Numerical optimization : theory and applications

Ammar Mian
Associate professor, LISTIC, Université Savoie Mont Blanc

Outline

Unconstrained Optimization Context

Our Goal

Starting from an initial point $\mathbf{x}_0$, generate a sequence of iterates:

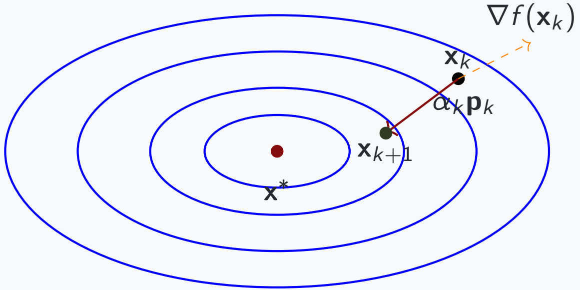
$$\{\mathbf{x}_k\}_{k=0}^{\infty}$$

such that $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ until convergence.

Key Question: How do we move from $\mathbf{x}_k$ to $\mathbf{x}_{k+1}$?

- Choose a direction $\mathbf{p}_k$
- Choose a step length α_k
- Update: $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$

Optimization on Level Sets



Level curves of $f(\mathbf{x})$

Line Search Strategy

Line Search Approach

1. Choose a search direction $\mathbf{p}_k$
2. Find step length α_k by approximately solving:

$$\min_{\alpha>0} f(\mathbf{x}_k + \alpha \mathbf{p}_k)$$

3. Update: $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$

Key insight: Fix direction first, then find distance.

- Exact line search is expensive and unnecessary
- Use inexact line search with appropriate conditions
- Generate limited number of trial step lengths

Trust Region Strategy

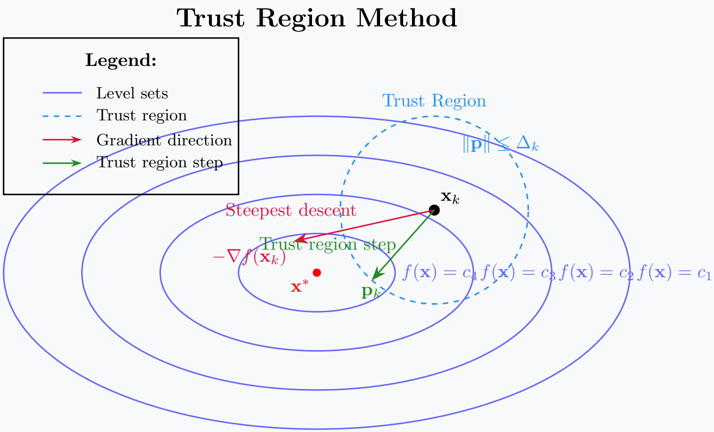
Trust Region Approach

1. Build local quadratic model $m_k(\mathbf{x}_k + \mathbf{p})$
2. Choose maximum distance Δ_k (trust region radius)
3. Solve: $\min_{\mathbf{p}} m_k(\mathbf{x}_k + \mathbf{p})$ subject to $\|\mathbf{p}\| \leq \Delta_k$
4. If step is successful, accept; otherwise shrink Δ_k

Key insight: Fix maximum distance first, then find best direction.

Quadratic model: $m_k(\mathbf{x}_k + \mathbf{p}) = f_k + \mathbf{p}^T \nabla f_k + \frac{1}{2}\mathbf{p}^T \mathbf{B}_k \mathbf{p}$

Visualization of Trust region



Trust region subproblem: $\min_{\mathbf{p}} m_k(\mathbf{p}) = f_k + \mathbf{g}_k^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \mathbf{B}_k \mathbf{p}$
subject to $\|\mathbf{p}\| \leq \Delta_k$

Line Search vs Trust Region

Aspect	Line Search	Trust Region
Order of choice	Direction $\rightarrow$ Distance	Distance $\rightarrow$ Direction
Search direction	Fixed per iteration	Changes when Δ_k changes
Step acceptance	Always accept	May reject and retry
Computational cost	Lower per iteration	Higher per iteration
Robustness	Good for well-scaled problems	Better for ill-conditioned

Focus of this lecture: Line search methods

Steepest Descent Direction

Theorem (Steepest Descent Direction)

The direction of steepest decrease is the solution to:

$$\min_{\mathbf{p}} \mathbf{p}^T \nabla f_k \quad \textit{subject to} \quad \|\mathbf{p}\| = 1$$

Solution: $\mathbf{p} = -\frac{\nabla f_k}{\|\nabla f_k\|}$

Proof sketch.

Using $\mathbf{p}^T \nabla f_k = \|\mathbf{p}\| \|\nabla f_k\| \cos\theta$:

- Minimize $\cos\theta$ subject to $\|\mathbf{p}\| = 1$
- Minimum occurs when $\cos\theta = -1$ (i.e., $\theta = \pi$)
- This gives $\mathbf{p} = -\nabla f_k / \|\nabla f_k\|$

□

General Descent Directions

Definition (Descent Direction)

A direction $\mathbf{p}_k$ is a **descent direction** if:

$$\mathbf{p}_k^T \nabla f_k < 0$$

Equivalently, the angle θ_k between $\mathbf{p}_k$ and $-\nabla f_k$ satisfies $\theta_k < \pi/2$.

Why Descent Directions Work

From Taylor expansion: $f(\mathbf{x}_k + \epsilon \mathbf{p}_k) = f(\mathbf{x}_k) + \epsilon \mathbf{p}_k^T \nabla f_k + O(\epsilon^2)$
If $\mathbf{p}_k^T \nabla f_k < 0$, then $f(\mathbf{x}_k + \epsilon \mathbf{p}_k) < f(\mathbf{x}_k)$ for sufficiently small $\epsilon > 0$.

The Step Length Tradeoff

Fundamental Challenge

We want to choose α_k to minimize $\phi(\alpha) = f(\mathbf{x}_k + \alpha \mathbf{p}_k)$, but:

- Exact minimization is too expensive
- Need substantial reduction in f
- Cannot spend too much time choosing α_k

Solution: Use inexact line search with appropriate termination conditions

- **Sufficient decrease:** Ensure adequate reduction in f
- **Curvature condition:** Prevent steps that are too short
- **Bracketing + interpolation:** Efficient implementation

Sufficient Decrease Condition (Armijo)

Definition (Armijo Condition)

$$f(\mathbf{x}_k + \alpha \mathbf{p}_k) \leq f(\mathbf{x}_k) + c_1 \alpha \nabla f_k^T \mathbf{p}_k$$

where $c_1 \in (0, 1)$ (typically $c_1 = 10^{-4}$).

- Ensures reduction proportional to step length and directional derivative
- Linear function $l(\alpha) = f(\mathbf{x}_k) + c_1 \alpha \nabla f_k^T \mathbf{p}_k$
- Since $c_1 < 1$, line $l(\alpha)$ lies above $\phi(\alpha)$ for small α
- **Problem:** Satisfied by all sufficiently small α

Curvature Condition

Definition (Curvature Condition)

$$\nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)^T \mathbf{p}_k \geq c_2 \nabla f_k^T \mathbf{p}_k$$

where $c_2 \in (c_1, 1)$.

Intuition:

- Left side is $\phi'(\alpha_k)$, right side is $c_2\phi'(0)$
- If slope $\phi'(\alpha)$ is strongly negative $\Rightarrow$ can reduce f more
- If slope is only slightly negative or positive $\Rightarrow$ terminate

Typical values:

- $c_2 = 0.9$ for Newton/quasi-Newton methods
- $c_2 = 0.1$ for conjugate gradient methods

Wolfe and Strong Wolfe Conditions

Wolfe Conditions

$$f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \leq f(\mathbf{x}_k) + c_1 \alpha_k \nabla f_k^T \mathbf{p}_k \quad (1)$$

$$\nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)^T \mathbf{p}_k \geq c_2 \nabla f_k^T \mathbf{p}_k \quad (2)$$

with $0 < c_1 < c_2 < 1$.

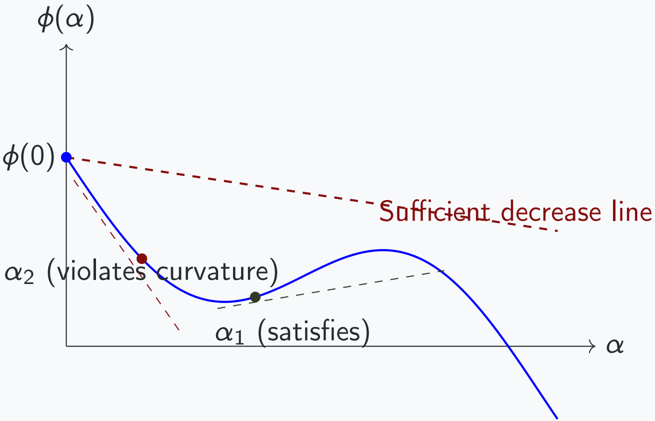
Strong Wolfe Conditions

Replace second condition with:

$$|\nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)^T \mathbf{p}_k| \leq c_2 |\nabla f_k^T \mathbf{p}_k|$$

Forces α_k to lie near stationary points of $\phi(\alpha)$.

Wolfe Conditions Illustration



Existence of Wolfe Step Lengths

Theorem (Existence Theorem)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, $\mathbf{p}_k$ is a descent direction, and f is bounded below along the ray $\{\mathbf{x}_k + \alpha \mathbf{p}_k \mid \alpha > 0\}$.
Then for $0 < c_1 < c_2 < 1$, there exist intervals of step lengths satisfying both the Wolfe conditions and the strong Wolfe conditions.

Key implications:

- Wolfe conditions are not too restrictive
- Always possible to find acceptable step lengths
- Line search algorithms are well-defined

Goldstein Conditions

Definition (Goldstein Conditions)

$$f(\mathbf{x}_k) + (1-c)\alpha_k \nabla f_k^T \mathbf{p}_k \leq f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \leq f(\mathbf{x}_k) + c\alpha_k \nabla f_k^T \mathbf{p}_k$$

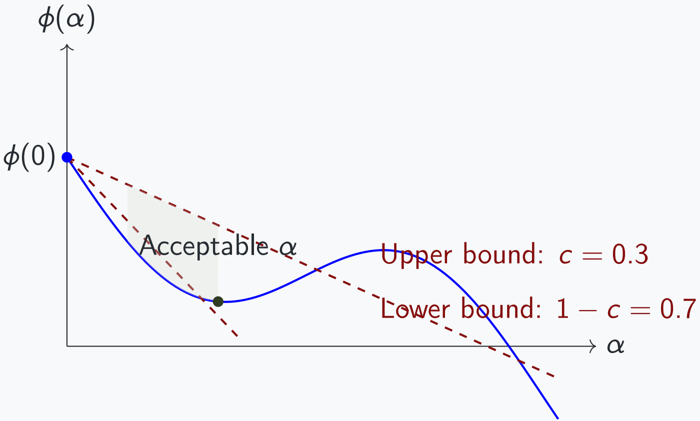
with $0 < c < \frac{1}{2}$.

- **Right inequality:** Sufficient decrease (same as Armijo)
- **Left inequality:** Controls step length from below
- Both conditions use same parameter c

Comparison with Wolfe:

- (Yes) Simpler (one parameter vs two)
- (No) May exclude minimizers of $\phi(\alpha)$
- (No) Not well-suited for quasi-Newton methods

Goldstein Conditions Illustration



Backtracking Line Search Algorithm

Algorithm Backtracking Line Search

Require: Choose $\bar{\alpha} > 0$, $\rho \in (0, 1)$, $c \in (0, 1)$

1: Set $\alpha \leftarrow \bar{\alpha}$
2: **while** $f(\mathbf{x}_k + \alpha \mathbf{p}_k) > f(\mathbf{x}_k) + c\alpha \nabla f_k^T \mathbf{p}_k$ **do**
3: $\alpha \leftarrow \rho\alpha$
4: **end while**
5: **return** $\alpha_k = \alpha$

Parameters:

- $\bar{\alpha} = 1$ for Newton/quasi-Newton methods
- $\rho \in [0.1, 0.8]$ (contraction factor)
- $c = 10^{-4}$ (sufficient decrease parameter)

Termination: Guaranteed in finite steps since α becomes small enough.

Properties of Backtracking

Key Properties

- **Simplicity:** Only uses sufficient decrease condition
- **Efficiency:** Cheap function evaluations
- **Robustness:** Always finds acceptable step
- **Flexibility:** Can use safeguarded interpolation for ρ

Why it works:

- Either accepts initial step $\bar{\alpha}$
- Or finds step short enough for sufficient decrease
- But not too short: within factor ρ of rejected step

Practical enhancement: Use polynomial interpolation to choose ρ adaptively.

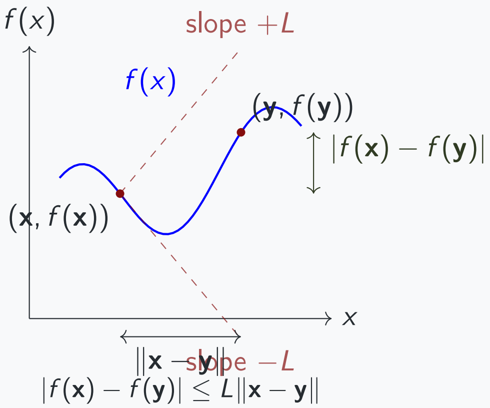
Lipschitz Continuous Functions

Definition (Lipschitz Continuity)

A function $f : \mathbb{R}^n \to \mathbb{R}$ is **Lipschitz continuous** on a set S if there exists a constant $L \geq 0$ such that:

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$$

for all $\mathbf{x}, \mathbf{y} \in S$.
The smallest such constant L is called the **Lipschitz constant**.



Key Properties:

- Lipschitz $\Rightarrow$ uniformly continuous
- Bounds the "steepness" of f
- If f is differentiable: $L = \sup \|\nabla f(\mathbf{x})\|$

Zoutendijk's Theorem

Theorem (Zoutendijk's Theorem)

Consider iterations $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$ where $\mathbf{p}_k$ is a descent direction and α_k satisfies the Wolfe conditions.
Suppose f is bounded below, continuously differentiable in a neighborhood of the level set $\mathcal{L} = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$, and ∇f is Lipschitz continuous on $\mathcal{L}$.
Then:

$$\sum_{k \geq 0} \cos^2 \theta_k \|\nabla f_k\|^2 < \infty$$

where $\cos \theta_k = \frac{-\nabla f_k^T \mathbf{p}_k}{\|\nabla f_k\| \|\mathbf{p}_k\|}$.

Global Convergence Result

Corollary (Global Convergence)

If our method for choosing the search direction $\mathbf{p}_k$ in the iteration ensures that the angle θ_k between the negative gradient ∇f_k and the search direction $\mathbf{p}_k$ is bounded away from 90°, there exists a $\delta > 0$ such that $\cos\theta_k \geq \delta > 0$ for all k.
Then:

$$\lim_{k\to\infty} \|\nabla f_k\| = 0$$

Applications:

- **Steepest descent:** $\cos\theta_k = 1$
- **Newton:** With proper modifications
- **Quasi-Newton:** With positive definite updates
- **Conjugate gradient:** With restarts

Note: "Global convergence" = convergence to stationary points, not global minima.

Rate of Convergence

Convergence Rates Depend on Method

- **Steepest descent:** Linear convergence, can be very slow
- **Newton's method:** Quadratic convergence near solution
- **Quasi-Newton:** Superlinear convergence
- **Conjugate gradient:** Finite termination (quadratic functions)

Key factors affecting rate:

- Condition number of the Hessian
- Choice of search direction
- Quality of line search
- Problem structure

Reference: Nocedal & Wright, Chapter 3, pages 47-51 for detailed analysis.

Exercice

Exercise: Line Search on Himmelblau Function

Consider the Himmelblau function:

$$f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$$

We want to find minima using a line search method with the following steps:

1. Implement a backtracking line search algorithm to find a step length α_k that satisfies the Wolfe conditions.
2. Use the steepest descent direction for the search direction $\mathbf{p}_k$.
3. Plot the convergence path on the level sets of f.