CS Space

Unconstrained Nonlinear Optimization Lecture 1: Fundamentals of Unconstrained Optimization

Feodor Pisnitchenko

CMCC/UFABC

Momoe Sakamori

IMECC/UNICAMP

Mathematical Background

- Vectors and matrices in $\mathbb{R}^n$
- Transpose, inner product, norm
- Eigenvalues of symmetric matrices
- Positive definite and semidefinite matrices
- Convergent sequences and subsequences
- Big-O and Little-o Notation
- Open, closed, and compact sets
- Continuity of functions
- 1st and 2nd order differentiability of functions
- Taylor series expansions
- Mean value theorem

Norm

A **norm** $\|\cdot\|$ on $\mathbb{R}^n$ is a mapping that assigns a scalar $\|x\|$ to every $x \in \mathbb{R}^n$ such that for all $x, y \in \mathbb{R}^n$ and $c \in \mathbb{R}$:

(a) $\|x\| \geq 0$ *(non-negativity)*.

(b) $\|c\,x\| = |c|\,\|x\|$ *(homogeneity)*.

(c) $\|x\| = 0$ if and only if $x = 0$ *(definiteness)*.

(d) $\|x + y\| \leq \|x\| + \|y\|$ *(triangle inequality)*.

For vector $x \in \mathbb{R}^n$, some common norms are:

Euclidean norm

$$\|x\| = \left(x^T x\right)^{1/2} = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$$

Maximum norm (sup-norm or l_∞-norm)

$$\|x\|_\infty = \max_i |x_i|$$

l_1-norm

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

Cauchy-Schwarz Inequality

For any two vectors $x, y \in \mathbb{R}^n$,

$$|x^T y| \leq \|x\| \, \|y\|,$$

with equality if and only if $x = \alpha y$ for some scalar α.

Matrix norms

Let $A \in \mathbb{R}^{m \times n}$ and $x \in \mathbb{R}^n$.

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

Common matrix norms:

1-*norm*

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |A_{ij}|$$

2-*norm*

$$\|A\|_2 = \text{largest eigenvalue of } (A^T A)^{1/2}$$

∞-norm

$$\|A\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} |A_{ij}|$$

Frobenius norm

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2 \right)$$

Condition Number

If A is a nonsingular matrix, the **condition number** of A (with respect to a given norm) is defined as

$$\kappa(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{\max}}{\sigma_{\min}},$$

where $\sigma_{\max}$ and $\sigma_{\min}$ denote the largest and smallest singular values of A, respectively.

Positive Definite Matrices

Let A be a symmetric $n \times n$ matrix.

Positive definite matrix

A is called positive definite if

$$x^T A x > 0, \text{for all} x \in \mathbb{R}^n.$$

Positive semidefinite or nonnegative definite matrix

A is called positive semidefinite or nonnegative definite if

$$x^T A x \geq 0, \text{for all} x \in \mathbb{R}^n.$$

Convergent Sequences

A sequence $\{x_k\}$ converge to some point x if for any $\epsilon > 0$ there exists some K such that $\|x_k - x\| \leq \epsilon$ for every $k \geq K$. Symbolically, $x_k \to x$ or $\lim_{k\to\infty} x_k = x$.

Accumulation point or limit point

We say that that $\bar{x} \in \mathbb{R}^n$ is an accumulation point or limit point for $\{x_k\}$ if there is an infinite set of indices $k_1, k_2, \ldots$ such that the subsequence $\{x_{k_i}\}$ converges to $\bar{x}$, i.e.

$$\lim_{i\to\infty} x_{k_i} = \bar{x}.$$

Alternatively, we say that for any $\epsilon > 0$ and all positive integers K, we have

$$\|x_k - x\| \leq \epsilon, \text{ for some } k \geq K.$$

Cauchy Sequences

Definition

A sequence $\{x_k\}$ is called a *Cauchy sequence* if for every $\epsilon > 0$, there exists an integer K such that

$$\|x_k - x_l\| < \epsilon \quad \text{for all } k, l \geq K.$$

In a **complete** metric space (e.g., $\mathbb{R}^n$), a sequence converges if and only if it is a Cauchy sequence.

Convergence Rates (I)

Let $\{x_k\}$ be a sequence that converges to x^*.

The convergence is **linear** if there exists a constant $0 < r < 1$ and an index N such that

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \leq r \quad \text{for all } k \geq N\,.$$

It is **superlinear** if

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0\,.$$

Convergence Rates (II)

The convergence is **quadratic** if there exists a constant $M > 0$ and an index N such that

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} \leq M \quad \text{for all } k \geq N\,.$$

In general, the **order** of convergence is $p > 1$ if there exists a constant $M > 0$ and an index N such that

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} \leq M \quad \text{for all } k \geq N\,.$$

Big-O and Little-o Notation

In the context of convergence analysis, we often use *Big-O* and *little-o* notation to describe how fast certain quantities go to zero (or grow) as $k \to \infty$.

Big-O Notation

Big-O

We write $g(k) = O(h(k))$ as $k \to \infty$ if there exist constants $C > 0$ and k_0 such that

$$|g(k)| \leq C\,|h(k)| \quad \text{for all } k \geq k_0.$$

Intuitively, $g(k)$ grows *at most as fast* as $h(k)$ up to a constant factor.

Example: If $x_{k+1} - x^* = O(\|x_k - x^*\|^2)$, it means the error shrinks *at least on the order* of the square of the previous error.

Little-o Notation

Little-o

We write $g(k) = o(h(k))$ as $k \to \infty$ if for every constant $\epsilon > 0$, there exists k_1 such that

$$|g(k)| \leq \epsilon |h(k)| \quad \text{for all } k \geq k_1.$$

Equivalently,

$$\lim_{k \to \infty} \frac{g(k)}{h(k)} = 0.$$

In other words, $g(k)$ grows *strictly slower* than $h(k)$ as $k \to \infty$.

Big-O and Little-o in Convergence Rates

- **Convergence rates** are often expressed using Big-O notation:

$$\|x_{k+1} - x^*\| = O(\|x_k - x^*\|^p).$$

- When analyzing asymptotic behavior, we use little-o notation to mean a *faster* decrease. For instance:

$$\|x_{k+1} - x^*\| = o(\|x_k - x^*\|)$$

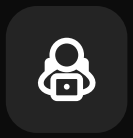
  implies that the sequence converges *superlinearly*.

- In many proofs, you will see limits such as

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} = \text{constant} < \infty,$$

  which is closely related to the $O(\cdot)$ concept.

Unconstrained Optimization



Nonlinear Programming Problem

$$\min_{x \in X} f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function of n variables.

- If $X = \mathbb{R}^n$, the problem is called unconstrained.
- If f is linear and X is polyhedral, the problem is a linear programming problem.

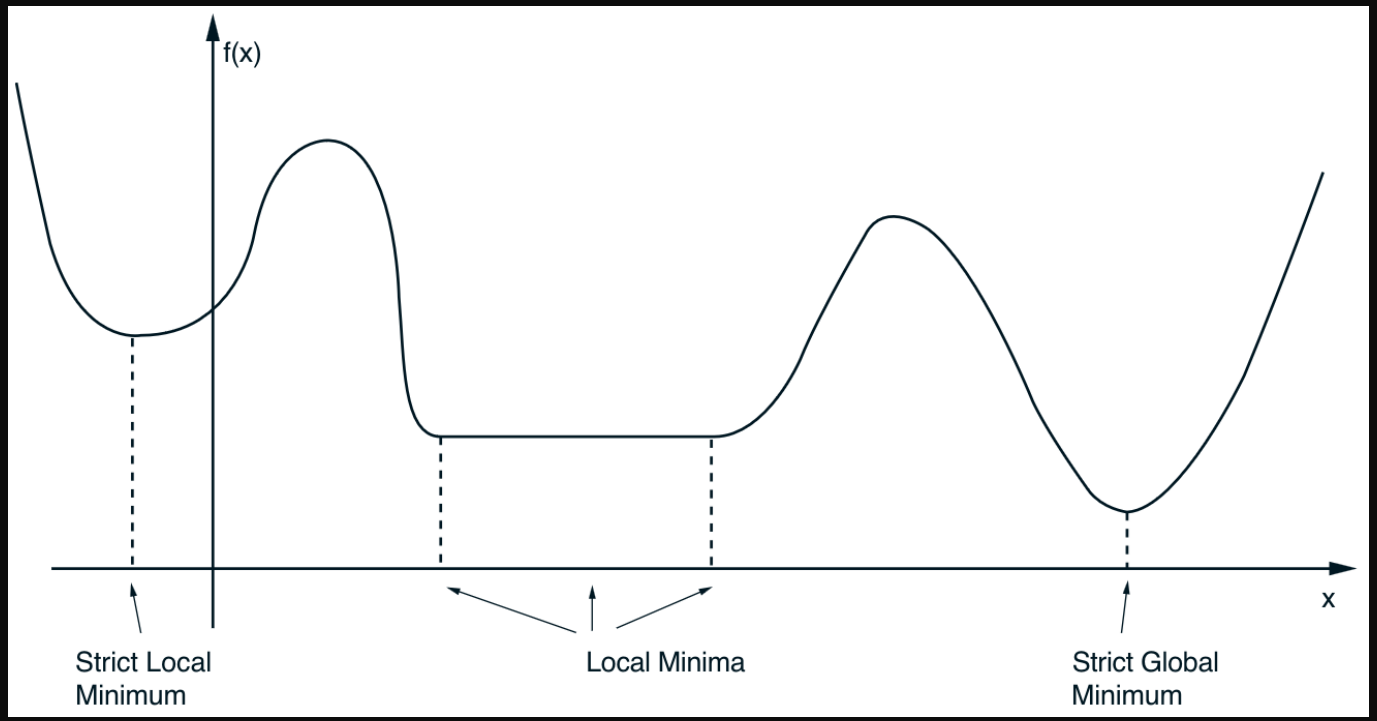
Unconstrained Optimization

Unconstrained Nonlinear Optimization Problem

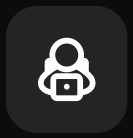
$$\min_x f(x),$$

where $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$.

Global and local optima



Global and local optima



Global

- x^* is a **global minimizer** if $f(x^*) \leq f(x)$ for all x.
- x^* is a **strict global minimizer** if $f(x^*) < f(x)$ for all x.

Local

- x^* is a **local minimizer** if there is a neighborhood N of x^* such that $f(x^*) \leq f(x)$ for all $x \in N$.
- x^* is a **strict local minimizer** if there is a neighborhood N of x^* such that $f(x^*) < f(x)$ for all $x \in N$.

Taylor's Theorem

Taylor's Theorem: Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable in a neighborhood of x, and let $p \in \mathbb{R}^n$. Then there exists some $t \in (0, 1)$ such that

$$f(x + p) = f(x) + \nabla f(x + t\,p)^T p\,.$$

Moreover, if f is twice continuously differentiable, then there exists some $t \in (0, 1)$ such that

$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{1}{2}\,p^T\big(\nabla^2 f(x + t\,p)\big)\,p\,.$$

First-Order Necessary Conditions

Theorem: If x^* is a local minimizer of a continuously differentiable function f, then $\nabla f(x^*) = 0$.

Proof (by contradiction): Assume $\nabla f(x^*) \neq 0$. Let $p = -\nabla f(x^*)$. Then

$$p^T \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0\,.$$

By continuity of ∇f, there exists $T > 0$ such that $p^T \nabla f(x^* + t\,p) < 0$ for all $t \in [0, T]$. For any $0 < \bar{t} \leq T$, Taylor's theorem (on $[0, \bar{t}]$) yields some $\xi \in (0, \bar{t})$ for which

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}\,p^T \nabla f(x^* + \xi p) < f(x^*)\,.$$

This contradicts the local minimality of x^*.

Second-Order Necessary Conditio

Theorem: If x^* is a local minimizer of f and $\nabla^2 f$ exists and is continuous in a neighborhood of x^*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

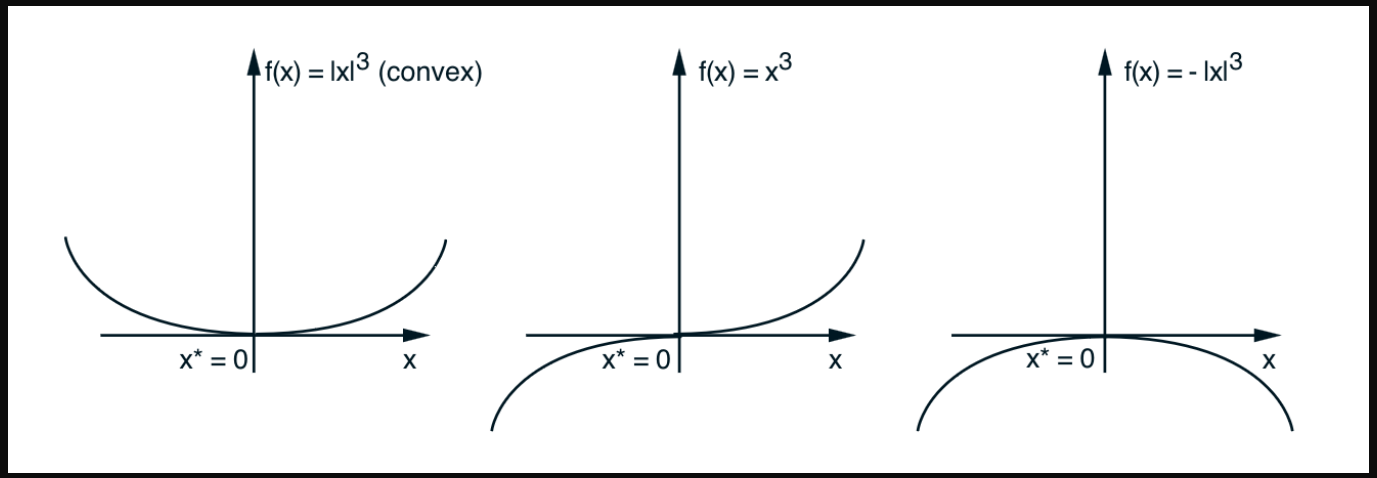
Proof (by contradiction): From the previous result, $\nabla f(x^*) = 0$. Now assume $\nabla^2 f(x^*)$ is *not* positive semidefinite. Then there exists a vector p such that $p^T \nabla^2 f(x^*)\, p < 0$. By continuity of $\nabla^2 f$, there is a $\delta > 0$ such that $p^T \nabla^2 f(x^* + t\, p)\, p < 0$ for all $0 < t \leq \delta$. For sufficiently small $\bar{t} \in (0, \delta]$, using the second-order Taylor expansion around x^* (with some $\xi \in (0, \bar{t})$) gives

$$f(x^* + \bar{t}p) = f(x^*) + \frac{1}{2}\, \bar{t}^2\, p^T \left(\nabla^2 f(x^* + \xi p)\right) p < f(x^*)\,,$$

contradicting the local minimality of x^*.

Stationary Points

Note: There may exist points (stationary points) that satisfy the first- and second-order necessary conditions but are not local minimizers. (Thus, these conditions are not sufficient for optimality.)



Second-Order Sufficient Condition

Theorem: Suppose $\nabla^2 f$ is continuous in a neighborhood of x^*. If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite, then x^* is a *strict* local minimizer of f.

Proof (sketch): Since $\nabla^2 f(x^*)$ is positive definite, there exists $\lambda > 0$ such that $p^T \nabla^2 f(x^*)\, p \geq \lambda \|p\|^2$ for all p. By continuity, we can find a neighborhood $D = \{z \mid \|z - x^*\| < r\}$ where $\nabla^2 f(x)$ remains positive definite for all $x \in D$. For any $x^* + p$ in this neighborhood, using Taylor's theorem (with $\nabla f(x^*) = 0$) we get

$$f(x^* + p) = f(x^*) + \frac{1}{2} p^T \big(\nabla^2 f(x^* + \xi p)\big)\, p\,,$$

for some $\xi \in (0, 1)$. Because $\nabla^2 f(x^* + \xi p)$ is positive definite, the quadratic term above is > 0 for all $p \neq 0$. Hence, $f(x^* + p) > f(x^*)$ for all small $p \neq 0$, which shows that x^* is a local minimizer.

Theorem

I. If f is convex, then any local minimizer x^* is a global minimizer of f.

II. Moreover, if f is differentiable and convex, then any stationary point x^* is a global minimizer of f.

Proof (I)

Suppose f is convex and x^* is a local minimizer but not global. Let z be such that $f(z) < f(x^*)$. By convexity of f, for all $\lambda \in (0, 1)$

$$\begin{aligned} f(\lambda z + (1-\lambda)x^*) &\leq \lambda f(z) + (1-\lambda) f(x^*) \\ &= \lambda \underbrace{[f(z) - f(x^*)]}_{<0} + f(x^*) \\ &< f(x^*) \end{aligned}$$

Thus, f takes values strictly lower than $f(x^*)$ on the line segment connecting z with x^*. Hence, x^* is not a local minimizer.

Proof (II)

Suppose f is convex and differentiable, and let x^* be a stationary point of f (so $\nabla f(x^*) = 0$). Assume for contradiction that x^* is not a global minimizer. Then there exists some z such that $f(z) < f(x^*)$. By convexity,

$$f(\lambda z + (1-\lambda)x^*) \;\leq\; \lambda f(z) + (1-\lambda) f(x^*)\,,$$

for all $\lambda \in [0, 1]$. Taking $\lambda \to 0^+$, we get

$$\nabla f(x^*)^T (z - x^*) \leq f(z) - f(x^*) < 0\,,$$

which contradicts $\nabla f(x^*) = 0$. Therefore, x^* must be a global minimizer of f.

Algorithm

We can summarize the main steps of an optimization algorithm as follows:

- Choose an initial point x_0.
- Generate a sequence of iterates $\{x_k\}_{k=0}^{\infty}$.
- At each iteration k, compute a new iterate x_{k+1} such that

$$f(x_{k+1}) < f(x_k)$$

  i.e., the objective value decreases.
- Two primary strategies to choose x_{k+1} are:
  - **Line search methods**
  - **Trust region methods**

Line Search Algorithms

- **Choose a descent direction** p_k at each iteration k.
- **Obtain the next iterate** by moving from the current point x_k in the direction p_k:

$$x_{k+1} = x_k + \alpha\, p_k,$$

where $\alpha > 0$ is the *step length*.

- **Approximate the optimal step length** by solving (inexactly, in practice):

$$\min_{\alpha>0} \; f(x_k + \alpha\, p_k).$$

  - Common techniques for determining α include:
    - **Backtracking line search**
    - **Zoom (More–Thuente) line search**

Trust Region Algorithms

- **Construct a local model:** Build a model function m_k that approximates f near the current iterate x_k.
- **Trust region subproblem:** Find a step p that (approximately) solves

$$\min_p \; m_k(x_k + p) \quad \text{subject to} \quad \|p\|_2 \leq \Delta,$$

  where $\Delta > 0$ is the *trust-region radius*.
- **Accept or reject the step:**
  - Compute the ratio of the actual decrease in f to the predicted decrease by m_k.
  - If the ratio is sufficiently high (indicating the model is accurate), accept the step: $x_{k+1} = x_k + p$.
  - Otherwise, reduce Δ (shrink the trust region) and re-solve the subproblem.
- **Iterate this process** until a convergence criterion is met.

Typical Trust Region Model

A common choice for m_k is a **quadratic approximation** around x_k:

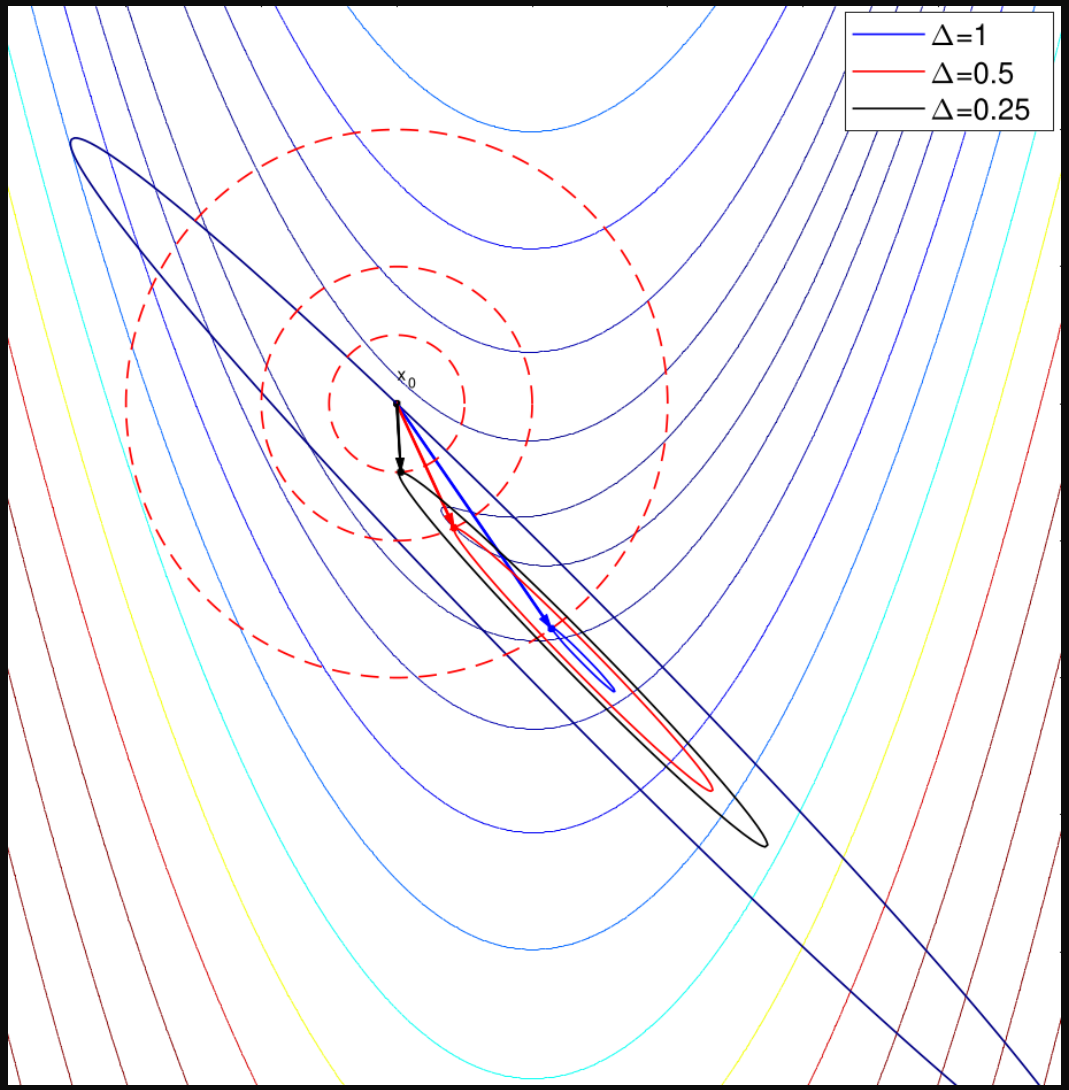
$$m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p,$$

where

- $f_k = f(x_k)$,
- $\nabla f_k = \nabla f(x_k)$,
- $B_k \approx \nabla^2 f(x_k)$ (the Hessian or an approximation).

The **trust region constraint** $\|p\|_2 \leq \Delta$ ensures p stays in the region where m_k is a good local approximation of f.

Trust Region Algorithms



Three possible trust regions (circles) and their corresponding steps p_k.