer is also a global minimizer.

Theorem 2.1: First-Order Necessary Condition (FONC)

If $\boldsymbol{x}^*$ is a local minimizer of f and f is continuously differentiable in an open neighborhood of $\boldsymbol{x}^*$, then the gradient at that point must be zero:

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0}$$

Proof. Proof by contradiction. Assume $\nabla f(\boldsymbol{x}^*) \neq \mathbf{0}$. Let $\boldsymbol{d} = -\nabla f(\boldsymbol{x}^*)$. Since f is differentiable, the directional derivative in direction $\boldsymbol{d}$ is $\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} = -\|\nabla f(\boldsymbol{x}^*)\|^2 < 0$. This means that for a small step $\alpha > 0$ in the direction $\boldsymbol{d}$, the function value will decrease. By Taylor's theorem:

$$f(\boldsymbol{x}^* + \alpha \boldsymbol{d}) = f(\boldsymbol{x}^*) + \alpha \nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} + o(\alpha) = f(\boldsymbol{x}^*) - \alpha \|\nabla f(\boldsymbol{x}^*)\|^2 + o(\alpha)$$

For a sufficiently small $\alpha > 0$, the negative linear term dominates the higher-order term, so $f(\boldsymbol{x}^* + \alpha \boldsymbol{d}) < f(\boldsymbol{x}^*)$. This contradicts the assumption that $\boldsymbol{x}^*$ is a local minimizer. Thus, we must have $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$. □

Theorem 2.2: Second-Order Necessary Condition (SONC)

If $\boldsymbol{x}^*$ is a local minimizer of f and $\nabla^2 f$ exists and is continuous in an open neighborhood of $\boldsymbol{x}^*$, then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and the Hessian matrix $\nabla^2 f(\boldsymbol{x}^*)$ is positive semidefinite.

Theorem 2.3: Second-Order Sufficient Condition (SOSC)

Suppose that $\nabla^2 f$ is continuous in an open neighborhood of $\boldsymbol{x}^*$. If $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and the Hessian matrix $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite, then $\boldsymbol{x}^*$ is a strict local minimizer of f.

Proof. Since $\nabla^2 f$ is continuous and $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite, there exists a radius $r > 0$ such that for any $\boldsymbol{x}$ with $\|\boldsymbol{x} - \boldsymbol{x}^*\| < r$, $\nabla^2 f(\boldsymbol{x})$ is also positive definite. For any such $\boldsymbol{x} \neq \boldsymbol{x}^*$, let $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{x}^*$. By Taylor's theorem, there is a $\boldsymbol{z}$ on the line segment between $\boldsymbol{x}^*$ and $\boldsymbol{x}$ such that:

$$f(\boldsymbol{x}) = f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} + \frac{1}{2} \boldsymbol{d}^T \nabla^2 f(\boldsymbol{z}) \boldsymbol{d}$$

Since $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{z})$ is positive definite (as $\boldsymbol{z}$ is within the radius r), we have:

$$f(\boldsymbol{x}) = f(\boldsymbol{x}^*) + \frac{1}{2} \boldsymbol{d}^T \nabla^2 f(\boldsymbol{z}) \boldsymbol{d} > f(\boldsymbol{x}^*)$$

The inequality is strict because $\boldsymbol{d} \neq \mathbf{0}$. Thus, $\boldsymbol{x}^*$ is a strict local minimizer. □

2.2 Structure of Iterative Methods

Most optimization algorithms are iterative. They generate a sequence of points $\{\boldsymbol{x}_k\}$ that ideally converge to a minimizer $\boldsymbol{x}^*$. The core idea is to move from the current point $\boldsymbol{x}_k$ to a new, better point $\boldsymbol{x}_{k+1}$.

The general structure of such a method is: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$, where:

- $\boldsymbol{d}_k$ is the **search direction**. It must be a descent direction.
- $\alpha_k > 0$ is the **step length** (or learning rate).

Definition 2.2: Descent Direction

A direction $\boldsymbol{d}_k$ is a **descent direction** from a point $\boldsymbol{x}_k$ if for a small step, the function value decreases. For a differentiable function, this is equivalent to the condition:

$$\nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k < 0$$

Geometrically, the descent direction must form an obtuse angle with the gradient vector.

There are two main strategies for choosing $\boldsymbol{d}_k$ and α_k:

- **Line Search Methods**: First, a descent direction $\boldsymbol{d}_k$ is chosen. Second, a step length α_k is found that minimizes f along that direction, i.e., solving $\min_{\alpha>0} f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$.
- **Trust Region Methods**: A model function (usually quadratic) is built to approximate f around $\boldsymbol{x}_k$ within a "trust region" of radius Δ_k. The direction and step length are determined simultaneously by minimizing the model within this region.

2.3 Convergence of Algorithms

Definition 2.3: Rate of Convergence

Let $\{\boldsymbol{x}_k\}$ be a sequence that converges to $\boldsymbol{x}^*$. The convergence is said to be:

- **Linear** if there is a constant $a \in (0, 1)$ such that $\lim_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|} = a$.
- **Superlinear** if the limit is $a = 0$.
- **Quadratic (or of order 2)** if there is a constant a such that $\lim_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2} = a$.

Quadratic and superlinear rates are much faster than linear convergence.

3 Line Search Methods

3.1 Step Length and Search Criteria

Given a descent direction $\boldsymbol{d}_k$, the one-dimensional subproblem is to find a step length $\alpha_k > 0$.

Definition 3.1: Exact Line Search

An **exact line search** finds the step length α_k that globally minimizes the one-dimensional function $\phi(\alpha) = f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ for $\alpha > 0$. This is equivalent to finding α_k such that $\phi'(\alpha_k) = 0$, which implies:

$$\nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)^T \boldsymbol{d}_k = 0$$

This means the gradient at the new point is orthogonal to the search direction. However, finding this exact minimum is often computationally expensive.

In practice, an **inexact line search** is used, which finds a step length that provides a sufficient decrease in the objective function without excessive computation. Simply requiring $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ is not enough to guarantee convergence to a minimizer, as the decrease might be negligible.

3.2 The Wolfe Conditions

The Wolfe conditions are a pair of inequalities that ensure both a sufficient decrease in the function value and that the step is not excessively short.

Definition 3.2: The Wolfe Conditions

For constants $0 < \rho < \sigma < 1$, a step length α_k satisfies the **Wolfe conditions** if the following two inequalities hold:

1. **Armijo Condition (Sufficient Decrease)**:

$$f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) \leq f(\boldsymbol{x}_k) + \rho \alpha_k \nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k$$

   This ensures the reduction in f is proportional to both the step length and the directional derivative.

2. **Curvature Condition**:

$$\nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)^T \boldsymbol{d}_k \geq \sigma \nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k$$

   This ensures the slope at the new point is less negative than the initial slope, preventing steps that are too short.

The **Strong Wolfe Conditions** replace the curvature condition with a stricter requirement:

$$|\nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)^T \boldsymbol{d}_k| \leq \sigma |\nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k|$$

This forces the step length to be closer to a stationary point of $\phi(\alpha)$.

Lemma 3.1 (Existence of Step Length for Wolfe Conditions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. If $\boldsymbol{d}_k$ is a descent direction at $\boldsymbol{x}_k$ and f is bounded below along the ray $\{\boldsymbol{x}_k + \alpha \boldsymbol{d}_k \mid \alpha > 0\}$, then there exist step lengths $\alpha > 0$ that satisfy the Wolfe conditions (for any $0 < \rho < \sigma < 1$).*

Proof. Let $\phi(\alpha) = f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$. Since $\boldsymbol{d}_k$ is a descent direction, $\phi'(0) = \nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k < 0$. The Armijo condition is $\phi(\alpha) \leq \phi(0) + \rho \alpha \phi'(0)$. The line $l(\alpha) = \phi(0) + \rho \alpha \phi'(0)$ has a negative slope. Since f is bounded below, $\phi(\alpha)$ is also bounded below. Thus, the line $l(\alpha)$ must eventually cross the graph of $\phi(\alpha)$, meaning there is a set of acceptable α values for the Armijo condition. Let $\alpha_{\max} = \sup\{\alpha \mid \phi(\alpha) \leq \phi(0) + \rho \alpha \phi'(0)\}$. Since ϕ is continuous, $\alpha_{\max}$ is well-defined and positive.
Now consider the curvature condition. By the Mean Value Theorem, for any $\alpha > 0$, there exists some $\xi \in (0, \alpha)$ such that

$$\phi'(\xi) = \frac{\phi(\alpha) - \phi(0)}{\alpha}$$

At $\alpha_{\max}$, we must have $\phi(\alpha_{\max}) = l(\alpha_{\max}) = \phi(0) + \rho \alpha_{\max} \phi'(0)$. Substituting this into the MVT expression (with $\alpha = \alpha_{\max}$), we get $\phi'(\xi) = \rho \phi'(0)$. Since $\rho < \sigma$ and $\phi'(0) < 0$, we have $\rho \phi'(0) > \sigma \phi'(0)$. Therefore,

$$\phi'(\xi) = \nabla f(\boldsymbol{x}_k + \xi \boldsymbol{d}_k)^T \boldsymbol{d}_k = \rho \phi'(0) > \sigma \phi'(0)$$

This shows that the step length $\xi \in (0, \alpha_{\max})$ satisfies the curvature condition. It also satisfies the Armijo condition since $\xi < \alpha_{\max}$ and $\phi'(\xi) < 0$, implying ϕ is still decreasing. Thus, a valid step length exists. □

3.3 Convergence of Line Search Methods

The Wolfe conditions are crucial for proving global convergence of line search methods. The following theorem, often attributed to Zoutendijk, is a cornerstone result.

Theorem 3.1: Zoutendijk's Theorem

Consider an iterative algorithm $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$, where $\boldsymbol{d}_k$ is a descent direction and α_k satisfies the Wolfe conditions. Suppose f is bounded below in $\mathbb{R}^n$ and is continuously differentiable in an open set containing the level set $\{\boldsymbol{x} \mid f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$, and that the gradient ∇f is Lipschitz continuous on this set. Then

$$\sum_{k \geq 0} \cos^2 \theta_k \|\nabla f(\boldsymbol{x}_k)\|^2 < \infty$$

where θ_k is the angle between the search direction $\boldsymbol{d}_k$ and the negative gradient $-\nabla f(\boldsymbol{x}_k)$.

Proof. From the second Wolfe condition:

$$(\nabla f_{k+1} - \nabla f_k)^T \boldsymbol{d}_k \geq (\sigma - 1)\nabla f_k^T \boldsymbol{d}_k$$

By the Mean Value Theorem, $(\nabla f_{k+1} - \nabla f_k) = \int_0^1 \nabla^2 f(\boldsymbol{x}_k + t\alpha_k \boldsymbol{d}_k)\alpha_k \boldsymbol{d}_k dt$. Using Lipschitz continuity of the gradient, $\|\nabla f_{k+1} - \nabla f_k\| \leq L\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| = L\alpha_k\|\boldsymbol{d}_k\|$. So, $L\alpha_k\|\boldsymbol{d}_k\|^2 \geq (\nabla f_{k+1} - \nabla f_k)^T \boldsymbol{d}_k \geq (\sigma - 1)\nabla f_k^T \boldsymbol{d}_k$. This gives a lower bound on the step size: $\alpha_k \geq \frac{\sigma-1}{L}\frac{\nabla f_k^T \boldsymbol{d}_k}{\|\boldsymbol{d}_k\|^2}$.
Now, sum the first Wolfe (Armijo) condition over all iterations:

$$f_{k+1} \leq f_k + \rho\alpha_k \nabla f_k^T \boldsymbol{d}_k$$

$$f_{N+1} - f_0 = \sum_{k=0}^{N}(f_{k+1} - f_k) \leq \rho \sum_{k=0}^{N} \alpha_k \nabla f_k^T \boldsymbol{d}_k$$

Since f is bounded below, as $N \to \infty$, the sum on the right must converge. Since $\rho > 0$ and $\nabla f_k^T \boldsymbol{d}_k < 0$, the series $\sum_{k=0}^{\infty} -\alpha_k \nabla f_k^T \boldsymbol{d}_k$ must converge. Substituting the lower bound for α_k:

$$\sum_{k=0}^{\infty} -\alpha_k \nabla f_k^T \boldsymbol{d}_k \geq \sum_{k=0}^{\infty} -\frac{1-\sigma}{L}\frac{(\nabla f_k^T \boldsymbol{d}_k)^2}{\|\boldsymbol{d}_k\|^2}$$

By definition, $\cos\theta_k = \frac{-\nabla f_k^T \boldsymbol{d}_k}{\|\nabla f_k\|\|\boldsymbol{d}_k\|}$. So $(\nabla f_k^T \boldsymbol{d}_k)^2 = \cos^2\theta_k \|\nabla f_k\|^2 \|\boldsymbol{d}_k\|^2$.

$$\sum_{k=0}^{\infty} \frac{1-\sigma}{L} \cos^2\theta_k \|\nabla f_k\|^2 < \infty$$

Since $(1 - \sigma)/L$ is a positive constant, the result follows. □

Corollary 3.1 (Convergence Corollary)**.** *If an algorithm produces search directions such that the angle θ_k is bounded away from $90°$ (i.e., $\cos\theta_k \geq \delta > 0$ for all k), then Zoutendijk's theorem implies that $\lim_{k\to\infty} \|\nabla f(\boldsymbol{x}_k)\| = 0$. This guarantees convergence to a stationary point. This condition on $\cos\theta_k$ is crucial and is satisfied by many algorithms, including the steepest descent and quasi-Newton methods.*

4 Trust Region Methods

Trust region methods are a powerful class of algorithms for unconstrained optimization. Unlike line search methods, which select a direction and then a step length, trust region methods build a local model of the objective function and restrict the search for the next iterate to a region around the current point where the model is considered reliable.

4.1 Basic Idea and Motivation

The core idea is: at each iteration, construct a (usually quadratic) model $q_k(\boldsymbol{d})$ of f near the current point $\boldsymbol{x}_k$, and only trust this model within a ball of radius Δ_k (the trust region). The next step $\boldsymbol{d}_k$ is chosen by (approximately) minimizing $q_k(\boldsymbol{d})$ subject to $\|\boldsymbol{d}\| \leq \Delta_k$.

Definition 4.1: Trust Region Subproblem

At iteration k, the trust region subproblem is:

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} \quad q_k(\boldsymbol{d}) = f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T \boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^T B_k \boldsymbol{d}$$

subject to $\|\boldsymbol{d}\| \leq \Delta_k$, where B_k is the Hessian $\nabla^2 f(\boldsymbol{x}_k)$ or its approximation.

4.2 Trust Region Models

The model $q_k(\boldsymbol{d})$ is typically quadratic, capturing local curvature information. The choice of B_k affects the algorithm's behavior:

- If $B_k = \nabla^2 f(\boldsymbol{x}_k)$, the model is locally accurate (Newton-type).
- If B_k is positive definite, the subproblem is easier to solve and ensures descent.
- If B_k is an approximation (e.g., quasi-Newton), the method is more robust for large-scale problems.

4.3 Solving the Trust Region Subproblem

The trust region subproblem is a constrained quadratic minimization. There are several solution strategies:

- **Cauchy Point**: Minimizes $q_k(\boldsymbol{d})$ along the steepest descent direction $-\nabla f(\boldsymbol{x}_k)$, clipped to the boundary of the trust region. Fast to compute, guarantees sufficient decrease.
- **Dogleg Method**: For positive definite B_k, combines the steepest descent direction and the Newton direction, choosing a step along a piecewise linear path (the "dogleg") within the trust region.
- **Exact Solution**: For small n, the subproblem can be solved exactly using eigenvalue decomposition or the Moré-Sorensen algorithm.
- **Truncated Conjugate Gradient**: For large-scale problems, an iterative method is used to approximately solve the subproblem.

Definition 4.2: Cauchy Point

The Cauchy point $\boldsymbol{d}_C$ is the minimizer of $q_k(\boldsymbol{d})$ along $-\nabla f(\boldsymbol{x}_k)$ within the trust region:

$$\boldsymbol{d}_C = -\tau_k \nabla f(\boldsymbol{x}_k)$$

where τ_k is chosen so that $\|\boldsymbol{d}_C\| = \min\left\{\frac{\|\nabla f(\boldsymbol{x}_k)\|^2}{\nabla f(\boldsymbol{x}_k)^T B_k \nabla f(\boldsymbol{x}_k)}, \Delta_k\right\}$.

Definition 4.3: Dogleg Method

The dogleg method constructs a path from the origin to the Cauchy point, then to the full Newton step. The step $\boldsymbol{d}_{DL}$ is chosen along this path such that $\|\boldsymbol{d}_{DL}\| \leq \Delta_k$.

4.4 Trust Region Radius Update

After computing $\boldsymbol{d}_k$, we evaluate how well the model predicted the actual reduction in f:

Definition 4.4: Trust Region Ratio

The ratio ρ_k is defined as

$$\rho_k = \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{d}_k)}{q_k(\boldsymbol{0}) - q_k(\boldsymbol{d}_k)}$$

It measures the agreement between the model and the true function.

The update strategy is:

- If $\rho_k \geq \eta_1$ (e.g., $\eta_1 = 0.75$), the model is good: accept the step and increase Δ_{k+1}.
- If $\eta_2 \leq \rho_k < \eta_1$ (e.g., $\eta_2 = 0.1$), accept the step, keep Δ_{k+1} unchanged.
- If $\rho_k < \eta_2$, reject the step, decrease Δ_{k+1}.

4.5 Global Convergence

Trust region methods can guarantee global convergence under mild assumptions:

Theorem 4.1: Global Convergence of Trust Region Methods

If f is bounded below and continuously differentiable, and the model q_k satisfies regularity conditions, then any limit point of the sequence $\{\boldsymbol{x}_k\}$ generated by the trust region method is a stationary point of f.