UNIVERSITY OF WATERLOO
Faculty of Mathematics

CO 367 Nonlinear Programming

Instructor: Fei Wang
Notes: Fei Wang, Juxin Fa

Fall 2021

Contents

1 Introduction

1.1 Mathematical Optimization

A mathematical optimization problem, or just optimization problem, has the form

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & c_i(x) = 0, i = 1, \ldots, m_1. \\ & c_i(x) \leq 0, i = m_1 + 1, \ldots, m. \end{array} \tag{1.1}$$

Here the vector $x = (x_1, \ldots, x_n)$ is the optimization variable of the problem, the function $f : \mathbb{R}^n \to \mathbb{R}$ is the object function, the function $c_i : \mathbb{R}^n \to ,i = 1, c \ldots, n$ are the constraint functions (equality constraints and inequality constraints).

Applications include Operational research (scheduling, supply chain, transportation), Image processing (matrix completion, compressed sensing), Statistics (deep learning, maximal likely-hood estimate), Finance (risk control), Biology, Control ...

1.2 Constrained and Unconstrained Optimization

If we have $m = 0$ constraints, then (1.1) is an unconstrained optimization problem. Otherwise it is a constrained optimization problem.

1.3 Feasible Region

In mathematical optimization, a feasible region, feasible set, search space, or solution space is the set of all possible points that satisfy the problem's constraints, i.e.,

$$\mathcal{X} = \{x \in \mathbb{R}^n \mid c_i(x) \leq 0, i = 1, ..., m, c_i(x) = 0, i = m + 1, ..., m + l\}$$

If $\mathcal{X}$ is empty, then (1.1) is infeasible, otherwise it is feasible.

For unconstrained optimization problem, the feasible region is $\mathcal{X} = \mathbb{R}$.

1.4 Optimal value

Notice that in $\mathcal{X}$, it is possible that the minimum or maximal value of the object function f may not exist. However, the infimum and supremum should always exists. Thus, when the maximal value or the minimal value does not exist, we will replace $min(max)$ with $inf(sup)$.

$$p^* = \inf\{f(x) \mid x \in \mathcal{X}\}$$

Special cases:

- $p^* = \infty$ if problem is infeasible.
- $p^* = -\infty$ if problem is unbounded.

1.5 Global and local minimizer

DEFINITION 1.1. (Open Ball & Closure)

The open ball of radius δ around $\bar{x}$ is $B_\delta(\bar{x}) = \{x \in \mathbb{R}^n, \|x - \bar{x}\| < \delta\}$. The closure of $B_\delta(\bar{x})$ is $\overline{B}_\delta(\bar{x}) = \{x \in \mathbb{R}^n, \|x - \bar{x}\| \leq \delta\}$

DEFINITION 1.2. (Minimizer)

Consider $f : \mathcal{X} \to R$. A point x^* is

- a global minimizer for f on $\mathcal{X}$ if $f(x^*) \leq f(x)$, $\forall x \in \mathcal{X}$
- a strict global minimizer for f on $\mathcal{X}$ if $f(x^*) < f(x)$, $\forall x \in \mathcal{X}, x \neq x^*$
- a local minimzer for f on $\mathcal{X}$ if there exists $\delta > 0$ such that $f(x^*) \leq f(x)$, $\forall x \in B_\delta(x^*) \cap \mathcal{X}$.
- a strict local minimizer for f on $\mathcal{X}$ if there exists $\delta > 0$ such that $f(x^*) < f(x)$, $\forall x \in B_\delta(x^*) \cap \mathcal{X}, x \neq x^*$.

1.6 Example

As a simple example, consider the problem

$$\begin{array}{ll} \min & (x_1 - 2)^2 + (x_2 - 1)^2 \\ \text{s.t.} & x_1^2 - x_2 \leq 0 \\ & x_1 + x_2 \leq 2 \end{array} \tag{1.2}$$

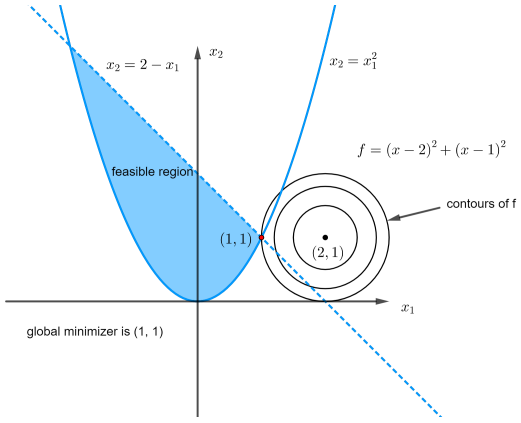


Figure 1: global minimizer

1.7 Continuous versus discrete optimization

In continuous optimization problems, variables are allowed to vary continuously (real numbers). In discrete optimization problems, the variables could be integers, binary variables or more abstract objects such as permutations of an order set. The key feature for discrete optimization problem is that feasible region is a finite set.

Continuous optimization problems are normally easier to solve because we can use information from a particular point x to study neighborhood points of x. In discrete optimization, the behaviour can change significantly from point to point.

1.8 Stochastic and deterministic optimization

Randomness appears in the problem formulation (random objective functions or random constraints).

Not covered in this course, we focus deterministic optimization problems.

1.9 Optimization algorithms

Optimization algorithms are iterative. They begin with an initial value of the variable x and generate a sequence of improved estimate. Hopefully they terminate in finitely many steps, such that either the last point is an optimal solution. or the limit of the sequence is an optimal solution.

Given an initial value x^0, denote the sequence of points generated by iterative algorithms as $\{x^k\}$. If $\{x^k\}$ satisfy $\lim_{k\to\infty} ||x^k - x^*|| = 0$, and x^* is a local(global) minimizer, then we say the algorithm (sequence) converges to a local(global) optimal solution.

1.9.1 Convergence rate

Let $\{x^k\}$ be a sequence in $\mathbb{R}^n$ that converges to x^*, we have the following definition of the convergence rate

DEFINITION 1.3. (Rate of Convergence (Quotient))

- The convergence is **Q-sublinear** if for k sufficiently large

$$\lim_{k\to\infty} \frac{\left\|x^{k+1} - x^*\right\|}{\|x^k - x^*\|} = 1.$$

- The convergence is **Q-linear** if for k sufficiently large

$$\lim_{k\to\infty} \frac{\left\|x^{k+1} - x^*\right\|}{\|x^k - x^*\|} \leq a, a \in (0,1).$$

- The convergence is **Q-superlinear** if for k sufficiently large

$$\lim_{k\to\infty} \frac{\left\|x^{k+1} - x^*\right\|}{\|x^k - x^*\|} = 0.$$

- The convergence is **Q-quadratic** if for k sufficiently large

$$\lim_{k\to\infty} \frac{\left\|x^{k+1} - x^*\right\|}{\|x^k - x^*\|^2} \leq a, a > 0.$$

1.10 Convexity

The optimization problem (1.1) is called a linear problem if the objective and constraint functions $f_0, ..., f_m$ are linear, i.e., it satisfies

$$f_i(\alpha x + \beta y) = \alpha f_i(x) + \beta f_i(x) \tag{1.3}$$

for all $x, y \in \mathbb{R}^n$ and all $\alpha, \beta \in \mathbb{R}$. If the optimization is not linear, it is called a nonlinear program which will be the main subject of this course.

An important class of optimization problem is called *convex optimization problems*. A convex optimization problem is one in which the objective and constraint functions are convex, which means they satisfy the inequality

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y) \tag{1.4}$$

for all $x, y \in \operatorname{dom} f_i$ and all $\alpha, \beta \in \mathbb{R}$ with $\alpha + \beta = 1$, $\alpha \geq 0, \beta \geq 0$.

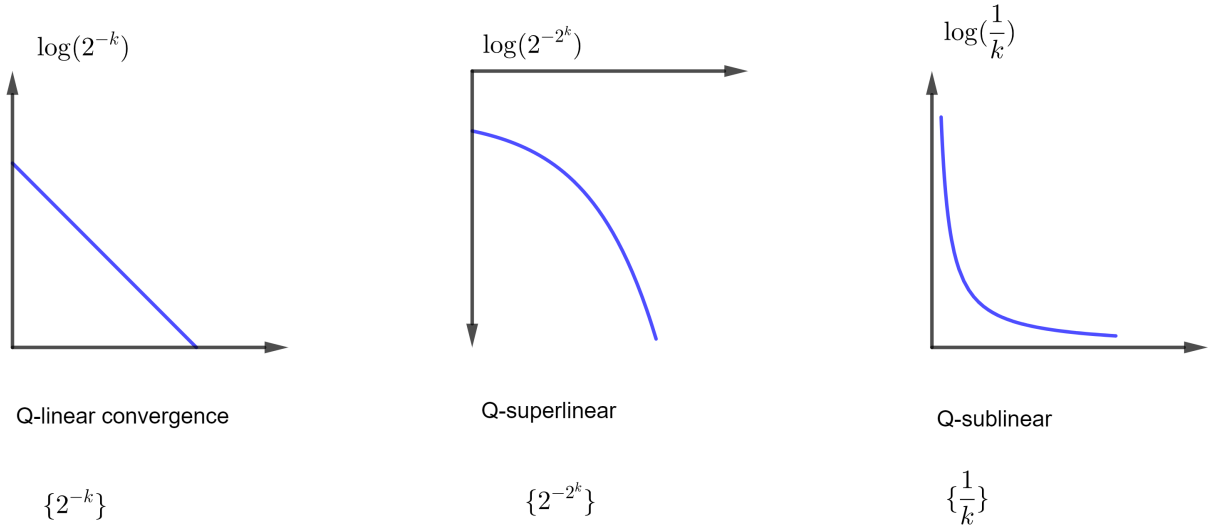


Figure 2: Comparison of different convergence rates

Convexity is more general than linearity. Any linear problem is a convex problem, but a convex problem is not necessarily linear (nonlinear).

2 Linear Algebra

2.1 Vector and Matrix Norm

DEFINITION 2.1. (Norm)

A norm $\|\cdot\|$ assigns a scalar $\|x\|$ to every $x \in \mathbb{R}^n$ such that

- $\|x\| \geq 0, \forall x \in \mathbb{R}^n$ and $\|x\| = 0 \iff x = 0$.
- $\|c \cdot x\| = |c| \cdot \|x\|$ for all $c \in \mathbb{R},\ x \in \mathbb{R}^n$.
- $\|x + y\| \leq \|x\| + \|y\|$.

Examples of norm

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

2-norm: $\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$

p-norm or ℓ^p norm: $\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$.

Infinite norm: $\|x\|_\infty = \max |x_i|$

THEOREM 2.2. (Schwartz Inequality)

For any $x, y \in \mathbb{R}^n$, we have $|x^T y| \leq \|x\|_2 \|y\|_2$

Proof. $|x^T y| = \|x\|_2 \|y\|_2 \cos(\theta) \leq \|x\|_2 \|y\|_2$ □

THEOREM 2.3. (Pythagorean)

If $x, y \in \mathbb{R}$ are orthogonal, then $\|x + y\|_2^2 = \|x\|_2^2 + \|y\|_2^2$

2.1.1 Induced Norm

DEFINITION 2.4. (Induced Norm)

Given a vector norm $\|\cdot\|$, the induced matrix norm assigns a scalar $\|A\|$ to every $A \in \mathbb{R}^{n \times n}$ with

$$A = \max_{\|x\|=1} \|Ax\|$$

.

Proposition 2.5. $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|=\|y\|=1} |y^T Ax|$

Proof. Apply Schwartz Inequality to $|y^T Ax|$, we get $\max_{\|x\|_2=1} \|Ax\|_2 \geq \max_{\|x\|=\|y\|=1} |y^T Ax|$

For the other direction, from the definition of induced norm, we know there exists x such that $\|x\|_2 = 1$ and $\|Ax\|_2 = \|A\|_2$.

Let $y = \frac{Ax}{\|Ax\|_2}$, then

$$|y^T Ax| = \frac{|x^T A^T Ax|}{\|Ax\|_2} = \frac{\|Ax\|_2^2}{|\,\|Ax\|_2} = \|Ax\|_2 = \|A\|_2 .$$

Therefore,

$\max_{\|x\|_2=1} \|Ax\|_2 \leq \max_{\|x\|=\|y\|=1} |y^T Ax|$ □

Proposition 2.6. $\|A\|_2 = \left\|A^T\right\|_2$.

Proof. Swap x and y in Proposition (2.5). □

Lemma 2.7. Let $A \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$, for any induced norm $\|\cdot\|$, then

$$\|Ax\| \leq \|A\| \|x\| .$$

Proposition 2.8. Let $A \in \mathbb{R}^{n\times n}$, then

$$\|A\|_2^2 = \left\|AA^T\right\|_2 = \left\|A^T A\right\|_2 .$$

2.2 Eigenvalues

DEFINITION 2.9. (Eigenvalue and Eigenvector)

The characteristic polynomial ϕ of a matrix $A \in \mathbb{R}^{n\times n}$ is defined as

$$\phi(\lambda) = \det(A - \lambda I).$$

The roots of ϕ are eigenvalues of A. The eigenvector x corresponding to an eigenvalue λ is $\{x \in \mathbb{R}^n | Ax = \lambda x\}$.

Proposition 2.10. Given a matrix $A \in \mathbb{R}^{n\times n}$, we have the following:

1. A is singular $\iff$ A has a zero eigenvalue.
2. If $S \in \mathbb{R}^{n\times n}$ is nonsingular and $B = SAS^{-1}$, then A, B has the same eigenvalues.
3. If the eigenvalues of A are $\lambda_1, ..., \lambda_n$, then
   - the eigenvalues of $A + cI$ are $\lambda_1 + c, ..., \lambda_n + c$.
   - the eigenvalues of A^k are $\lambda_1^k, ..., \lambda_n^k$.
   - the eigenvalues of A^{-1} are $\lambda_1, ..., \lambda_n$.
   - the eigenvalues of A^T is are also $\lambda_1, ..., \lambda_n$.
   - the algebraic multiplicity of an eigenvalue λ is greater or equal to the geometric multiplicity.

DEFINITION 2.11. (Spectral Radius)

The spectral radius $\rho(A)$ of $A \in \mathbb{R}^{n\times n}$ is

$$\max\{|\lambda| \mid \lambda \text{ is eigenvalue of } A\}$$

Proposition 2.12. For any induced norm $\|\cdot\|$, we have $\rho(A) \le \|A^k\|^{\frac{1}{k}}$.

Proof. Let λ be any eigenvalue of A, x is the corresponding eigenvector.

$$\begin{aligned}
\left\|A^k\right\| &= \max_{\|y\|=1} \left\|A^k y\right\| = \max_{\|y\|\neq 0} \frac{1}{\|y\|}\|Ay\| \\
\left\|A^k\right\| &\ge \frac{1}{\|y\|}\left\|A^k y\right\| \\
&= \frac{1}{\|y\|}\|A_1 \cdot A_2 \cdots A_y\| \\
&= \frac{1}{\|y\|}\left\|\lambda^k y\right\| \\
&= \left|\lambda^k\right| \\
\Rightarrow \left\|A^k\right\|^{\frac{1}{k}} &\ge P(A)
\end{aligned}$$

□

Proposition 2.13. $\lim_{k\to\infty} \left\|A^k\right\|^{\frac{1}{k}} = \rho(A)$

2.3 Symmetric Matrices

Proposition 2.14. Let $A \in \mathbb{R}^{n\times n}$ be symmetric, then

1. Its eigenvalues are all real.
2. For each eigenvalue, its algebraic multiplicity is equal to its geometric multiplicity.
3. Its eigenvectors are n mutually orthogonal nonzero real vectors.
4. A can be decomposed as $A = \sum_{i=1}^{n} \lambda_i x_i x_i^T$ where x_i is the corresponding eigenvector for eigenvalue λ_i and $\|x\|_2 = 1$.

Proposition 2.15. Let $A \in \mathbb{R}^{n\times n}$ be symmetric, then $\|A\|_2 = \rho(A)$.

Proof. We know that $\rho(A) \le \|A\|_2$ from Proposition (2.12), so it remains to show $\|A\|_2 \le \rho(A)$.

Let $\lambda_1, ..., \lambda_n$ be the n eigenvalues of A (counting multiplicity). Let $\{x_i : i = 1, ..., n\}$ be the corresponding n mutually orthogonal eigenvectors of A which are normalized, i.e., $||x_i||_2 = 1$ (why does A have such eigenvectors ?).

Let y be any vector in $\mathbb{R}^n$ such that $\|y\|_2 = 1$, then we can write y as $y = \sum_{i=1}^{n} a_i x_i$ for some $a_i \in \mathbb{R}$ ($\sum a_i^2 = 1$ why?). Then

$$\begin{aligned} \|Ay\|_2^2 &= \left\|A \sum a_i x_i\right\|_2^2 \\ &= \left\|\sum a_i A x_i\right\|_2^2 \\ &= \left\|\sum a_i \lambda_i x_i\right\|_2^2 \\ &= \sum a_i^2 \lambda_i^2 \|x_i\|_2^2 \quad \text{(by Pythagorean Theorem)} \\ &= \sum a_i^2 \lambda_i^2 \\ &\le \rho(A)^2 \quad \left(\text{since } \sum a_i^2 = 1\right) \end{aligned} \tag{2.1}$$

Therefore we have $\|Ay\|_2 \le \rho(A)$ for any $\|Ay\|_2 = 1$. Hence,

$$\|A\|_2 = \max_{\|y\|_2=1} \|Ay\|_2 \le \rho(A). \tag{2.2}$$

□

Proposition 2.16. Let $A \in \mathbb{R}^n$ be symmetric, then $\left\|A^k\right\|_2 = \|A\|_2^k, \forall k = 1, ..., n$.

Proposition 2.17. Let $A \in \mathbb{R}^n$, then $\left\|A^{-1}\right\|_2 = \dfrac{1}{\lambda_{\min}}$

2.4 Positive Semidefinite Matrix

DEFINITION 2.18. (Positive Definite and Positive Semidefinite Matrix)

A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semidefinite (PSD) if $x^T Ax \geq 0, \forall x \in \mathbb{R}^n$.

It is positive definite (PD) if $x^T Ax > 0, \forall x \in \mathbb{R}^n, x \neq 0$.

Proposition 2.19.

- A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semidefinite if and only if all its eigenvalues are non-negative.
- A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive definite if and only if all its eigenvalues are positive.

Proposition 2.20. A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semidefinite if and only if $A = BB^T$ for some matrix $B \in \mathbb{R}^{n\times r}$ of full column rank where $r = \text{rank}(A)$.

2.5 Singular Value Decomposition

Let $A \in \mathbb{R}^{m\times n}$ be a matrix, the *singular value* of A is defined as the nonnegative square root of eigenvalue of AA^T.

Proposition 2.21. (Singular Value Decomposition)

Let $A \in \mathbb{R}^{m\times}$ be a matrix of rank r, then

$$A = U\Sigma V^T$$

where

- $\Sigma_{m\times n}$ is the **unique** rectangular matrix with r diagonal entries consisting of singular values of A in a nonincreasing order.
- $U_{m\times m}$ and $V_{n\times n}$ are orthogonal matrices.
- The columns of V form an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvalues of $A^T A$. These columns are called the right singular vectors of A.
- The columns of U form an orthonormal basis of $\mathbb{R}^m$ consisting of eigenvectors of AA^T . These columns are called the left singular vectors of A.
- If A is positive semidefinite(symmetric), each singular value is equal to its eigenvalue. If in addition A is positive definite, then $U = V$ (eigenvalue decomposition and singular value decomposition of a positive definite matrix is the same.

3 Convexity

3.1 Introduction

The optimization problem (1.1) is called a linear problem if the objective and constraint functions $f_0, ..., f_m$ are linear, i.e., it satisfies

$$f_i(\alpha x + \beta y) = \alpha f_i(x) + \beta f_i(x) \tag{3.1}$$

for all $x, y \in \mathbb{R}^n$ and all $\alpha, \beta \in \mathbb{R}$. If the optimization is not linear, it is called a nonlinear program which will be the main subject of this course.

An important class of optimization problem is called **convex optimization problems**.

A convex optimization problem is one in which the objective and constraint functions are convex, which means they satisfy the inequality

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y) \tag{3.2}$$

for all $x, y \in \mathbb{R}^n$ and all $\alpha, \beta \in \mathbb{R}$ with $\alpha + \beta = 1$, $\alpha \geq 0, \beta \geq 0$.

Convexity is more general than linearity. Any linear problem is a convex problem, but a convex problem is not necessarily linear (nonlinear).

3.2 Basic Definitions

DEFINITION 3.1. (Affine Set)

A set A is said to be an affine set if for any two distinct points, the line passing through these points lie in the set A. i.e.,

$$x_1, x_2 \in A \Rightarrow \theta x_1 + (1-\theta)x_2 \in A, \ \forall \theta \in \mathbb{R}$$

DEFINITION 3.2. (Convex Set)

A set C is said to be a *convex set* if for any two distinct points, the line segment passing through these points lie in the set C. i.e.,

$$x_1, x_2 \in C \Rightarrow \theta x_1 + (1-\theta)x_2 \in C, \ \forall \theta \in [0, 1].$$

Figure 3: Example - Convex and non-convex sets

Example. $B_\delta(\bar{x}) = \{x - \|x - \bar{x}\| \leq \delta\}$. Let $x_1, x_2 \in B_\delta(\bar{x})$.

$$\begin{aligned} \|\theta x_1 + (1-\theta)x_2 - \bar{x}\| &\leq \delta \\ \|\theta(x_1 - \bar{x}) + (1-\theta)(x_2 - \bar{x})\| &\leq \|\theta(x_1 - \bar{x})\| + \|(1-\theta)(x_2 - \bar{x})\| \\ &\leq \theta\delta + (1-\theta)\delta = \delta \end{aligned}$$

Proposition 3.3. The closed and open balls are convex

DEFINITION 3.4. (Affine Combination)

Given a finite number of points $x_1, x_2, \ldots, x_n$ in a real vector space $\mathbb{R}^m$, a convex combination of these points is a point of the form

$$x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

where the real numbers θ_i satisfy $\theta_1 + \theta_2 + \cdots + \theta_n = 1$..

DEFINITION 3.5. (Convex Combination)

Given a finite number of points $x_1, x_2, \ldots, x_n$ in a real vector space $\mathbb{R}^m$, a convex combination of these points is a point of the form

$$x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

where the real numbers θ_i satisfy $\theta_i \geq 0$ and $\theta_1 + \theta_2 + \cdots + \theta_n = 1$..

DEFINITION 3.6. (Affine Hull)

Let C be a set in $\mathbb{R}^n$, then the *affine hull* of C, denoted as $\operatorname{aff} C$, is defined as

$$\{x \mid x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n\}$$

where θ_i satisfy $\theta_1 + \theta_2 + \cdots + \theta_n = 1$..

DEFINITION 3.7. (Convex Hull)

Let C be a set in $\mathbb{R}^n$, then the *convex hull* of C, denoted as $\operatorname{cov} C$, is defined as

$$\{x \mid x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n\}$$

where θ_i satisfy $\theta_1 + \theta_2 + \cdots + \theta_n = 1.$ and $\theta_i \geq 0$.

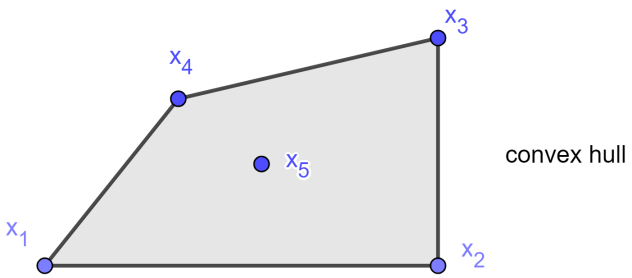


Figure 4: Example - Convex hull

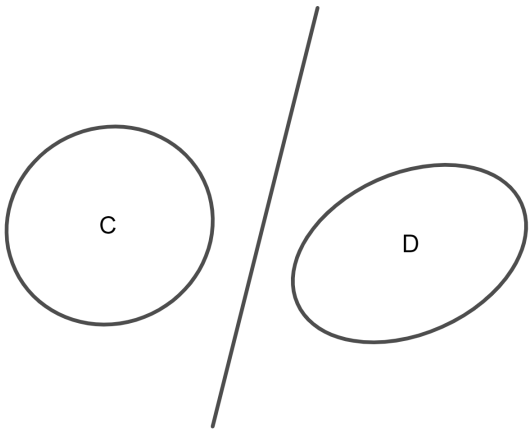
Proposition 3.8. The following holds:

1. For any collection of $\{Ci : i \in I\}$ of convex sets, their intersection $\cap_{i\in I} C_i$ is convex.
2. The vector (Minkowski) sum $\{x + y : x \in C_1, y \in C_2\}$ of two convex sets C_1, C_2 is convex.
3. The image of a convex set under an affine transformation is a convex set.

Proof. Exercise □

THEOREM 3.9. (Hyperplane Separation Theorem)

Suppose C and D are nonempty disjoint convex sets, i.e., $C \cap D \neq \emptyset$. Then there exist $a \neq 0$ such that $a^T x \leq b$ for all $x \in C$ and $a^T x \geq b$ for all $x \in D$.



$$C \cap D = \emptyset$$

Figure 5: Example - Hyperplane Separation Theorem

3.3 Convex Function

A Minkowski sum is defined as:

$$\{x + y : x \in C_1, y \in C_2\}$$

DEFINITION 3.10. (Convex Function)

A function $f : \mathbb{R}^n \to R$ is said to be convex if

1. its domain dom f is a convex set.
2. $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y), \forall\, x, y \in \text{dom}\, f, \forall\, 0 \leq \lambda \leq 1$

A function is said to be strictly convex if a strict inequality ($<$) holds as well.

We say f is **concave** if $-f$ is convex, and **strictly concave** if $-f$ is strictly convex.

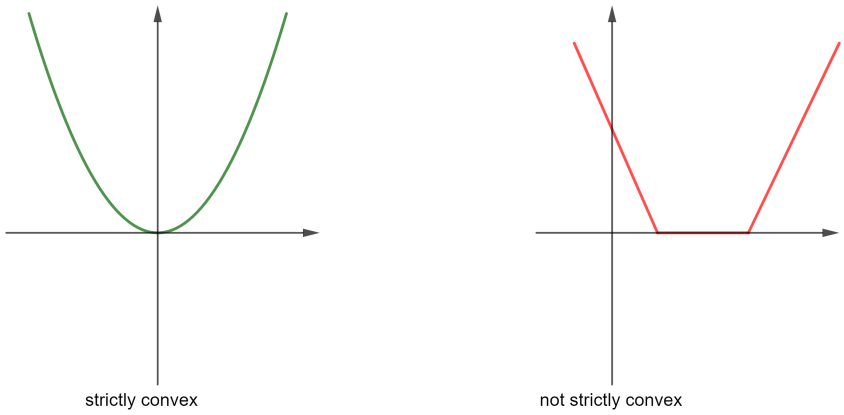


Figure 6: Example - Convexity

Example 1.
$y = x^2$ is convex,

$y = -x^2$ is not convex (concave)

$y = -x^3$ not convex, not a concave.

DEFINITION 3.11. (Level set)

The $\alpha - level\ set$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$C_\alpha = \{x \in \text{dom}\, f \mid f(x) = \alpha\}$$

DEFINITION 3.12. (Sublevel Set)

The $\alpha - sublevel\ set$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$C_\alpha = \{x \in \text{dom}\, f \mid f(x) \leq \alpha, \alpha \in \mathbb{R}\}$$

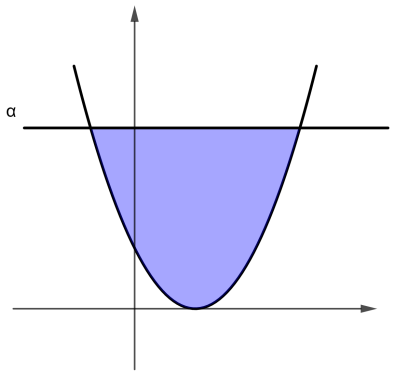


Figure 7: Example - Sublevel set

DEFINITION 3.13. (Epigraph)

The graph of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\{(x, f(x)) \mid x \in \operatorname{dom} f\}$$

which is a subset of $\mathbb{R}^n$. The *epigraph* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\operatorname{epi} f = \{(x, t) \mid x \in \operatorname{dom} f,\ f(x) \leq t\}$$

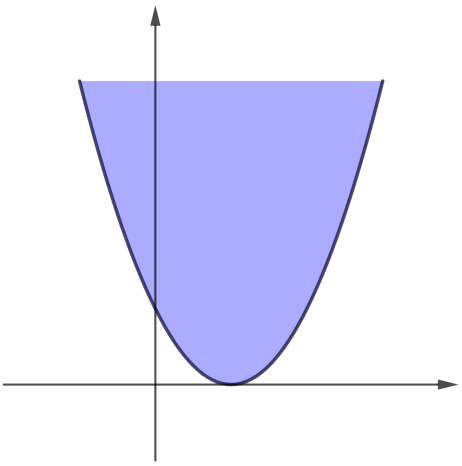


Figure 8: Example - Epigraph

Proposition 3.14. The following holds:

1. If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then its sublevel sets are convex as well. (the converse is not true)
2. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if its epigraph is a convex set

Proof. From the definition of convexity.

1. Let $x, y \in C_\alpha$, then
$$\begin{aligned} f(\lambda x + (1-\lambda)y) &\leq \lambda f(x) + (1-\lambda) f(y) \\ &= \lambda\alpha + (1-\lambda)\alpha = \alpha \end{aligned}$$
Hence $\lambda x + (1-\lambda)y \in C_\alpha$, its sublevel set is convex.
2. ($\Rightarrow$) Let (x_1, t_1) and $(x_2, t_2) \in \operatorname{epi} f$. Then
$$\lambda(x_1, t_1) + (1-\lambda)(x_2, t_2) = (\lambda x_1 + (1-\lambda)x_2,\ \lambda t_1 + (1-\lambda)t_2)$$
Since $f((\lambda x_1 + (1 - \lambda x_2)) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \leq \lambda t_1 + (1-\lambda)t_2$, we have
$$(\lambda x_1 + (1-\lambda)x_2,\ \lambda t_1 + (1-\lambda)t_2) \in \operatorname{epi} f.$$
($\Leftarrow$) Now assume $\operatorname{epi} f$ is convex, then consider
$$\lambda(x_1, f(x_1)) + (1-\lambda)(x_2, f(x_2)) = (\lambda x_1 + (1-\lambda)x_2, \lambda f(x_1) + (1-\lambda) f(x_2) \in \operatorname{epi} f.$$
Therefore, $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$

□

Proposition 3.15.

1. Any affine function is convex.
2. If f is a convex function, then λf is also a convex function.
3. The sum of two convex function is also convex.
4. The maximum of two convex function is also convex.

Lemma 3.16. Any vector norm is convex.

Proof.
$$\begin{aligned} \|\lambda x + (1-\lambda)y\| &\leq \|\lambda(x)\| + \|(1-\lambda)y\| \\ &\leq \lambda \|x\| + (1-\lambda)\|y\| \quad \lambda \geq 0, (1-\lambda) \geq 0 \end{aligned}$$

□

Proposition 3.17. (Pointwise Maximum/Supremum of Convex Functions) The pointwise maximum of m convex functions $f_1, ..., fm$ is convex
$$f_{\max}(x) := \max_{1 \leq i \leq m} f_i(x).$$
The pointwise supremum of a family of convex functions index by a set $\mathcal{I}$ is convex
$$f_{\sup}(x) := \sup_{i \in \mathcal{I}} f_i(x)$$

Proof. For any $0 \leq \theta \leq 1$ and any $x, y \in \mathbb{R}$

$$\begin{aligned} f_{\sup}(\theta x + (1-\theta)y) &= \sup_{i \in \mathcal{I}} f_i(\theta x + (1-\theta)y) \\ &\leq \sup_{i \in \mathcal{I}} \theta f_i(x) + (1-\theta) y f_i(y) \\ &\leq \theta \sup_{i \in \mathcal{I}} f_i(x) + (1-\theta) \sup_{j \in \mathcal{I}} f_j(x) \\ &= \theta f_{\sup}(x) + (1-\theta) f_{\sup}(y). \end{aligned}$$

□

Proposition 3.18. Any induced norm is convex.

Proof.

$$\|A\| = \max_{i \in I,\ I = \{\|i\| = 1\}} \|Ai\|$$

Fix i,

$$\|(\lambda(A + (1-\lambda)B)i\| \leq \lambda \|A_i\| + (1-\lambda) \|B_i\|.$$

□

THEOREM 3.19. Consider an optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \Omega \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and Ω is a convex set. Then any local optimum is also a global optimum.

Proof. Let $\bar{x} \in \Omega$ be a local minimizer, then there exists $\epsilon > 0$ such that

$$f(\bar{x}) \leq f(x),\ \forall x \in B_\epsilon(\bar{x}) \cap \Omega$$

Now suppose $\bar{x}$ is not global minimum, then $\exists z \in \omega$ with $f(z) < f(\bar{x})$.

But because of convexity of Ω, we have

$$\lambda \bar{x} + (1-\lambda) z \in \Omega,\ \forall \lambda \in [0, 1]$$

By convexity of f we have

$$\begin{aligned} f(\lambda \bar{x} + (1-\lambda) z) &\leq \lambda f(\bar{x}) + (1-\lambda) f(z) \\ &< \lambda f(\bar{x}) + (1-\lambda) f(\bar{x}) \\ &= f(\bar{x}) \end{aligned}$$

But as $\lambda \to 1$, $\lambda \bar{x} + (1-\lambda) z \to \bar{x}$, so $\lambda \bar{x} + (1-\lambda) z \in B_\epsilon(\bar{x}) \cap \Omega$ for some λ close to 1, which is a contradiction of the local optimality of $\bar{x}$. □

4 Multivariate Calculus

DEFINITION 4.1. (C^k Class)

A function $f : D \to \mathbb{R}$ is called a C^k class or C^k-smooth function ($f \in C^k$), over D, if all its kth derivatives are continuous over D (Usually assume D is open).

Note a function is differentiable at a point implies it must be continuous at the point.

Example. $f := \begin{cases} 0 & x \leq 0 \\ x^2 & x \geq 0 \end{cases}$. Here f'' is not defined at $x = 0$ (f' is not differentiable on 0), so it belongs to C^1 but not in C^2.

$y = |x|^3$ is in C^2 but not in C^3.

DEFINITION 4.2. (Gradient)

Let $f \in C^1 : \mathbb{R}^n \to \mathbb{R}$. Its gradient $\nabla f \in C^0 : \mathbb{R}^n \to \mathbb{R}^\ltimes$ is given by

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$

DEFINITION 4.3. (Hessian)

Let $f \in C^2 : \mathbb{R}^n \to \mathbb{R}$. Its Hessian $\nabla^2 f \in C^1 : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ is given by

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Note the partial derivatives commute if $f \in C^2$ (Schwarz's theorem or Clairaut's theorem).

THEOREM 4.4. (Taylor Theorem for Univariable Functions)

Let $f : \mathbb{R} \to \mathbb{R}$ be a function that has $n+1$ continuous derivatives in some neighbourhood U of $x = a$. Then for any $x \in U$,

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots \frac{f^{(n)}(a)}{n!}(x-a)^n + R_{n,a}(x),$$

where $R_{n,a}(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{(n+1)}$, c is between x and a.

THEOREM 4.5. (Taylor Theorem for Multivariate Functions)

- (First order) Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then

$$f(x+h) = f(x) + h^T \nabla f(x) + \phi(h).$$

  where $\lim_{h\to 0} \frac{\phi(h)}{\|h\|} = 0$.

- (First order explicit form) Let $f \in C^1 : \mathbb{R}^n \to \mathbb{R}$. Then

$$f(x+h) = f(x) + h^T \nabla f(x+th), 0 < t < 1.$$

- (Second order) Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable. Then

$$f(x+h) = f(x) + h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x) h + \phi(h),$$

  where $\phi(h) = \frac{1}{2} h^T \nabla^2 f(x+\lambda h) h$, $0 \le \lambda \le 1$ with $\lim_{h\to 0} \frac{\phi(h)}{\|h\|^2} = 0$.

- (Second order explicit form) Let $f \in C^2 : \mathbb{R}^n \to \mathbb{R}$. Then

$$f(x+h) = f(x) + h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x+th) h, \ 0 < t < 1$$

DEFINITION 4.6. (Directional Derivative)

The directional derivative of f in the direction of h is

$$\nabla_h f(x) = \lim_{\alpha \to 0} \frac{f(x+\alpha h) - f(x)}{\alpha}.$$

THEOREM 4.7. Let $f \in C^1$, then $\nabla_h f = h^T \nabla f$.

Proof.

$$\begin{aligned}
\nabla_h f &= \lim_{\alpha \to 0} \frac{f(x+\alpha h) - f(x)}{\alpha} \\
&= \lim_{\alpha \to 0} \frac{f(x) + \alpha h^T \nabla f(x) + \phi(\alpha h) - f(x)}{\alpha} \\
&= h^T \nabla f(x) + \lim_{\alpha \to 0} \frac{\phi(\alpha h)}{\alpha} \\
&= h^T \nabla f(x) + \lim_{\|\alpha h\| \to 0} \frac{\phi(\alpha h)}{\|\alpha h\|} \|h\| \\
&= h^T \nabla f(x)
\end{aligned}$$

□

Proposition 4.8. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $\operatorname{dom} f$ is convex. Then f is convex if and only if

$$\forall x, y \in C,\ f(y) \geq f(x) + (y - x)^T \nabla f(x).$$

Proof. Assume f is a convex function, then for any $x, y \in \operatorname{dom} f$ and $t \in [0, 1]$, we have

$$tf(y) + (1 - t)f(x) \geq f(x + t(y - x)).$$

Divided by t, we have

$$f(y) - f(x) \geq \frac{f(x + t(y - x)) - f(x)}{t}$$

Let $t \to 0$, then

$$f(y) - f(x) \geq \lim_{t \to 0} \frac{f(x + t(y - x)) - f(x)}{t} = \nabla f(x)^T (y - x)$$

For the other direction, for any $x, y \in \operatorname{dom} f$ and any $t \in [0, 1]$, define $z = tx + (1 - t)y$ (because $\operatorname{dom} f$ is convex). Then

$$f(x) \geq f(z) + \nabla f(z)^T (x - z) \qquad (1)$$
$$f(y) \geq f(z) + \nabla f(z)^T (y - z) \qquad (2)$$

t times (1) + $(1 - t)$ times (2):

$$tf(x) + (1 - t)f(y) \geq f(z) + 0 = f(tx + (1 - t)y).$$

which is exactly the definition of a convex function. □

Proposition 4.9. Let $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2$, and $\operatorname{dom} f$ is a convex set. Then

- $\nabla^2 f(x) \succeq 0, \forall x \in \operatorname{dom} f$ if and only if f is convex function.
- if $\nabla^2 f(x) \succ 0, \forall x \in \operatorname{dom} f$, then f is strictly convex.

Proof. First assume that $\nabla^2 f(x) \not\succeq 0$, i.e., $\exists v \neq 0 \in \mathbb{R}^n$ such that $v^T \nabla^2 f(x) v < 0$. By Taylor expansion, we have

$$f(x + tv) = f(x) + t\nabla^2 f(x)v + \frac{t^2}{2} v^T \nabla^2 f(x) v + O(t^2)$$

where $O(t^2)$ is a higher order term than t^2. Then divided by t^2, we have

$$\frac{f(x + tv) - f(x) - t\nabla f(x)^T v}{t^2} = \frac{1}{2} v^T \nabla^2 f(x) v + O(1)$$

Hence as $t \to 0$, $\dfrac{f(x + tv) - f(x) - t\nabla f(x)^T v}{t^2} < 0$, which is a contradiction to Proposition (4.8).

Now assume $\forall x \in \operatorname{dom} f, \nabla^2 f(x) \succeq 0$. Then for any $x, y \in \operatorname{dom} f, x \neq y$, by Taylor expansion (Second order explicit form), we have

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x + t(y - x))(y - x), \text{ for some } t \in (0, 1)$$

Then since $\operatorname{dom} f$ is convex, $x + t(y - x) \in \operatorname{dom} f$, so $\nabla^2 f(x + t(y - x)) \succeq 0$. Therefore,

$$\frac{1}{2}(y - x)^T \nabla^2 f(x + t(y - x))(y - x) \geq 0.$$

Hence,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x),$$

which implies $f(x)$ is a convex function. In addition, if $\nabla^2 f(x) \succ 0, \forall x \in \operatorname{dom} f$, then

$$f(y) > f(x) + \nabla f(x)^T (y - x).$$

Hence f is strictly convex. □

5 Optimality Conditions for Unconstrained Optimization Problem

DEFINITION 5.1. (Critical/Stationary Points)

All points x such that $\nabla f(x) = 0$ are called critical or stationary points

THEOREM 5.2. (First Order Necessary Conditions for Local Optimality)

Let $f \in C^1 : \mathbb{R}^n \to \mathbb{R}$. If x^* is a local minimizer, then $\nabla f(x^*) = 0$.

Proof. For any $h \in \mathbb{R}^n$, consder the taylor expansion at $x = x^*$

$$f(x^* + \alpha h) = f(x^*) + \alpha h^T \nabla f(x^*) + O(\alpha)$$

and divide by α,

$$\frac{f(x^* + \alpha h) - f(x^*)}{\alpha} = h^T \nabla f(x^*) + O(1)$$

Because x^* is a local minimizer, we have

$$\lim_{\alpha \to 0^+} \frac{f(x^* + \alpha h) - f(x^*)}{\alpha} = h^T \nabla f(x^*) \geq 0$$

$$\lim_{\alpha \to 0^-} \frac{f(x^* + \alpha h) - f(x^*)}{\alpha} = h^T \nabla f(x^*) \leq 0$$

Therefore, $h^T \nabla f(x^*) = 0$ for any h, which implies $\nabla f(x^*) = 0$. □

THEOREM 5.3. (Second Order Necessary Conditions for Local Optimality)

Let $f \in C^2 : \mathbb{R}^n \to \mathbb{R}$. If x^* is a local minimizer, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is a PSD matrix.

Proof. Suppose on the contrary that $\nabla^2 f(x^*) \not\succeq 0$, i.e., $\nabla^2 f(x^*)$ has negative eigenvalues. Let d be the eigenvector corresponding to a negative eigenvalue λ.

Consider the second order Taylor expansion of $f(x)$ at x^*, since x^* is a local minimizer, we have $\nabla f(x^*) = 0$. Hence

$$\frac{f(x^* + d) - f(x^*)}{\|d\|^2} = \frac{1}{2} \frac{d^T}{\|d\|} \nabla^2 f(x^*) \frac{d}{\|d\|} + O(1)$$

Since $\dfrac{d}{\|d\|}$ is a unit vector, we have

$$\frac{f(x^* + d) - f(x^*)}{\|d\|^2} = \frac{1}{2}\lambda + O(1)$$

When $\|d\|$ is sufficiently small, we have $f(x^* + d) < f(x^*)$, which is a contradiction of x^* being a local minimizer. □

THEOREM 5.4. (Second Order Sufficient Conditions for Local Optimality)
Let $f \in C^2 : \mathbb{R}^n \to \mathbb{R}$. If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is a PD matrix, then x^* is a strict local minimizer.

Proof. Follow the proof the previous theorem, assume $\nabla^2 f(x^*) \succ 0$, then for any $\|d\|_2 \neq 0$, we have

$$d^T \nabla^2 f(x^*) d \geq \lambda_{\min} \|d\|_2^2$$
$$\frac{d^T}{\|d\|_2} \nabla^2 f(x^*) \frac{d}{\|d\|_2} \geq \lambda_{\min} \|d\|_2^2$$

(for any $0 \neq d \in \mathbb{R}^n$, d can be expressed as $d = \sum_{i=1}^{n} v_i$ where v_i is the i-th eigenvector of $\nabla^2 f(x^*)$, and v_i are orthogonal to each other). By Taylor's expansion,

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

Hence,

$$\frac{f(x^* + d) - f(x^*)}{\|d\|_2^2} \geq \frac{1}{2} \lambda_{\min} + O(1).$$

Therefore $f(x^* + d) > f(x^*)$ when $\|d\|_2$ is sufficiently small. □

THEOREM 5.5. [Stationary point of convex function] Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and convex. If x^* is a stationary point, then x^* is also a global minimizer.

Proof. Since f is convex, we have $f(y) \geq f(x^*) + \nabla f(x^*)^T (y - x^*)$ for any y. Since x^* is a stationary point, $\nabla f(x^*) = 0$. Therefore $f(y) \geq f(x)$ for any y. □

Remark 5.6. In Theorem 5.3 and Theorem 5.4, we require f to be C^2 instead of twice differentiable (weaker), that is because we need the partial derivatives to commute, so that the Hessian is a symmetric matrix (Recall Schwarz's theorem or Clairaut's theorem).

$\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is PD $\Rightarrow$

(1) x^* is a strict local minimizer.

(2) x^* is a local minimizer

(3) $\begin{cases} \nabla f(x^*) = 0 \\ \nabla^2 f(x^*) \succeq 0 \end{cases}$

Converse of (1), (2), (3) are all false. Counterexamples:

(1) $f(x) = x^4$ at $x^* = 0$, $\nabla^2 f(x) = 12x^2 = 0$

(2) $f(x) = 1$ at $x^* = 0$. minimizer but not strict.

(3) $f(x) = x^3$ at $x^* = 0$. $\nabla f(x^*) = 0$

5.1 Existence of optimal solution

DEFINITION 5.7. (Bounded Set, Closed Set, Compact Set)

A set S is *bounded* if $S \subseteq B_\delta(0)$ for some δ.

A set S is *closed* if for any sequence $x_1, x_2, ... \in S$ such that $\lim_{i\to\infty} x_i$ exists, then $\lim_{i\to\infty} x_i \in S$.

A set S is *compact* if it is bounded and closed.

THEOREM 5.8. (Weierstrass Extreme Value Theorem)

Every continuous function on a *compact set* S attains its extreme value at some $x \in S$. That is, there exist $x^* \in S$ such that $f(x^*) = \sup_S(f)$ or $\inf_S(f)$

THEOREM 5.9. If f is continuous, then its sublevel sets are closed.

Proof. Let $C_\alpha = \{x \,|f(x) \leq \alpha\}$ be a sublevel set. Let $\{x_k\} \subset C_\alpha$ be a sequence that $\lim_{k\to\infty}(x_k) = x^*$. Then

$$f(x^*) = f(\lim_{k\to\infty} x_k) = \lim_{k\to\infty} f(x_k) \leq \alpha.$$

Hence $x^* \in C_\alpha$ and C_α is closed. □

THEOREM 5.10. If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and has at least one bounded nonempty sublevel set, then f has a global optimizer.

Proof. Let $C_\alpha = \{x|f(x) \leq \alpha\}$ be a bounded sublevel set. Since f is continuous, it must be closed, therefore it must be compact. By Weierstrass Extreme Value Theorem, f must have a global optimizer x^* over C_α. Now for any $x \notin C_\alpha$, we must have $f(x) > \alpha \geq f(x^*)$. Hence x^* is a global optimizer over $\mathbb{R}^n$. □

Example. Functions without global minimizer:

1. Affine function $f(x) = Ax + b$
2. $f(x) = e^x$.

DEFINITION 5.11. (Coercive Function)

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be coercive if for every sequence $\{x_k\} \subset \mathbb{R}^n$ with $\|x_k\| \to \infty$, it must be the case that $f(x^k) \to \infty$.

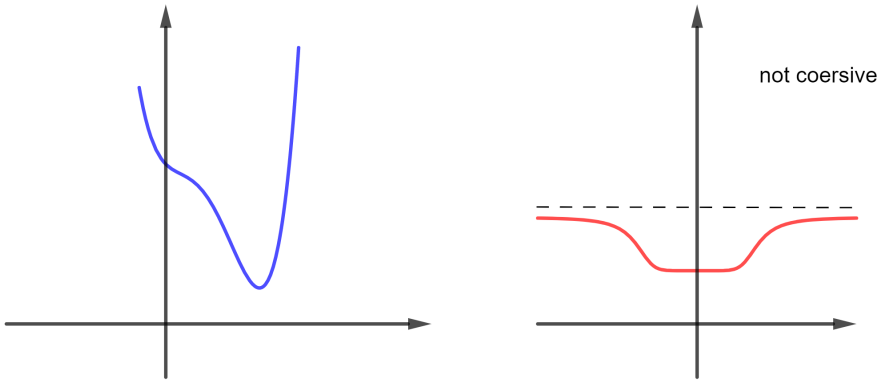


Figure 9: Coercive and not coercive

THEOREM 5.12. (Coercivity and Compactness)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous on $\mathbb{R}^n$. Then f is coercive if and only if all its sublevel sets are compact (or bounded).

Proof. We first show that the coercivity of f implies compactness of the sets $\{x \in \mathbb{R}^n | f(x) \leq \alpha\}$. We begin by noting that the continuity of f implies the closeness of the sets $\{x \in \mathbb{R}^n | f(x) \leq \alpha\}$. Thus, it remains only to show that any set of the form $\{x \in \mathbb{R}^n | f(x) \leq \alpha\}$ is bounded. We show this by contradiction. Suppose to the contrary that there is an α such that the set $S = \{x \in \mathbb{R}^n | f(x) \leq \alpha\}$ is unbounded. Then there must exist a sequence $\{x_k\} \subset S$ with $||x_k|| \to \infty$. But then, by the coercivity of f, we must also have $f(x_k) \to \infty$. This contradicts the fact that $f(x_k) \leq \alpha$ for all $k = 1, 2, ...$ Therefore the set S must be bounded.

Let us now assume that each of the sets $\{x | f(x) \leq \alpha\}$ is bounded and let $\{x_k\} \subset \mathbb{R}^n$ be such that $||x_k|| \to \infty$. Let us suppose that there exists a subsequence of the integers $J \subset N$ such that the set $\{f(x_k)\}_J$ is bounded above. Then there exists $\alpha \in \mathbb{R}^n$ such that $\{x_k\}_J \subset \{x | f(x) \leq \alpha\}$. But this cannot be the case since each of the sets $\{x | f(x) \leq \alpha\}$ is bounded while every subsequence of the sequence $\{x_k\}$ is unbounded by definition. Therefore, the set $\{f(x_k)\}_J$ cannot be bounded, and so the sequence contains no bounded subsequence, i.e. $f(x_k) \to \infty$. □

Proposition 5.13. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous on $\mathbb{R}^n$, If f is coercive, then f has at least one global minimizer.

Example. Let $A \in \mathbb{R}^{m \times n}$ with $\text{rank}(A) = n$, then $f(x) = \|Ax - b\|_2$ is coercive.

Solution: By triangle inequality, $\|Ax - b\|_2 \geq \|Ax\| - \|b\| = \sqrt{x^T A^T A x} - \|b\|_2$.

Since $\text{rank}(A^T A) = \text{rank}(A) = n$, we know $A^T A$ is a PD matrix. Similarly as the proof of Theorem 5.4, we know

$$\sqrt{x^T A^T A x} \geq \sqrt{\lambda_{\min} \|x\|_2^2} = \sqrt{\lambda_{\min}} \|x\|_2 .$$

Therefore $f(x) \to \infty$ as $\|x\|_2 \to \infty$ □

THEOREM 5.14. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, a local minimizer of f is also a global minimizer. If f is strictly convex, then there is at most one global minimizer.

5.2 Application to Least Squares problem

Given m points $x^{(1)}, ..., x^{(m)} \in \mathbb{R}^n$, and m points $y^{(1)}, ..., y^{(m)} \in \mathbb{R}$. The goal is to find a line $y = a^T x$ (more specifically, find the coefficients $a \in \mathbb{R}^n$) that best approximate the given data (See figure 10).

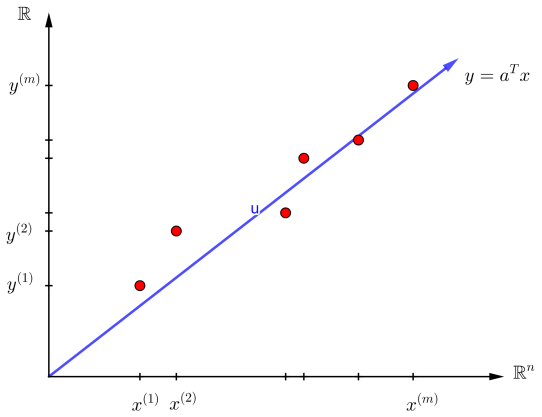


Figure 10: Least Squares

We minimize the following function

$$f(a) = \sum_{i=1}^{m}((x^{(i)})^T a - y^{(i)})^2 = (Xa - y)^T(Xa - y) = \|Xa - y\|_2^2$$

The linear least squares problem is the following unconstrained minimization problem

$$\min_{a \in \mathbb{R}^n} \|Xa - y\|_2^2 \tag{5.1}$$

where $X \in \mathbb{R}^{m \times n}$ is a matrix of the form:

$$X = \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix}$$

and $y \in \mathbb{R}^m$ is a vector of the form:

$$y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

THEOREM 5.15.

If $\text{rank}(X) = n$, then it has one unique global minimizer.

Proof. Recall

$$\begin{aligned} f(a) &= \|Xa - y\|_2^2 \\ &= a^T X^T X a - y^T X a - a^T X^T y + y^T y \\ &= a^T X^T X a - 2y^T X a + y^T y \end{aligned}$$

Since $\operatorname{rank}(X) = n$, by Proposition 5.13, it is coercive and coercivity implies it has at least one global minimizer. Since $X^T X$ is a $n \times n$ matrix, and $\operatorname{rank}(X^T X) = \operatorname{rank}(X) = n$, it must be a positive definite matrix. Therefore it is strictly convex, which implies the global minimizer is unique. □

Remark 5.16. Note if we want to find a line that does not go through the origin, i.e., a line of the form $y = a^T x + c$. We can rewrite the line equation as $y = [a^T, c][x^T, 1]^T$. Therefore we can simply add 1's to the matrix X such that the least squares problem has the same formulation:

$$\min_{a \in \mathbb{R}^n,\, c \in \mathbb{R}} \left\| \bar{X}[a^T, c]^T - y \right\|_2^2 \tag{5.2}$$

where $\bar{X} \in \mathbb{R}^{m \times n}$ is a matrix of the form:

$$\bar{X} = \begin{bmatrix} (x^{(1)})^T & 1 \\ \vdots & \vdots \\ (x^{(m)})^T & 1 \end{bmatrix}$$

and $y \in \mathbb{R}^m$ is a vector of the form:

$$y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

6 Unconstrained Quadratic Optimization

6.1 Quadratic Function

DEFINITION 6.1. (Quadratic Function)

A quadratic function is a function of the form $q(x) = x^T Ax + b^T x + c$ for any $A \in \mathbb{R}^n, b \in \mathbb{R}, c \in \mathbb{R}$.

WLOG we can assume that A is symmetric.

Lemma 6.2. Let $A \in \mathbb{R}^n$ and let $G = \frac{1}{2}(A + A^T)$. Then

$$q(x) = x^T Gx + bx + c, \quad \forall x \in \mathbb{R}^n.$$

Proof. $x^T Ax = \frac{x^T Ax}{2} + \frac{x^T Ax}{2} = \frac{x^T Ax}{2} + \frac{x^T A^T x}{2} = \frac{1}{2} x^T \left(A + A^T\right) x = x^T Gx$ □

6.2 Matrix calculus

Let $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^n, x \in \mathbb{R}^n$, then

1. $\frac{\partial(x^T)}{\partial x} = I$,
2. $\frac{\partial(x^T b)}{\partial x} = b$,
3. $\frac{\partial(x^T A^T)}{\partial x} = A^T$.

Similarly,

1. $\frac{\partial(x)}{\partial x^T} = I$,
2. $\frac{\partial(b^T x)}{\partial x^T} = b^T$,
3. $\frac{\partial(Ax)}{\partial x^T} = A$.

Lemma 6.3. (Chain Rule) Let $g \in \mathbb{R}^m, h \in \mathbb{R}^n$. Then

$$\frac{\partial f(g,h)}{\partial x} = \frac{\partial(g(x)^T)}{\partial x} \frac{\partial f(g,h)}{\partial g} + \frac{\partial(h(x)^T)}{\partial x} \frac{\partial f(g,h)}{\partial h}.$$

Lemma 6.4. Let $A \in \mathbb{R}^n$. Then

$$\frac{\partial(x^T Ax)}{\partial x} = (A + A^T)x.$$

Proof. Let $g(x) = x$ and $h(x) = Ax$. Then

$$\begin{aligned}\frac{\partial(x^T Ax)}{\partial x} &= \frac{\partial x^T}{\partial x}\frac{\partial x^T Ax}{\partial x} + \frac{\partial((Ax)^T)}{dx}\frac{\partial x^T Ax}{\partial Ax} \\ &= Ax + \frac{\partial(x^T A^T)}{dx}\frac{\partial((Ax)^T x)}{\partial Ax} \\ &= Ax + A^T x\end{aligned}$$

□

The gradient: $\nabla x^T Ax = \dfrac{\partial x^T Ax}{\partial x}$.

The Hessian: $\nabla^2 x^T Ax = \dfrac{\partial^2 x^T Ax}{\partial x \partial x^T} = \dfrac{\partial(Ax + A^T x)}{\partial x^T} = A + A^T$.

DEFINITION 6.5. (Generalized Inverse)

A matrix $A^- \in \mathbb{R}^{n\times m}$ is a *generalized inverse* of matrix $A \in \mathbb{R}^{m\times n}$ if $AA^-A = A$.

Lemma 6.6. Let $b \in \text{Range}(A)$, then $\bar{x} = A^- b$ is a solution of the linear system $Ax = b$.

Lemma 6.7. Let $A \in \mathbb{R}^{m\times n}$, then for any $b \in \mathbb{R}^n$, b can be uniquely written as $b = y + z$ where $y \in \text{Range}(A^T)$ and $z = \text{Null}(A)$.

Proof. Assume $\dim(\text{Range}(A^T)) = n_1$ and $\dim(\text{Null}(A)) = n_2$. Then by the Rank-Nullity Theorem, $n_1 + n_2 = n$. Let $a_1, ..., a_{n_1}$ be an orthogonal basis of $\text{Range}(A^T)$ and $e_1, ..., e_{n_2}$ be an orthogonal basis of $\text{Null}(A)$. From the definition we know that $\text{Range}(A^T) \perp \text{Null}(A)$, then it is clear that $\langle a_i, e_j\rangle = 0$ for any $1 \le i \le n_1, 1 \le j \le n_2$. It then implies $a_1, ..., a_{n_1}, e_1, ..., e_{n_2}$ must be linearly independent. (Why? prove it using the orthogonality of the basis)

Hence $\{a_1, ..., a_{n_1}, e_1, ..., e_{n_2}\}$ spans the whole $\mathbb{R}^n$ and we can write b as $b = \sum c_i a_i + \sum d_j e_j$, then simply let $y = \sum c_i a_i$ and $z = \sum d_j e_j$. The uniqueness comes from the linear independence of the basis. □

THEOREM 6.8. Let $q(x) = x^T Ax + b^T x + c$ and $A \in \mathbb{R}^n$ be symmetric and $b \in \mathbb{R}, c \in \mathbb{R}$.

- If A is positive definite, then $q(x)$ has a unique global minimizer $x^* = -\dfrac{1}{2}A^{-1}b$.

- If A is positive semidefinite and $b \in \text{Range}(A)$, then $q(x)$ has global minimizer x^* such that $Ax^* = -\frac{1}{2}b$. A particular solution is $x^* = A^- b$.
- For all the other cases, $q(x)$ is unbounded, i.e., $q(x) = -\infty$ for some $\|x\| \to \infty$.

Proof.

1. Since $\nabla f(x^*) = 0$ and $\nabla^2 f(x) = 2A \succ 0$, by proposition 4.9, f must be strictly convex. Then by the second order sufficient condition, x^* is a local minimizer. Since f is convex, any local minimzer is a global minimizer. Since f is strictly convex, x^* must be a unique global minimizer.
2. Now assume A is positive semidefinite but not positive definite. Then let x^* be a stationary point such that $2Ax^* + b = 0$, we will show that for any $y \in \mathbb{R}^n$, $q(x^*) \leq q(y)$.

$$\begin{aligned} q(y) &= q(y - x^* + x^*) \\ &= (y - x^* + x^*)^T A(y - x^* + x^*) + b^T(y - x^* + x^*) + c \\ &= (y - x^*)^T A(y - x^*) + 2(y - x^*)^T A(x^*) + (x^*)^T A(x^*) + b^T(y - x^*) + b^T x^* + c \\ &= (y - x^*)^T A(y - x^*) - (y - x^*)^T b + (x^*)^T A(x^*) + b^T(y - x^*) + b^T x^* + c \\ &= (y - x^*)^T A(y - x^*) + (x^*)^T A(x^*) + b^T x^* + c \\ &= (y - x^*)^T A(y - x^*) + q(x^*) \end{aligned}$$

   Since A is PSD, $(y - x^*)^T A(y - x^*) \geq 0$. Hence $q(y) \geq q(x^*)$.
3. For the other cases, note $b \in \text{Range}(A)$ in both case 1 and case 2. So there are two cases left: (a) $b \notin \text{Range}(A)$. (b) $b \in \text{Range}(A)$ but A is not PSD.

   For part (a), by lemma 6.7, b can be uniquely written as

$$b = y + z,$$

   where $y \in \text{Range}(A^T)$ and $z \in \text{Null}(A)$. Since $b \notin \text{Range}(A^T) = \text{Range}(A)$, it implies $z \neq 0$. Now let $x = \lambda z$, then

$$q(\lambda z) = \lambda^2 z^T A z + \lambda(y + z)^T z + c = 0 + 0 + \lambda z^T z + c.$$

   Therefore, $q(\lambda x) \to -\infty$ as $\lambda \to -\infty$.

   For part (b), since A is not PSD, there exists some $v \neq 0$ such that $v^T A v < 0$. Let $x = \lambda v$, then

$$q(\lambda v) = \lambda^2 v^T A v + \lambda b^T v + c.$$

   Since $v^T A v < 0$, it implies $q(\lambda v) \to -\infty$ as $\lambda \to \infty$ (The quadratic term λ^2 grows faster than the linear term λ).

□

7 Equivalent Norms

DEFINITION 7.1. (Equivalent Norms)

We say that two norms $\|\cdot\|_F$ and $\|\cdot\|_G$ on $\mathbb{R}^n$ are equivalent if there exists some constant $C_1 > 0, C_2 > 0$ such that

$$C_1 \|x\|_G \leq \|x\|_F \leq C_2 \|x\|_G .$$

Lemma 7.2. Given two equivalent norms $\|\cdot\|_F$ and $\|\cdot\|_G$, we have the following:

1. $\lim_{k\to\infty} \|x_k\|_F \to \infty$ if and only if $\lim_{k\to\infty} \|x_k\|_G \to \infty$.
2. $\lim_{k\to\infty} \|x_l\|_F \to 0$ if and only if $\lim_{k\to\infty} \|x_k\|_G \to 0$.

THEOREM 7.3. Any two norms on a finite dimensional vector space are equivalent.

Proof. It suffices to prove it in the case the finite-dimensional space is $\mathbb{R}^n$ and $\|\cdot\|_G$ is the 2-norm. Let e_i be the unit vector such that the ith element is 1, and all the other elements are zero. Then $\{e_1, ..., e_n\}$ is a basis of $\mathbb{R}^n$ and we can write $x \in \mathbb{R}^n$ as $x = \sum_i x_i e_i$ where $x_i \in \mathbb{R}$.

Let $\mu = \max_{1\leq i\leq n} \|e_i\|_F$, then

$$\begin{aligned}
\|x\|_F &= \left\| \sum_i x_i e_i \right\|_F \\
&\leq \sum_i |x_i| \|e_i\|_F \quad \text{(Triangle Inequality of norms)} \\
&\leq \mu \sum_i |x_i| \\
&\leq \mu\sqrt{n} \|x\|_2 \quad \text{(Cauchy-Schwartz Inequality to the vectors } |x| \text{ and } \textstyle\sum_i e_i\text{)}.
\end{aligned}$$

Thus

$$\|x\|_F - \|y\|_F \leq \|x - y\|_F \leq \mu\sqrt{n} \|x - y\|_2 ,$$

so $\|\cdot\|_F$ is continuous. Hence $\|\cdot\|_F$ restricted to the unit sphere $\{x : \|x\|_2 = 1\}$ attains its minimum at some point p (Weierstrass Extreme Value Theorem) and $\|p\|_F \leq \infty$. We then claim that

$$\|x\|_F \geq \|p\|_F \|x\|_2 .$$

When $x = 0$, this is true.

When $x \neq 0$, then

$$\|x\|_F = \|x\|_2 \left\| \frac{x}{\|x\|_2} \right\|_F \geq \|x\|_2 \|p\|_F ,$$

since p is the minimizer in the unit sphere, and $\frac{x}{\|x\|_2}$ is also in the unit sphere. Hence,

$$C_1 \|x\|_2 \leq \|x\|_F \leq C_2 \|x\|_2 ,$$

where $C_1 = \|p\|_F$ and $C_2 = \mu\sqrt{n}$. □

8 Algorithms for Unconstrained Optimization

There are mainly two classes of algorithms for solving unconstrained optimization:

- Line search type algorithms
- Trust region type algorithms

8.1 Line Search Algorithm

General framework for Line Search Algorithm:

DEFINITION 8.1. (Line Search Step)

(1) Choose a search direction d^k.

(2) Choose a step size $\alpha_k \geq 0$.

(3) Update $x^{k+1} = x^k + \alpha_k d^k$.

8.1.1 Descent Direction

How to choose a search direction? Usually we take the descent direction.

DEFINITION 8.2. (Descent Direction)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $x^k \in \mathbb{R}^n$ be such that $\nabla f(x^k) \neq 0$. Then d^k is called a *descent direction* if $\nabla f(x^k)^T d^k < 0$.

Lemma 8.3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and d^k be a direction such that $\nabla f(x^k)^T d^k < 0$, then $f(x^k + \epsilon d^k) \leq f(x^k)$ for sufficiently small ϵ.

Proof. Using Taylor Theorem. □

8.1.2 Steepest Descent

If we choose the direction d^k to be the negative of the gradient. Then locally it is the "steepest" direction in the following sense.

THEOREM 8.4. (Steepest Descent)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $\nabla f(x^k) \neq 0$. Let $d_\alpha = \arg\min_d \{f(x^k + \alpha d) : \|d\|_2 = 1\}$. Then

$$\lim_{\alpha \to 0^+} d_\alpha = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|_2}.$$

Proof. Denote $g^k = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|_2}$. Let $p \neq g^k$ and $\|p\| = 1$.

By Cauchy-Schwartz Inequality, we have

$$|\nabla f(x^k)^T p| = c|\nabla f(x^k)^T g^k|,$$

where $c = |\cos\theta|$ and θ is the angle between p and g^k. Since $p \neq g^k$ we must have $c < 1$.

Now consider the Taylor expansion of $f(x^k + \alpha p)$ and $f(x^k + \alpha g^k)$

$$f(x^k + \alpha p) = f(x^k) + \alpha \nabla f(x^k)^T p + O(\alpha p),$$

$$f(x^k + \alpha g^k) = f(x^k) + \alpha \nabla f(x^k)^T g^k + O(\alpha g^k).$$

Now we subtract the first equation from the second one

$$f(x^k + \alpha g^k) - f(x^k + \alpha p) = \alpha(\nabla f(x^k)^T g^k - \nabla f(x^k)^T p) + O(\alpha g^k) - O(\alpha p)$$
$$\frac{f(x^k + \alpha g^k) - f(x^k + \alpha p)}{\alpha} = \nabla f(x^k)^T g^k - \nabla f(x^k)^T p + \frac{O(\alpha g^k) - O(\alpha p)}{\alpha}.$$

Since $\nabla f(x^k)^T g^k$ is a negative number and $0 \leq c < 1$, we must have $\nabla f(x^k)^T g^k - \nabla f(x^k)^T p$ being a strictly negative number.

From Taylor Expansion we have

$$\lim_{\alpha \to 0^+} \frac{O(\alpha g^k)}{\alpha} = \lim_{\alpha \to 0^+} \frac{O(\alpha g^k)}{\|\alpha g^k\|_2} = 0, \quad \lim_{\alpha \to 0^+} \frac{O(\alpha p)}{\alpha} = \lim_{\alpha \to 0^+} \frac{O(\alpha p)}{||\alpha p||_2} = 0.$$

Hence for $\alpha > 0$ sufficiently small, we always have $f(x^k + \alpha g^k) - f(x^k + \alpha p) < 0$, therefore p can not be the minimizer, and g^k is the only minimizer as $\alpha \to 0^+$. □

8.2 Line Search Rules

Construct a function as following

$$\phi(\alpha) = f(x^k + \alpha d^k).$$

A natural choice is to find α_k such that

$$\alpha_k = \arg\min_{\alpha > 0} \phi(\alpha),$$

i.e., α is the best possible step size. This is called *exact line search*.

However, exact line search is usually computationally expensive, therefore not used much in practice. Instead we use *inexact line search*.

Consider the following example.

Example.

$$\min_x f(x) = x^2,$$

The initial point is $x^0 = 1$. Choose $d^k = -\text{sign}(x^k)$ and require $f(x^k + \alpha_k d^k) < f(x^k)$. Consider the following two step sizes,

$$\alpha_{k,1} = \frac{1}{3^{k+1}}, \quad \alpha_{k,2} = 1 + \frac{2}{3^{k+1}}$$

By simple calculation, we get

$$x_1^k = \frac{1}{2}\left(1 + \frac{1}{3^k}\right), \quad x_2^k = \frac{(-1)^k}{2}\left(1 + \frac{1}{3^k}\right).$$

Although both $\{f(x_1^k)\}$ and $\{f(x_2^k)\}$ are monotone decreasing, none of them converges to the minimum point.

8.2.1 Armijo Rule

DEFINITION 8.5. (Armijo Rule)

Let d^k be a descent direction, if

$$f(x^k + \alpha d^k) \leq f(x^k) + c\alpha \nabla f(x^k)^T d^k,$$

we say α satisfies **Armijo condition**, where $c \in (0, 1)$ is a constant.

Armijo condition is also called sufficient decrease condition.

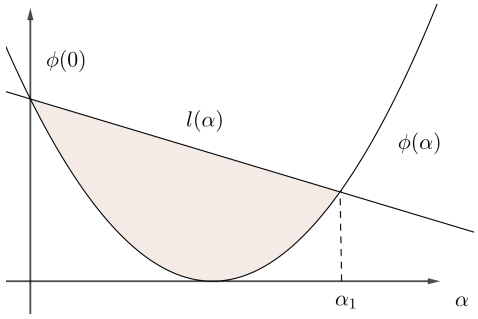


Figure 11: Armijo Rule

$$\begin{cases} \phi(\alpha) = f(x^k + \alpha d^k) \\ \phi'(\alpha) = \nabla f(x^k + d^k)^T d^k \\ \phi'(0) = \nabla f(x^k)^T d^k \\ l(\alpha) = f(x^k) + c \cdot \alpha \nabla f(x^k)^T d^k,\ 0 < c < 1 \\ l'(\alpha) = c \cdot \phi'(\alpha) \end{cases}$$

Drawback of Armijo Rule: $\alpha = 0$ satisfies Armijo Rule $\Rightarrow l(\alpha) = f(x^k)$, step size could be too small.

Algorithm 1: Back Track Armijo Method

Input: Choose parameters $\gamma, c \in (0, 1)$. Initialize $\alpha \leftarrow \bar{\alpha}$.
while $f(x^k + \alpha d^k) > f(x^k) + c\alpha \nabla f(x^k)^T d^k$ **do**
| Set $\alpha \leftarrow \gamma\alpha$
end
Output: $\alpha_k = \alpha$.

8.2.2 Goldstein Rule

DEFINITION 8.6. (Goldstein Rule)

Let d^k be a descent direction, if

$$f(x^k + \alpha d^k) \leq f(x^k) + c\alpha \nabla f(x^k)^T d^k,$$
$$f(x^k + \alpha d^k) \geq f(x^k) + (1-c)\alpha \nabla f(x^k)^T d^k$$

We say α satisfies **Goldstein Condition**, where $c \in \left(0, \frac{1}{2}\right)$ is a constant.

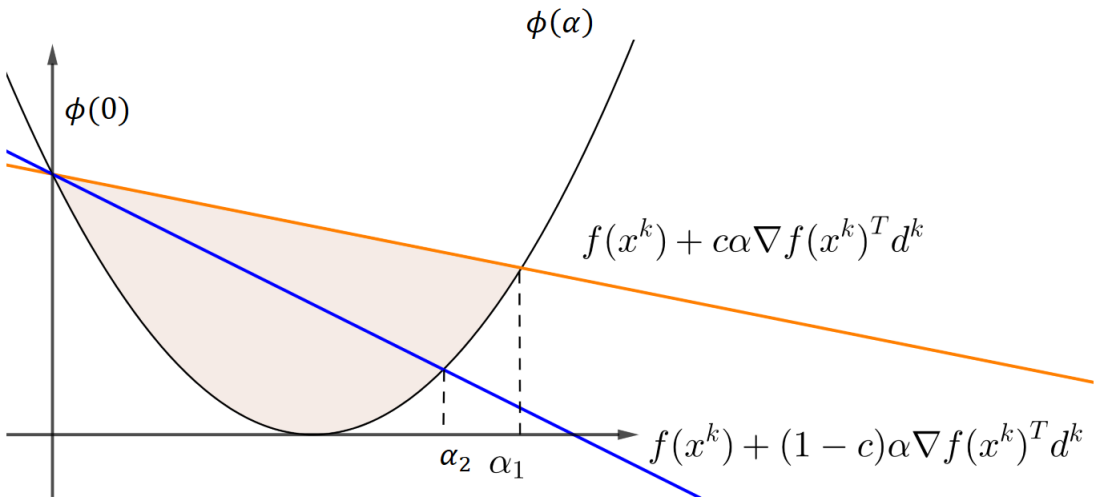


Figure 12: Goldstein Condition

Goldstein rule may avoid the optimal step size.

8.2.3 Wolfe Rule

DEFINITION 8.7. (Wolfe Rule)

Let d^k be a descent direction, if

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^T d^k, \tag{8.1}$$
$$\nabla f(x^k + \alpha d^k)^T d^k \geq c_2 \nabla f(x^k)^T d^k \quad \text{(Curvature condition)} \tag{8.2}$$

we say α satisfies **Wolfe Condition**, where $c_1, c_2 \in (0,1)$ are constants and $c_1 < c_2$.

Note: Since an optimal α^* must be a stationary point (i.e., $\nabla f(x^k + \alpha^* d^k)^T d^k = 0$), and d^k is a descent direction, therefore $\alpha \nabla f(x^k)^T d^k < 0$. So α^* must satisfy the Wolfe condition.

DEFINITION 8.8. (Strong Wolfe Rule)

Let d^k be a descent direction, if

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^T d^k,$$
$$|\nabla f(x^k + \alpha d^k)^T d^k| \leq c_2 |\nabla f(x^k)^T d^k|,$$

we say α satisfies **Strong Wolfe Condition**, where $c_1, c_2 \in (0, 1)$ are constants and $c_1 < c_2$.

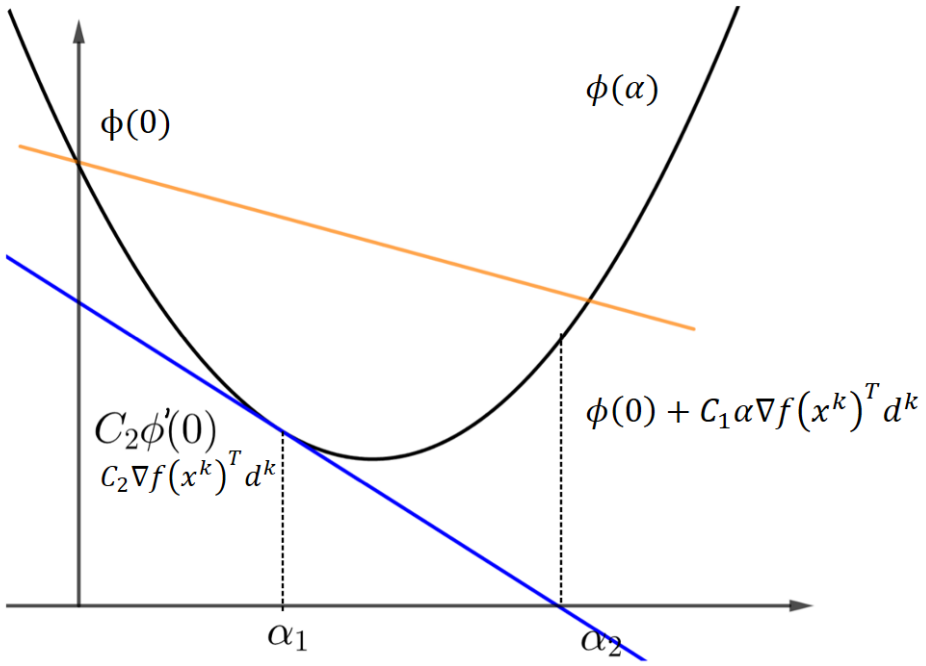


Figure 13: Wolfe Condition

THEOREM 8.9. Suppose that $f : \mathbb{R}^n \to \mathbb{R} \in C^1$, let d^k be a descent direction at x^k. Assume $\phi(\alpha)$ is bounded below for $\alpha > 0$. If $0 < c_1 < c_2 < 1$, then there exists intervals of step lengths α satisfying the Wolfe Condition and Strong Wolfe Condition.

Proof. Since $0 < c_1 < 1$, the line $l(\alpha) = f(x^k) + \alpha c_1 \nabla f(x^k)^T d^k$ is unbounded below. Also $\phi(\alpha)$ is bounded below, hence they must intersect at least once. Let α_1 be the smallest intersecting value of α, that is

$$f(x^k + \alpha_1 d^k) = f(x^k) + \alpha_1 c_1 \nabla f(x^k)^T d^k$$

The Armijo condition holds for all step lengths less than α_1.

By first order Taylor expansion, there exists $\alpha_2 \in (0, \alpha_1)$ such that

$$f(x^k + \alpha_1 d^k) - f(x^k) = \alpha_1 \nabla f(x^k + \alpha_2 d^k)^T d^k$$

By combining the two equations above, since $0 < c_1 < c_2 < 1$ and $\nabla f(x^k)^T d^k < 0$, we obtain

$$\nabla f(x^k + \alpha_2 d^k)^T d^k = c_1 \nabla f(x^k)^T d^k > c_2 \nabla f(x^k)^T d^k.$$

Therefore α_2 satisfies the Wolfe condition, and the inequalities holds strictly in both (8.1) and (8.2). Hence by the smoothness assumption of f, there is an interval around α_2 such that the Wolfe condition holds. Moreover, since $\nabla f(x^k + \alpha_2 d^k)^T d^k = c_1 \nabla f(x^k)^T d^k < 0$, the Strong Wolfe Condition also holds in the same interval. □

8.3 Convergence Analysis of Armijo backtracking line search

DEFINITION 8.10. (Lipschitz Continuous)

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *Lipschitz continuous* with constant $L \geq 0$ if for any $x, y \in \mathbb{R}^n$,

$$|f(y) - f(x)| \leq L \|y - x\| .$$

Note: Lipschitz continuous implies uniformly continuous, but not the the way around.

For example $y = \sqrt{|x|}$ is uniformly continuous but not Lipschitz continuous.

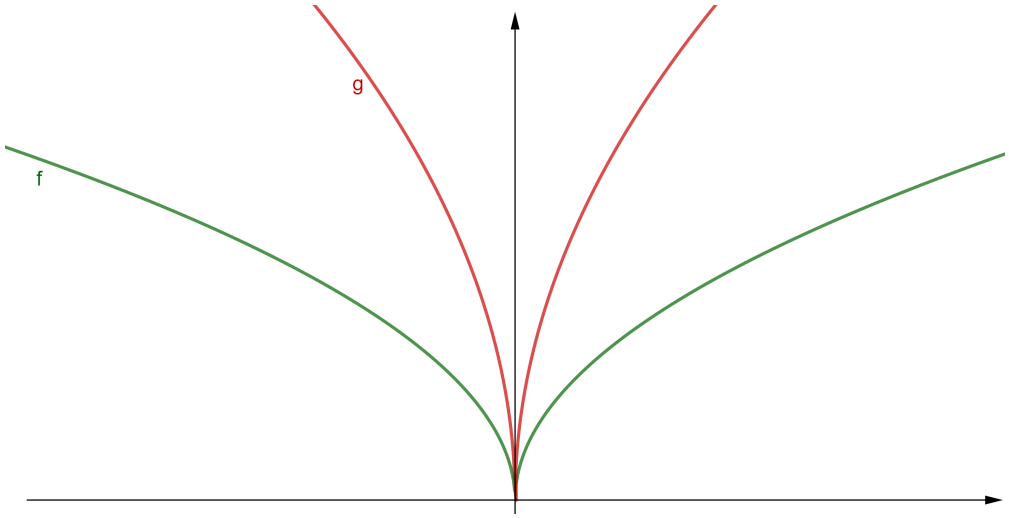


Figure 14: Uniform but not Lipschitz continuous functions

Lemma 8.11. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function ($f \in C^1$). Then f is Lipschitz continuous with constant L if and only if for all $x \in D$, $\|\nabla f(x)\| \leq L$.

Proof. Let $\|\nabla f(x)\| \leq L$ for any $x \in \mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$ we have

$$f(y) = f(x) + \nabla f(x + \theta(y - x))^T (y - x),\ 0 < \theta < 1.$$

Hence

$$\begin{aligned} |f(y) - f(x)| &= |\nabla f(x + \theta(y - x))^T (y - x)| \\ &\leq \|\nabla f(x + \theta(y - x)\| \cdot \|y - x\| \quad \text{(Cauchy-Schwartz Inequality)} \\ &\leq L \|y - x\| \end{aligned}$$

For the other direction, let's assume $|f(y) - f(x)| \leq L \|y - x\|$ for all $x, y \in \mathbb{R}^n$.

Let $y = x + \alpha \nabla f(x), \alpha > 0$. Then $f(y) = f(x) + \alpha \nabla f(x + \theta(y - x))^T \nabla f(x)$ for some $\theta \in (0, 1)$. Hence

$$|f(y) - f(x)| = |\alpha| \cdot |\nabla f(x + \theta(y - x))^T \nabla f(x)| \leq L \|y - x\| = L|\alpha| \cdot \|\nabla f(x)\|.$$

By dividing α from both sides

$$|\nabla f(x + \alpha\theta \nabla f(x))^T \nabla f(x)| \leq L \|\nabla f(x)\|.$$

Let $\alpha \to 0^+$, then by continuity of $\nabla f(x)$ ($f \in C^1$), we have

$$\begin{aligned} \lim_{\alpha \to 0} |\nabla f(x + \alpha\theta \nabla f(x))^T \nabla f(x)| &= |\nabla f(x + \lim_{\alpha \to 0} \alpha\theta \nabla f(x))^T \nabla f(x)| \\ &= \|\nabla f(x)\|^2 \leq L \|\nabla f(x)\| \end{aligned}$$

Hence $\|\nabla f(x)\| \leq L$, the other direction is proved. □

DEFINITION 8.12. (Lipschitz continuous gradient)

A function $f : \mathbb{R}^n \to \mathbb{R} \in C^1$ has a *Lipschitz continuous gradient* with constant $L \geq 0$ if for any $x, y \in \mathbb{R}^n$,

$$|\nabla f(y) - \nabla f(x)| \leq L \|y - x\|$$

THEOREM 8.13. (Convergence for Armijo Backtracking Line Search)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be such that f is differentiable and the gradient $\nabla f(x)$ is Lipschitz continuous with constant L, i.e., for any $x, y \in \mathbb{R}^n$,

$$|\nabla f(y) - \nabla f(x)| \leq L \|y - x\|$$

Let $x^{k+1} = x^k + \alpha_k d^k$ where step size α_k is obtained by Armijo backtracking and d^k is a descent direction. Assume

1. $f(x^k) > -\infty$. ($f(x^k)$ is bounded below)
2. The descent direction is bounded, i.e., $\left\|d^k\right\| < +\infty$.

Then we have

$$\lim_{k \to \infty} \nabla f(x^k)^T d^k = 0.$$

Proof. Suppose on the contrary the conclusion does not hold. Then there must exist a subsequence $J \subseteq \mathbb{Z}^+$ such that

$$\sup_{k \in J} \nabla f(x^k)^T d^k = \eta < 0.$$

From the fact the Armijo condition is satisfied we have

$$f(x^{k+1}) - f(x^k) \leq c\alpha_k \nabla f(x^k)^T d^k.$$

Since both sides are negative and $c \in (0, 1)$, we have

$$|f(x^{k+1}) - f(x^k)| \geq c|\alpha_k \nabla f(x^k)^T d^k|.$$

Since $f(x^k)$ is a bounded decreasing sequence (d^k is a descent direction), it converges by the Monotone Convergence Theorem. Then, we also have $f(x^{k+1}) - f(x^k) \to 0$ since convergent sequences are Cauchy. Then we can get

$$f(x^{k+1}) - f(x^k) \to 0,$$

which implies

$$\alpha_k \nabla f(x^k)^T d^k \to 0.$$

Since $\nabla f(x^k)^T d^k = \eta < 0$ for $k \subseteq J$, we must have $\alpha_k \to 0$. So we can assume $\alpha_k < 1$ for k sufficiently large, as a consequence, we must have done at least one backtracking step inside the while loop in the Armijo backtrack method (Definition 1). (Why? WLOG the initial step size can be chosen as 1.)

Therefore $\alpha_k \gamma^{-1}$ must be a step size that violates the Armijo condition, let $\alpha_k \gamma^{-1} = \beta_k$, we have

$$f(x^k + \beta_k d^k) > f(x^k) + c\beta_k \nabla f(x^k)^T d^k \tag{8.3}$$

By Taylor Expansion, there exists $0 < \theta < 1$ such that

$$\begin{aligned}
f(x^k + \beta_k d^k) - f(x^k) &= \beta_k \nabla f(x^k + \theta\beta_k d^k)^T d^k \\
&= \beta_k \nabla f(x^k)^T d^k + \beta_k \nabla f(x^k + \theta\beta_k d^k)^T d^k - \beta_k \nabla f(x^k)^T d^k \\
&= \beta_k \nabla f(x^k)^T d^k + \beta_k (\nabla f(x^k + \theta\beta_k d^k) - \nabla f(x^k))^T d^k \\
&\leq \beta_k \nabla f(x^k)^T d^k + \beta_k \left\| (\nabla f(x^k + \theta\beta_k d^k) - \nabla f(x^k)) \right\| \cdot \left\| d^k \right\| \quad \text{(Cauchy-Schwarz)} \\
&\leq \beta_k \nabla f(x^k)^T d^k + \beta_k L \left\| \theta\beta_k d^k \right\| \cdot \left\| d^k \right\| \quad \text{(Lipschitz gradient)} \\
&= \beta_k \nabla f(x^k)^T d^k + \beta_k^2 L\theta \left\| d^k \right\|^2
\end{aligned}$$

Combining with equation (8.3), we have

$$\beta_k \nabla f(x^k)^T d^k + \beta_k^2 L\theta \left\| d^k \right\|^2 > c\beta_k \nabla f(x^k)^T d^k$$

which means

$$(1-c)\nabla f(x^k)^T d^k + \beta_k L\theta \left\| d^k \right\|^2 > 0 \tag{8.4}$$

Since $\nabla f(x^k)^T d^k < \eta$ and $1 - c > 0$, we have

$$(1-c)\eta + \beta_k L\theta \left\| d^k \right\|^2 > 0$$

Let $k \to \infty$, then $\beta_k \to 0$, also $\left\| d^k \right\|$ is bounded, which implies

$$(1-c)\eta \geq 0$$

which is a contradiction to $\eta < 0$, hence the theorem is proved. □

Corollary 8.14. Convergence of normalized gradient descent

Let $f : \mathbb{R}^n \to \mathbb{R}$ be such that f is differentiable and the gradient $\nabla f(x)$ is Lipschitz continuous with constant L. Let $d^k = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}$ in the Armijo backtracking line search algorithm, then every accumulation point x^*, of the sequence $\{x^k\}$ is a stationary point (i.e., $\nabla f(x^*) = 0$).

Proof. First since d^k is a descent direction, therefore the sequence $\{f(x^k)\}$ is decreasing. If x^* is any

accumulation point of the sequence $\{x^k\}$, then we claim that $f(x^*)$ is a lower bound for the sequence $\{f(x^k)\}$. (why? prove this claim as an exercise.)

Since the sequence $\{f(x^k)\}$ is bounded below, Theorem 8.13 applies. That is

$$\begin{aligned} 0 &= \lim_{k\to\infty} \nabla f(x^k)^T \left(-\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|} \right) \\ &= \lim_{k\to\infty} -\left\|\nabla f(x^k)\right\|. \end{aligned}$$

Since ∇f is continuous, we have $\nabla f(x^*) = 0$. □

Corollary 8.15. Let $f : \mathbb{R}^n \to \mathbb{R}$ be such that f is differentiable and the gradient $\nabla f(x)$ is Lipschitz continuous with constant L. Assume that d^k is the descent direction at step k in the Armijo backtracking line search algorithm and $\|d_k\| = 1$ (The step size is chosen by Armijo backtracking line search). Let $\theta_k \in [0, \pi/2]$ be the angle between the negative gradient direction $-\nabla f(x^k)$ and the descent direction d^k. If $\theta_k \leq \frac{\pi}{2} - \epsilon$ for some constant $\epsilon > 0$ when k is sufficiently large. Then every accumulation point x^*, of the sequence $\{x^k\}$ is a stationary point (i.e., $\nabla f(x^*) = 0$)

Proof. Exercise. □

In fact, we can obtain the same result without assuming the descent direction d^k is normalized.

THEOREM 8.16. Let $f : \mathbb{R}^n \to \mathbb{R}$ be such that f is differentiable and the gradient $\nabla f(x)$ is Lipschitz continuous with constant L. Let d^k be a descent direction at step k in the Armijo backtracking line search algorithm. Let $\theta_k \in [0, \pi/2]$ be the angle between the negative gradient direction $-\nabla f(x^k)$ and the descent direction d^k. If $\theta_k \leq \frac{\pi}{2} - \epsilon$ for some constant $\epsilon > 0$ when k is sufficiently large. Then every accumulation point x^*, of the sequence $\{x^k\}$ (generated by Armijo backtracking line search algorithm) is a stationary point (i.e., $\nabla f(x^*) = 0$).

Proof. First since d^k is a descent direction, therefore the sequence $\{f(x^k)\}$ is decreasing. If x^* is any accumulation point of the sequence $\{x^k\}$, then we know that $f(x^*)$ is a lower bound for the sequence $\{f(x^k)\}$.

From the proof of Theorem 8.13, we can derive a lower bound for α^k. From equation (8.4), we have

$$\beta_k \theta > \frac{(c-1)\nabla f(x^k)^T d^k}{L \left\|d^k\right\|^2}$$

Therefore

$$\alpha_k \gamma^{-1} = \beta_k \geq \beta_k \theta > \frac{(c-1)\nabla f(x^k)^T d^k}{L \left\|d^k\right\|^2}$$

Hence

$$\alpha_k > \frac{\gamma(c-1)\nabla f(x^k)^T d^k}{L \left\|d^k\right\|^2}$$

By the assumption of θ_k we know $-\nabla f(x^k)^T d^k = \cos(\theta_k) \left\|\nabla f(x^k)\right\| \left\|d^k\right\| \geq \delta \left\|\nabla f(x^k)\right\| \left\|d^k\right\|$ for some $\delta > 0$. Then by sufficient decrease condition, we have

$$f(x^{k+1}) - f(x^k) \leq c\alpha^k \nabla f(x^k)^T d^k < \frac{\gamma c(c-1)\delta^2}{L} \left\|\nabla f(x^k)\right\|^2 \leq 0$$

Hence

$$|f(x^{k+1}) - f(x^k)| \geq \frac{\gamma c(1-c)\delta^2}{L} \left\|\nabla f(x^k)\right\|^2 \geq 0 \tag{8.5}$$

Therefore, if $f(x^k)$ is bounded below (guaranteed by the existence of an accumulation point), we have $\lim_{k\to\infty} |f(x^{k+1}) - f(x^k)| = 0$, which implies $\nabla f(x^k) \to 0$ as $k \to \infty$. □

Corollary 8.17. Convergence of gradient descent

Let $f : \mathbb{R}^n \to \mathbb{R}$ be such that f is differentiable and the gradient $\nabla f(x)$ is Lipschitz continuous with constant L. If $d^k = -\nabla f(x^k)$ in Armijo backtracking line search algorithm, then every accumulation point x^*, of the sequence $\{x^k\}$ (generated by Armijo backtracking line search algorithm) is a stationary point (i.e., $\nabla f(x^*) = 0$).

Proof. The angle θ_k between d^k and the negative gradient is 0 in this case, therefore the result follows from Theorem 8.16. □

8.4 Convergence Analysis, Zoutedijk's Theorem

THEOREM 8.18. (Zoutendijk Condition)

Consider $x^{k+1} = x^k + \alpha_k d^k$, where α is the step size and d^k is a descent direction. and the iterations satisfy Wolfe condition. Assume the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is bounded below, continuously differentiable (C^1) and $\nabla f(x)$ is Lipschitz continuous with constant L. Then

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \left\|\nabla f(x^k)\right\|^2 < +\infty,$$

where $\cos\theta_k$ is cosine of the angle between the negative gradient $-\nabla f(x^k)$ and the descent direction d^k, i.e.,

$$\cos\theta_k = \frac{-\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\| \, \|d^k\|}.$$

Proof. From the curvature condition, we have

$$(\nabla f(x^{k+1}) - \nabla f(x^k))^T d^k \geq (c_2 - 1)\nabla f(x^k)^T d^k.$$

By Cauchy-Schwarz inequality an Lipschitz continuity of the gradient, we have

$$(\nabla f(x^{k+1}) - \nabla f(x^k))^T d^k \leq \left\|\nabla f(x^{k+1}) - \nabla f(x^k)\right\| \left\|d^k\right\| \leq \alpha_k \left\|d^k\right\|^2.$$

Combining the two equality above, we have

$$\alpha_k \geq \frac{c_2 - 1}{L} \frac{\nabla f(x^k)^T d^k}{\|d^k\|^2}.$$

Note $\nabla f(x^k)^T d^k < 0$, substitute α_k into the sufficient decrease condition, we have

$$f(x^{k+1}) \leq f(x^k) + c_1 \frac{c_2 - 1}{L} \frac{(\nabla f(x^k)^T d^k)^2}{\|d^k\|^2}.$$

By the definition of θ_k, this is equivalent to

$$f(x^{k+1}) \leq f(x^k) + c_1 \frac{c_2 - 1}{L} \cos^2 \theta_k \left\|\nabla f(x^k)\right\|^2.$$

Sum over all k, we have

$$f(x^{k+1}) \leq f(x^0) - c_1 \frac{1 - c_2}{L} \sum_{j=0}^{k} \cos^2 \theta_j \left\|\nabla f(x^j)\right\|^2.$$

Since $f(x)$ is bounded below, and from $0 < c_1 < c_2 < 1$ we have $c_1(1 - c_2) > 0$, therefore when $k \to \infty$

$$\sum_{j=0}^{\infty} \cos^2 \theta_j \left\|\nabla f(x^j)\right\|^2 < +\infty.$$

□

Corollary 8.19. For a liner search iteration $x^{k+1} = x^k + \alpha_k d^k$, let θ_k be the angle between the negative gradient $-\nabla f(x^k)$ and the descent direction d^k, and assume for any k, there exists a constant $\gamma > 0$ such that

$$\theta_k \leq \frac{\pi}{2} - \gamma,$$

Then under the assumption of Theorem 8.18, we have

$$\lim_{k \to \infty} \nabla f(x^k) = 0$$

Proof. Assume on the contrary that the conclusion is not true, then there exists a subsequence $\{k_l\}$ and positive constant $\delta > 0$ such that

$$\left\|\nabla f(x^{k_l})\right\| \geq \delta, \quad l = 1, 2, ...$$

By the assumption of θ_k, for any k

$$\cos \theta_k > \sin \gamma > 0.$$

Therefore we have

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \left\|\nabla f(x^k)\right\|^2 \geq \sum_{l=1}^{\infty} \cos^2 \theta_{k_l} \left\|\nabla f(x^{k_l})\right\|^2 \geq \sum_{l=1}^{\infty} (\sin^2 \gamma)\delta^2 \to \infty,$$

which is a contradiction to Theorem 8.18, therefore the corollary is proved. □

8.5 Convergence of Gradient Descent Algorithm

Consider the *gradient descent* step

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k),$$

where α_k is a step size at step k, we assume $\alpha_k \geq 0$.

In corollary 8.17 and corollary 8.19, we know that if the function f being minimized has a Lipschitz continuous gradient, then any accumulation point generated by gradient descent method is a stationary point, provided that the step size is carefully chosen (Armijo backtracking or Wolfe condition). Now we study the behavior of gradient descent method under the assumption that function f is convex. In fact, under convexity assumption, we can use a constant step size and obtain some results about the convergence rate of the gradient descent algorithm.

Lemma 8.20. Suppose function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is differentiable, $\operatorname{dom} f$ is convex, and its gradient $\nabla f(x)$ is Lipschitz continuous with constant L, then the following holds

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2, \forall x, y \in \operatorname{dom} f \tag{8.6}$$

Proof. For any $x, y \in \operatorname{dom} f$, we construct a function

$$g(t) = f(x + t(y - x)), t \in [0, 1].$$

Obviously, $g(0) = f(x), g(1) = f(y)$, and

$$g'(t) = \nabla f(x + t(y - x))^T (y - x).$$

In particular

$$g'(0) = \nabla f(x)^T (y - x).$$

By Newton-Lebnitz Theorem, we have

$$g(1) - g(0) = \int_0^1 g'(t) dt.$$

Note $g(t)$ is well-defined for $t \in [0, 1]$ since $\operatorname{dom} f$ is convex. Therefore

$$\begin{aligned}
f(y) - f(x) - \nabla f(x)^T (y - x) &= \int_0^1 g'(t) dt - g'(0) \\
&= \int_0^1 ((g'(t) - g'(0)) dt \\
&= \int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x) dt \\
&\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \|y - x\| dt \quad \text{(Cauchy–Schwarz Inequality)} \\
&\leq \int_0^1 L \|y - x\|^2 t dt \quad \text{(Lipschitz continuity of gradient)} \\
&= \frac{L}{2} \|y - x\|^2
\end{aligned}$$

□

THEOREM 8.21. (Convergence of Gradient Descent Algorithm for Convex Function)

Suppose $f(x)$ is a convex function, and $\nabla f(x)$ is Lipschitz continuous with constant L. Assume there exists a point $x^* \in \operatorname{dom} f$ such that

$$f(x^*) = \inf_{x \in \operatorname{dom} f} f(x)$$

(the infimum can be attained). If the step size α_k is chosen to be a constant α such that $0 < \alpha < \frac{1}{L}$. Then the sequence $\{f(x^k)\}$ obtained by iteration $x^{k+1} = x^k - +\alpha_k \nabla f(xd^k)$ where d^k is a descent direction is converging to the optimal value, i.e.,

$$\lim_{k \to \infty} f(x^k) = f(x^*),$$

and the convergence rate is $O(\frac{1}{k})$ (in terms of optimal value).

Proof. Since $\nabla f(x)$ is Lipschitz continuous, then for any $x^k \in \operatorname{dom} f$, Let $x^{k+1} = x^k - \alpha \nabla f(x^k)$, by equation (8.6), we have

$$f(x^{k+1}) \leq f(x^k) - \alpha \left(1 - \frac{L\alpha}{2}\right) \left\|\nabla f(x^k)\right\|^2 \tag{8.7}$$

Since $0 < \alpha < \dfrac{1}{L}$, we have $1 - \dfrac{L\alpha}{2} > \dfrac{1}{2}$.

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) - \frac{\alpha}{2} \left\|\nabla f(x^k)\right\|^2 \quad \text{(by equation (8.7))} \\
&\leq f(x^*) + \nabla f(x^k)^T (x^k - x^*) - \frac{\alpha}{2} \left\|\nabla f(x^k)\right\|^2 \quad \text{(by Proposition 4.8, } f(x) \text{ is convex)} \\
&= f(x^*) + \nabla f(x^k)^T (x^k - x^* - \frac{\alpha}{2} \nabla f(x^k)) \\
&= f(x^*) + \frac{1}{2\alpha} ((x^k - x^*) - (x^k - x^* - \alpha \nabla f(x^k)))^T (2(x^k - x^*) - \alpha \nabla f(x^k)) \\
&= f(x^*) + \frac{1}{2\alpha} ((x^k - x^*) - (x^k - x^* - \alpha \nabla f(x^k)))^T ((x^k - x^*) + (x^k - x^* - \alpha \nabla f(x^k)) \\
&= f(x^*) + \frac{1}{2\alpha} (\left\|x^k - x^*\right\|^2 - \left\|x^k - x^* - \alpha \nabla f(x^k)\right\|^2) \\
&= f(x^*) + \frac{1}{2\alpha} (\left\|x^k - x^*\right\|^2 - \left\|x^{k+1} - x^*\right\|^2)
\end{aligned}$$

Now we make a summation of the above inequality for $k = 0, 1, ..., t-1$, i.e.,

$$\begin{aligned}
\sum_{k=0}^{t-1} \left(f(x^{k+1}) - f(x^*)\right) &\leq \frac{1}{2\alpha} \sum_{k=0}^{t-1} \left(\left\|x^k - x^*\right\|^2 - \left\|x^{k+1} - x^*\right\|^2\right) \\
&= \frac{1}{2\alpha} \left(\left\|x^0 - x^*\right\|^2 - \left\|x^t - x^*\right\|^2\right) \\
&\leq \frac{1}{2\alpha} \left(\left\|x^0 - x^*\right\|^2\right)
\end{aligned}$$

(From the above inequalities, what can you say about the sequence $\{x^t\}$?) Since $\{f(x^k)\}$ is a non-increasing sequence, we have

$$f(x^t) - f(x^*) \leq \frac{1}{t} \sum_{k=0}^{t-1} (f(x^k) - f(x^*)) \leq \frac{1}{2t\alpha} \left\|x^0 - x^*\right\|^2.$$

Hence we have $f(x^t) - f(x^*) \sim O\left(\frac{1}{t}\right)$. Moreover, by continuity of f, we have

$$0 \leq -f(x^*) + \lim_{t\to\infty} f(x^t) = \lim_{t\to\infty} \left(f(x^t) - f(x^*)\right) \leq \lim_{t\to\infty} \frac{1}{2t\alpha} \left\|x^0 - x^*\right\|^2 = 0,$$

which implies that

$$f(x^*) = \lim_{t\to\infty} f(x^t),$$

as required. □

8.6 Strongly Convex Function

Lemma 8.22. (Gradient of Convex Function Is Monotone)

Let $f : \mathbb{R}^n \to \mathbb{R}$ to be a differentiable function. Then f is convex if and only if f has a *monotone gradient*,

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0, \forall x, y \in \mathbb{R}^n$$

Proof. Assume $f(x)$ is convex, then

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

and

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

Adding the above inequalities together, we have

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0.$$

The other direction is left as an exercise. □

DEFINITION 8.23. (Strongly Convex Function)

A function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is *M-strongly convex* if there exists a constant $M > 0$ such that the function

$$g(x) = f(x) - \frac{M}{2} \|x\|^2$$

is also convex. That is, a strongly convex function is a convex function plus a quadratic term.

Lemma 8.24.

- Let $f(x)$ be a differentiable function. Then $f(x)$ is M-strongly convex if and only if

$$\forall x, y,\ (\nabla f(x) - \nabla f(y))^T (x - y) \geq M \|x - y\|_2^2$$

- Let $f(x)$ be a C^2-class function. Then $f(x)$ is M-strongly convex if and only if

$$\forall x,\ \nabla^2 f(x) - MI \succeq 0$$

Proof. Exercise □

8.7 Lipschitz Continuity of Gradient

Lemma 8.25. Let $f : \mathbb{R}^n \to \mathbb{R}$ to be a differentiable and convex function. Then the following statements are equivalent

- $\nabla f(x)$ is Lipschitz continuous with constant L.
- $g(x) = \frac{L}{2}\|x\|^2 - f(x)$ is convex.
- For any $x, y \in \mathbb{R}^n$, we have

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 .$$

Proof.
From 1 to 2:

We just need to check $\nabla g(x)$ is monotone. Since $\nabla g(x) = Lx - \nabla f(x)$ and $\nabla g(y) = Ly - \nabla f(y)$, we have

$$\begin{aligned}(\nabla g(x) - \nabla g(y))^T (x - y) &= (L(x - y)) - (\nabla f(x) - \nabla f(y))^T (x - y) \\ &= L\|x - y\|^2 - (\nabla f(x) - \nabla f(y))^T (x - y) \\ &\geq L\|x - y\|^2 - L\|x - y\|^2 = 0. \quad \text{(by Lipschitz)}\end{aligned}$$

Therefore $g(x)$ is convex.

From 2 to 3:

Given any $x \in \mathbb{R}^n$, we construct function

$$g_x(z) = \frac{L}{2}\|z\|^2 - f(z) + \nabla f(x)^T z$$

By our assumption $g(z)$ is convex, a convex function plus a linear term is also convex, therefore $g_x(z)$ is convex.

By convexity of $g_x(z)$, we have

$$\forall z_1, z_2 \in \mathbb{R}^n,\ g_x(z_2) \geq g_x(z_1) + \nabla g_x(z_1)^T (z_2 - z_1), \tag{8.8}$$

Denote $f_x(z) = f(z) - \nabla f(x)^T z$, then we can write $g_x(z) = \frac{L}{2}\|z\|^2 - f_x(z)$.

Substitute $g_x(z)$ into equation (8.8), we have

$$\frac{L}{2}\|z_2\|^2 - f_x(z_2) \geq \frac{L}{2}\|z_1\|^2 - f_x(z_1) + (Lz_1 - \nabla f_x(z_1))^T(z_2 - z_1)$$

which is equivalent to

$$f_x(z_2) \leq f_x(z_1) + \nabla f_x(z_1)^T(z_2 - z_1) + \frac{L}{2}\|z_2 - z_1\|^2$$

Let $z_2 = z$, $z_1 = y$, we have

$$f_x(z) \leq f_x(y) + \nabla f_x(y)^T(z - y) + \frac{L}{2}\|z - y\|^2.$$

Notice $\nabla f_x(z) = 0$ at $z = x$, i.e., $\nabla f_x(x) = 0$ and $f_x(z)$ is convex function (a convex function plus a linear term is also convex). Therefore x is a global minimizer of $f_x(z)$. By problem 2 in the practice mid term, we can show

$$f_x(x) \leq f_x(y) - \frac{1}{2L}\|\nabla f_x(y)\|^2.$$

Since $\nabla f_x(y) = \nabla f(y) - \nabla f(x)$, $f_x(x) = f(x) - \nabla f(x)^T x$ and $f_x(y) = f(y) - \nabla f(x)^T y$, we have

$$\frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \leq f(y) - f(x) - \nabla f(x)^T(y - x)$$

By interchanging x and y, we have

$$\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(x) - f(y) - \nabla f(y)^T(x - y)$$

Adding the above two inequalities together, we have the desired inequality

$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

From 3 to 1:

Apply Cauchy-Schwarz Inequality. Easy □

8.8 Convergence of Gradient Descent Algorithm for Strongly Convex Function

THEOREM 8.26. (Gradient Descent, Strongly Convex)

Suppose $f(x)$ is a M-strongly convex function and $\nabla f(x)$ is L-Lipschitz continuous. Let x^* be such that $f(x^*) = \inf_x f(x)$. Then if the step size $\alpha \in \left(0, \frac{2}{M+L}\right)$, the sequence $\{x^k\}$ produced by gradient descent algorithm converges to x^* Q-linearly.

Proof. First since $f(x)$ is M-strongly convex function, by definition of strongly convex function, we have

$$g(x) = f(x) - \frac{M}{2}\|x\|^2$$

is a convex function.

Since $f(x)$ is Lipschitz continuous, we have $\frac{L}{2}\|x\|^2 - f(x)$ is convex by Lemma 8.25. Therefore

$$\frac{L}{2}\|x\|^2 - f(x) = \frac{L-M}{2}\|x\|^2 - g(x)$$

is convex.

Since $\frac{L-M}{2}\|x\|^2 - g(x)$ is convex, again by Lemma 8.25, we have

$$(\nabla g(x) - \nabla g(y))^T(x-y) \geq \frac{1}{L-M}\|\nabla g(x) - \nabla g(y)\|^2.$$

Plugging in $g(x) = f(x) - \frac{M}{2}\|x\|^2$ we have

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{ML}{M+L}\|x-y\|^2 + \frac{1}{M+L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Let $x = x^k$, $y = x^*$ and the fact that x^* is a global minimizer implies $\nabla f(x^*) = 0$, we have

$$\nabla f(x^k)^T(x^k - x^*) \geq \frac{ML}{M+L}\left\|x^k - x^*\right\|^2 + \frac{1}{M+L}\left\|\nabla f(x^k)\right\|^2. \tag{8.9}$$

Let $\alpha \in \left(0, \frac{2}{M+L}\right)$. Then

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &= \left\|x^k - \alpha\nabla f(x^k) - x^*\right\|^2 \\
&= \left\|x^k - x^*\right\|^2 - 2\alpha\nabla f(x^k)^T(x^k - x^*) + \alpha^2\left\|\nabla f(x^k)\right\|^2 \\
&\leq (1 - \alpha\frac{2ML}{M+L})\left\|x^k - x^*\right\|^2 + \alpha(\alpha - \frac{2}{M+L})\left\|\nabla f(x^k)\right\|^2 \quad \text{(by (8.9))} \\
&\leq (1 - \alpha\frac{2ML}{M+L})\left\|x^k - x^*\right\|^2. \quad (\text{since } \alpha - \tfrac{2}{M+L} < 0)
\end{aligned}$$

It is easy to see $\left(1 - \alpha\frac{2ML}{M+L}\right) \in (0,1)$, therefore the convergence is Q-linear. □

8.9 Newton's Method

Newton step:

$$x^{k+1} = x^k - \nabla^2 f(x^k)^{-1}\nabla f(x^k) \tag{8.10}$$

The step size α is always 1, this is the classic Newton's method.

8.9.1 Convergence of Newton's method

THEOREM 8.27. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a C^2-class function, and the Hessian matrix $\nabla^2 f(x)$ is Lipschitz-continuous in an open ball $B_\delta(x^*)$ of the optimal point x^*, i.e., there exists constant L such that

$$\forall x, y \in B_\delta(x^*),\ \left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq L\|x-y\|$$

If in addition $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succ 0$. Then we have the following conclusion for the sequence generated by (8.10).

1. If the initial point x_0 is close enough to x^*, then the sequence $\{x^k\}$ is converging to x^*.

2. $\{x^k\}$ converges to x^* Q-quadratically.
3. $\{\|\nabla f(x^k)\|\}$ converges to 0 Q-quadratically.

Proof. By definition of Newton' step and $\nabla f(x^*) = 0$ at the optimal point x^*, we have

$$\begin{aligned} x^{k+1} - x^* &= x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k) - x^* \\ &= \nabla^2 f(x^k)^{-1} [\nabla^2 f(x^k)(x^k - x) - (\nabla f(x^k) - \nabla f(x^*)] \end{aligned} \tag{8.11}$$

By Taylor theorem, we have

$$\nabla f(x^k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^k + t(x^* - x^k))(x^k - x^*) dt$$

Therefore we have:

$$\begin{aligned} \left\|\nabla^2 f(x^k)(x^k - x^*) - (\nabla f(x^k) - \nabla f(x^*))\right\| &= \left\| \int_0^1 [\nabla^2 f(x^k + t(x^* - x^k)) - \nabla^2 f(x^k)](x^k - x^*) dt \right\| \\ &\leq \int_0^1 \left\|\nabla^2 f(x^k + t(x^* - x^k)) - \nabla^2 f(x^k)\right\| \left\|x^k - x^*\right\| dt \\ &\leq \left\|x^k - x^*\right\|^2 \int_0^1 L t dt \\ &= \frac{L}{2} \left\|x^k - x^*\right\|^2 \end{aligned} \tag{8.12}$$

Since $\nabla^2 f(x^*)$ is positive definite and f is C^2 class, there exists $r > 0$ such that when $\|x - x^*\| \leq r$, $\nabla^2 f(x)$ is invertible and

$$\left\|\nabla^2 f(x)^{-1}\right\| \leq 2 \left\|\nabla^2 f(x^*)^{-1}\right\|$$

Combining equation (8.11) and equation (8.11), we have

$$\begin{aligned} \left\|x^{k+1} - x^*\right\| &\leq \left\|\nabla^2 f(x^k)^{-1}\right\| \left\|\nabla^2 f(x^k)(x^k - x^*) - (\nabla f(x^k) - \nabla f(x^*))\right\| \\ &\leq L \left\|\nabla^2 f(x^*)^{-1}\right\| \left\|x^k - x^*\right\|^2 \end{aligned}$$

Let $C = L \left\|\nabla^2 f(x^*)^{-1}\right\|$, then $\left\|x^{k+1} - x^*\right\| \leq C \left\|x^k - x^*\right\|^2$. Therefore the convergence rate is Q-quadratic if the sequence is converging. Hence

$$\left\|x^k - x^*\right\| \leq C^{2^k - 1} \left\|x^0 - x^*\right\|^{2^k} = (C \left\|x^0 - x^*\right\|)^{2^k - 1} \left\|x^0 - x^*\right\|$$

Therefore let $C \left\|x^0 - x^*\right\| \leq \frac{1}{2}$, we can guarantee the convergence of the sequence, which gives

$$\left\|x^0 - x^*\right\| \leq \frac{1}{2L \left\|\nabla^2 f(x^*)^{-1}\right\|}$$

Therefore, when initial point x_0 satisfies

$$\left\|x^0 - x^*\right\| \leq \min\left\{\delta, r, \frac{1}{2L \left\|\nabla^2 f(x^*)^{-1}\right\|}\right\}$$

the sequence $\{x^k\}$ is in the open ball $B_\delta(x^*)$ and is converging to x^* Q-quadratically.

Since $\nabla f(x)$ is continuous, $\nabla f(x^k)$ is converging to 0 as well.

To study the convergence rate of the gradient, let $d^k = -\nabla^2 f(x^k)^{-1}\nabla f(x^k)$, then $x^{k+1} = x^k + d^k$, we have

$$\begin{aligned}
\left\|\nabla f(x^{k+1})\right\| &= \left\|\nabla f(x^{k+1} - \nabla f(x^k) - \nabla^2 f(x^k)d^k\right\| \\
&= \left\|\int_0^1 \nabla^2 f(x^k + td^k)d^k dt - \int_0^1 \nabla^2 f(x^k)d^k dt\right\| \\
&\leq \int_0^1 \left\|\nabla^2 f(x^k + td^k) - \nabla^2 f(x^k)d^k\right\| dt \\
&\leq \int_0^1 \left\|\nabla^2 f(x^k + td^k) - \nabla^2 f(x^k)\right\| \left\|d^k\right\| dt \\
&\leq \int_0^1 Lt \left\|d^k\right\|^2 dt \quad \text{Lipschitz} \\
&\leq \frac{1}{2}\left\|d^k\right\|^2 \leq \frac{1}{2}L\left\|\nabla^2 f(x^k)^{-1}\right\|^2 \left\|\nabla f(x^k)\right\|^2 \\
&\leq 2L\left\|\nabla^2 f(x^*)^{-1}\right\|^2 \left\|\nabla f(x^k)\right\|^2
\end{aligned}$$

Therefore, the convergence is Q-quadratic. □

Remark 8.28. From Theorem 8.27, we can see Newton's method has fast convergence rate, but its convergence is conditional: First, Newton's method only has local convergence instead of global convergence. If the initial point is too far away from the optimal solution, Newton's method can fail to converge. Secondly, The Hessian matrix $\nabla^2 f(x^*)$ needs to be positive definite. If $\nabla^2 f(x^*)$ is singular, it may fail to converge or converge very slow.

Therefore, in practice, people often combine Newton method with other first order methods, for example, the gradient descent method. One can use gradient descent method to obtain a solution which is not very accurate but close to the optimal point, then switch to Newton's method to obtain a very accurate solution.

8.9.2 Quasi-Newton Method

*** Additional reading not covered in this course ***

9 Trust Region Methods

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^2-class function. Consider the second order Taylor expansion of function f.

$$f(x^k + d) = f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T \nabla^2 f(x^k + td)d, \quad t \in (0, 1)$$

Similar to Newton's method, we consider second order approximation of $f(x^k + d)$. Define

$$m_k(d) = f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T B^k d,$$

where B^k is a symmetric matrix. We would like B^k to be an approximation of the Hessian matrix $\nabla^2 f(x^k)$. If B^k happens to be equal to $\nabla^2 f(x^k)$, then the error of the approximation is $O(\|d\|^3)$.

Hence function m_k only approximates $f(x)$ well when $\|d\|$ is small, when $\|d\|$ is big, the approximation may be very bad. Therefore we need to add some constraints to it. Consider the approximation of $f(x^k + d)$ in the following region (closed ball):

$$\Omega_k = \{x^k + d \mid \|d\| \leq \Delta_k\}$$

We call Ω_k the "*trust region*" and Δ_k the "*trust region radius*". It literally means we "trust" function $m_k(d)$ to be a good approximation of $f(x^k + d)$ in the trust region, and Δ_k measures the size of this trust region. Therefore we need to solve the following *trust region subproblem* at every iteration of trust region method.

$$\begin{array}{ll} \min\limits_{d \in \mathbb{R}^n} & m_k(d) \\ \text{s.t.} & \|d\| \leq \Delta_k \end{array} \tag{9.1}$$

In trust region methods, it is important to choose the right trust region region to ensure convergence. If $m_k(d)$ is a good approximation and the optimal d is on the boundary of the trust region, then it is reasonable to increase the trust region radius to make the algorithm more efficient. Else if the error is too big, then we should decrease the trust region radius. Otherwise, the trust region radius stays the same.

We introduce the following parameter ρ_k to measure how "good" the approximation is

$$\rho_k = \frac{f(x^k) - f(x^k + d)}{m_k(0) - m_k(d^k)}$$

The trust region method is described in Algorithm 2:

Algorithm 2: Trust Region Method

Input: the maximum radius $\Delta_{\max}$, initial radius Δ_0, initial point x^0, $k \leftarrow 0$.
Parameters: $0 \leq \eta < \rho_1 < \rho_2 < 1$, $\gamma_1 < 1 < \gamma_2$
In practice, one can choose $\bar{\rho}_1 = \frac{1}{4}, \bar{\rho}_2 = 0.75$ and $\gamma_1 = 0.25, \gamma_2 = 2$.
while *Stopping criteria not satisfied* **do**
  Solve the trust region subproblem to obtain d^k
  Compute the measurement ρ_k
  Updating the trust region radius ρ_k

$$\Delta = \begin{cases} \gamma_1 \Delta_k & \rho_k < \bar{\rho}_1 \\ \min\{\gamma_2 \Delta_k, \Delta_{\max}\} & \rho_k > \bar{\rho}_2 \text{ and } \left\|d^k\right\| = \Delta_k \\ \Delta_k & \text{otherwise} \end{cases}$$

  Updating x^k:

$$x^{k+1} = \begin{cases} x^k + d^k & \rho_k > \eta \\ x^k & \text{otherwise} \end{cases}$$

  $k \leftarrow k + 1$
end

Note we use $\Delta_{\max}$ to control the maximum radius of the trust region, that is because ρ_k only reflects relative error, when $\|d\|$ is big, the error is big, even ρ_k is close to 1.

Also besides $\rho_k > \bar{\rho}_2$, we also require $\|d\|^k = \Delta_k$ in order to increase the radius of the trust region. That is because if the optimal d is in the interior of the trust region, then increasing the trust region radius will yield the same optimal solution for the trust region subproblem.

9.1 Solving the Trust Region Subproblem

The trust region subproblem is a constrained quadratic minimization problem. The following condition is the optimal condition for the trust region subproblem.

THEOREM 9.1. Vector d^* is the global optimal solution for the trust region subproblem

$$\begin{array}{ll} \min\limits_{d \in \mathbb{R}^n} & c + b^T d + \frac{1}{2} d^T B d = m(d) \\ \text{s.t.} & \|d\| \leq \Delta \end{array}$$

if and only if d^* is feasible and there exists $\lambda \geq 0$ such that

$$(B + \lambda I) d^* = -b, \tag{9.2a}$$
$$\lambda(\Delta - \|d^*\|) = 0, \tag{9.2b}$$
$$(B + \lambda I) \succeq 0. \tag{9.2c}$$

Proof. We only prove it is the sufficient condition, we will delay the complete proof of this theorem to the next chapter.

We define function

$$\hat{m}(d) = c + b^T d + \frac{1}{2} d^T (B + \lambda I) d$$

Then $\hat{m}(d)$ is a convex function by equation (9.2c). Also d^* is a stationary point of $\hat{m}(d)$ by equation (9.2a), therefore d^* is a global minimizer of $\hat{m}(d)$. Note $\hat{m}(d) = m(d) + \frac{\lambda}{2} d^T d$. Therefore for any feasible d, we have

$$\begin{aligned} \hat{m}(d) &\geq \hat{m}(d^*) \\ m(d) + \frac{\lambda}{2}\|d\|^2 &\geq m(d^*) + \frac{\lambda}{2}\|d^*\|^2 \\ m(d) &\geq m(d^*) + \frac{\lambda}{2}(\|d^*\|^2 - \|d\|^2) \end{aligned}$$

By equation (9.2b), we have $\lambda(\Delta^2 - \|d^*\|^2) = 0$, and

$$m(d^*) + \frac{\lambda}{2}(\|d^*\|^2 - \|d^*\|^2) = m(d^*) + \frac{\lambda}{2}(\Delta^2 - \|d\|^2)$$

Since $\|d\| \leq \Delta$, we have

$$m(d^*) + \frac{\lambda}{2}(\Delta^2 - \|d\|^2) \geq m(d^*)$$

Therefore $m(d) \geq m(d^*)$ for any feasible d and d^* is a global minimizer.

□

We now describe how to solve the trust region subproblem.

9.1.1 Case a

If $b = 0$, then we already know how to solve the trust region subproblem from Assignment 2. In fact since $b = \nabla f(x^k)^T$, we would have already obtained a stationary point if $b = 0$, so the trust region method can just terminate. Therefore we can assume $b \neq 0$

9.1.2 Case b

We solve the trust region subproblem without the constraints, i.e., let $\hat{d}$ be a global minimizer of $m(d)$. If $\left\|\hat{d}\right\| < \Delta$, then $\hat{d}$ is automatically the optimal solution of the constrained trust region subproblem.

9.1.3 Case c

Assume $m(d)$ has a global minimzer $\left\|\hat{d}\right\| \geq \Delta$, if this happens, then we claim there must exist a global optimal solution of the trust region subproblem on the boundary.

Lemma 9.2. Let $\hat{d}$ be a global minimizer of $m(d)$. If $\left\|\hat{d}\right\| \geq \Delta$, then there exists a global minimizer d^* of the trust region subproblem such that $\|d^*\| = \Delta$.

Proof. Let d^* be a global minimizer of the constrained trust region subproblem (its existence is guaranteed by Weierstrass extreme value theorem). If $\|d^*\| = \Delta$, then we are done. So let's assume $\|d^*\| < \Delta$, then there is a small open ball in the trust region which contains d^*, so $\|d^*\|$ is also a local minimizer of the unconstrained optimization problem $m(d)$. By Second Order Necessary Condition for local optimality, d^* must be a stationary point of $m(d)$ and $\nabla^2 m(d^*) = B$ must be positive semidefinite. Since B is positive semidefinite, $m(d)$ must be a convex function by Proposition 4.9, therefore d^* must be a global minimizer of

$m(d)$. So we now have two global minimizers of $m(d)$: $\hat{d}$ and d^*. We can then draw a line segment connecting $\hat{d}$ and d^*. Any point in this line segment is a convex combination of d^* and $\hat{d}$, which is also a stationary point, and therefore a global minimizer of $m(d)$ by Theorem 5.5. So the intersection of this line segment with the boundary $\|d\| = \Delta$ is a global minimizer of the trust region subproblem. □

In the other case, if $m(d)$ is not bounded below, so there exists no global minimizer of $m(d)$, we can also similarly show that there exists an optimal solution of the trust region subproblem on the boundary, this is left as an exercise.

Before we derive a complete algorithm, we need to define a function $d(\lambda)$ to help analyze the problem.

Let $B = Q\Lambda Q^T$ be the eigenvalue decomposition of B where Q is orthonormal and

$$\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$$

with $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ being the eigenvalues of B.

Let $Q = [q_1, q_2, ..., q_n]$ where q_i is the eigenvector corresponding to λ_i and define function

$$d(\lambda) = -\sum_{j=1}^{n} \frac{q_j^T b}{\lambda_j + \lambda} q_j$$

One can check if $\lambda > -\lambda_1$ (or more general $\lambda \neq \lambda_j$), then $d(\lambda) = -(B + \lambda I)^{-1} b$ and

$$(B + \lambda I) d(\lambda) = -b \tag{9.3}$$

which satisfies condition (9.2a).

The squared norm of $d(\lambda)$ is the following

$$\|d(\lambda)\|^2 = \sum_{j=1}^{n} \frac{(q_j^T b)^2}{(\lambda_j + \lambda)^2}$$

Since $b \neq 0$ by assumption, there must exists some j such that $q_j^T b \neq 0$. Therefore $\|d(\lambda)\|$ is a monotone decreasing function on the interval $[-\lambda_1, \infty)$. (If $\lambda = -\lambda_1$ and $q_j^T b \neq 0$ for some $\lambda_j = \lambda_1$, we can simply define $\|d(-\lambda_1)\| = +\infty$).

Now we can make the following two assumptions:

1. $b \neq 0$. (which implies $\|d(\lambda)\|$ is a monotone decreasing function on the interval $[-\lambda_1, \infty)$)
2. The optimal d^* satisfies $\|d^*\| = \Delta$.

We consider two cases under the above assumptions

- Non-degenerate case (easy case)
    1. If $q_j^T b \neq 0$ for some j such that $\lambda_j = \lambda_1$. Then

    $$\lim_{\lambda \to -\lambda_1^+} \|d(\lambda)\| = +\infty$$

    and

    $$\lim_{\lambda \to +\infty} \|d(\lambda)\| = 0$$

    Therefore there must exists $\lambda^* \in (-\lambda_1, \infty)$ such that $\|d(\lambda^*)\| = \Delta$.

    From equation (9.3), $(\lambda^*, d(\lambda^*))$ satisfies optimal condition (9.2a). Therefore it is an optimal pair.

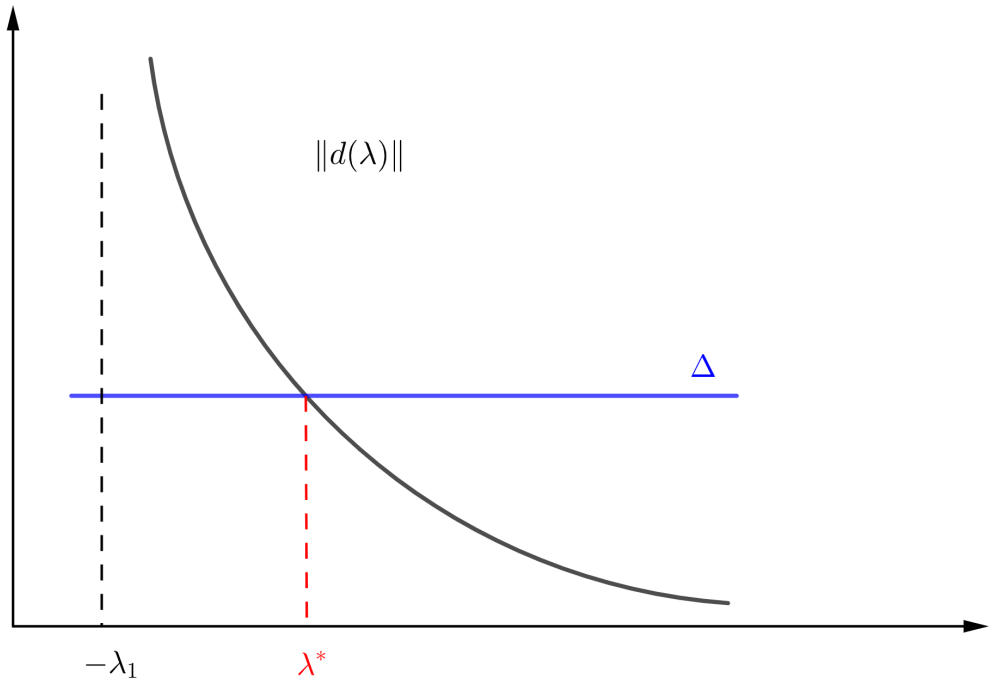


Figure 15: Non-degenerate case: part 1

2. If $q_j^T b = 0$ for all j such that $\lambda_j = \lambda_1$ and $\|-d(\lambda)\| > \Delta$. (In this case $d(\lambda) = \sum_{j:\lambda_j \neq \lambda_1} \frac{q_j^T b}{\lambda_j + \lambda} q_j$ so $d(\lambda)$ is well defined at $-\lambda_1$)

   Again $\lim_{\lambda \to +\infty} \|d(\lambda)\| = 0$ and there must exist $\lambda^* \in (-\lambda_1, \infty)$ such that $\|d(\lambda^*)\| = \Delta$ and it is easy to check the pair $(\lambda^*, d(\lambda^*))$ satisfies all the three optimal conditions.

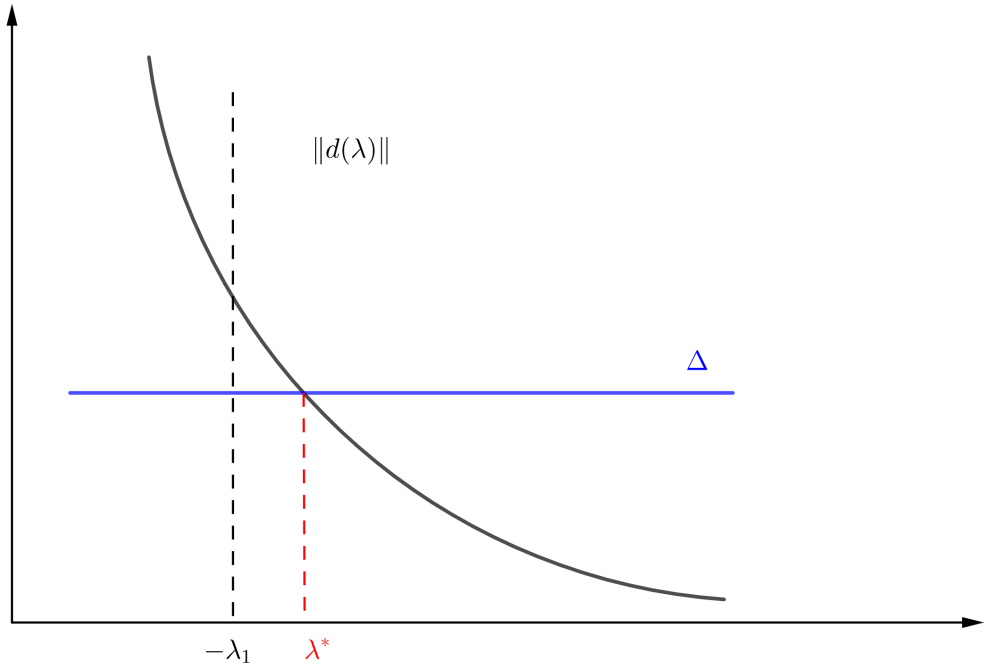


Figure 16: Non-degenerate case: part 2

- Denegerate case (hard case)
  - If scenarios in the non-degenerate case do not happen, then we have $q_j^T b = 0$ for all j such that $\lambda_j = \lambda_1$ and $\|d(-\lambda_1)\| \leq \Delta$.

    In this case, we claim the optimal multiplier $\lambda^* = -\lambda_1$, i.e., there exists some d^* such that $(-\lambda_1, d^*)$ is an optimal pair.

    Suppose the claim is not true, then $\lambda^* > -\lambda_1$ (from the optimal conditions, the optimal λ^* must be in the interval $[-\lambda_1, +\infty)$) which implies $B + \lambda^* I$ must be positive definite. From the first optimal condition $(B + \lambda^* I)d^* = -b$, we have

$$d^* = -(B + \lambda^* I)^{-1} b = -\sum_{j=1}^{n} \frac{q_j^T b}{\lambda_j + \lambda^*} q_j = d(\lambda^*)$$

    So d^* is exactly $d(\lambda^*)$! Since d^* is optimal, by assumption 2 we have $\|d^*\| = \Delta = \|d(\lambda^*)\|$. But we already know $\|d(-\lambda_1)\| \leq \Delta$. So this is a contradiction to the monotone decreasing property of $\|d(\lambda)\|$ from assumption 1. Hence $\lambda^* > -\lambda_1$ can not happen, and we have $\lambda^* = -\lambda_1$. The optimal d^* is then given by

$$d^* = -\sum_{j:\lambda_j \neq \lambda_1} \frac{q_j^T b}{\lambda_j - \lambda_1} q_j + \tau q_1 \tag{9.4}$$

    where τ is computed from

$$\Delta^2 = \sum_{j:\lambda_j \neq \lambda_1} \frac{(q_j^T b)^2}{(\lambda_j - \lambda_1)^2} + \tau^2$$

    It is easy to check $-\lambda_1$ and d^* in equation (9.4) satisfies optimal conditions (9.2c) and (9.2b). It is left as an exercise to check d^* does indeed satisfy condition (9.2a).

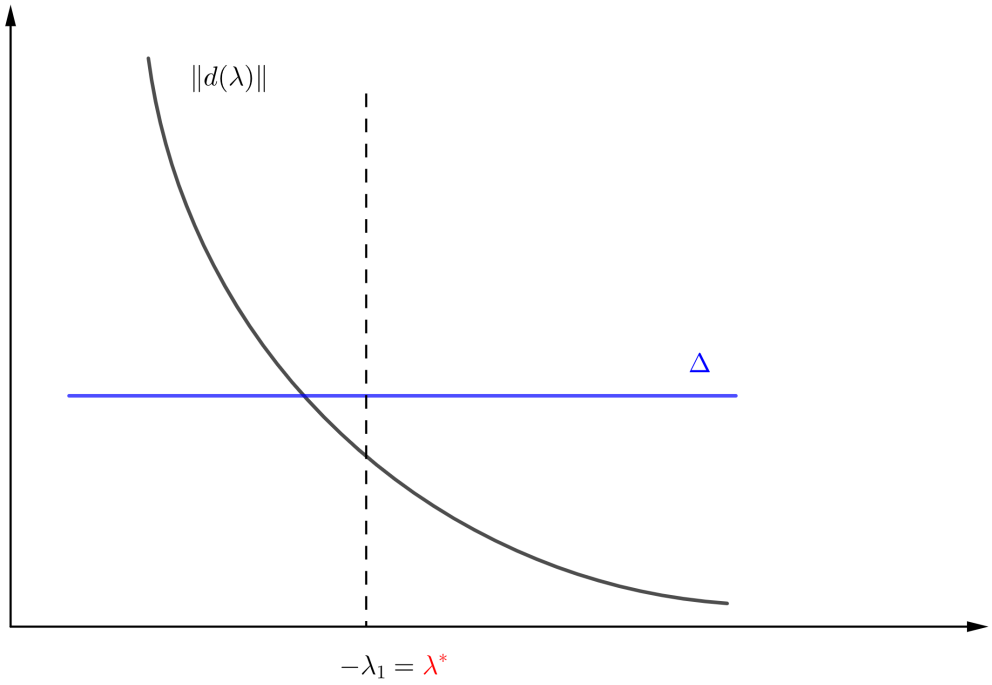


Figure 17: Degenerate case

Remark 9.3. Note in the above discussion, we didn't explicitly show $\lambda^* \geq 0$ as required by Theorem 9.1. This may cause trouble if $B \succ 0$ ($-\lambda_1 < 0$).

In fact, if $B \succ 0$, then we can find the unique optimal solution of the unconstrained optimization problem which is $\hat{d} = -B^{-1}b$. Since we assume there exists an optimal solution of the trust region subproblem on the boundary, it must be the case that $\left\|\hat{d}\right\| \geq \Delta$. Observe that $d(0) = -B^{-1}b$, so $\|d(0)\| \geq \Delta$. Therefore there must exist $\lambda^* \geq 0$ such that $\|d(\lambda^*)\| = \Delta$.

To summarize, if B is positive definite, then we must have $\|d(0)\| \geq \Delta$, therefore the degenerate case can not happen, and we are guaranteed to find $\lambda^* \geq 0$.

9.1.4 Lower and Upper Bound in the Nondegenerate Case

In the nondegenerate case, $\|d(\lambda^*)\| = \Delta$ leads to a lower and upper bound for the multiplier λ^*. Since $\lambda_i + \lambda^* \geq \lambda_1 + \lambda^*,\ 1 \leq i \leq n$, we have

$$\Delta^2 = \sum_{j=1}^{n} \frac{(q_j^T b)^2}{(\lambda_j + \lambda^*)^2} \leq \sum_{j=1}^{n} \frac{(q_j^T b)^2}{(\lambda_1 + \lambda^*)^2} = \frac{\|b\|^2}{(\lambda_1 + \lambda^*)^2}.$$

Since $\lambda_1 + \lambda^*$, it follows that

$$\lambda^* \leq \frac{\|b\|}{\Delta} - \lambda_1 := \lambda_u$$

To obtain a lower bound, observe that

$$\Delta^2 = \sum_{j=1}^{n} \frac{(q_j^T b)^2}{(\lambda_j + \lambda^*)^2} \geq \frac{1}{(\lambda_1 + \lambda^*)^2} \sum_{j:\lambda_j = \lambda_1} (q_j^T b)^2,$$

which yields the relation

$$\lambda^* \geq -\lambda_1 + \frac{1}{\Delta} \left(\sum_{j:\lambda_j = \lambda_1} (q_j^T b)^2 \right)^{\frac{1}{2}} := \lambda_l$$

Hence $\lambda^* \in [\max\{0, \lambda_l\}, \lambda_u]$. We can then use bisection method or root-finding Newton method to solve for λ^*

9.2 A Complete Algorithm

We now have a complete algorithm for solving the trust region subproblem.

Algorithm 3: Trust Region Subproblem

Input: a symmetric matrix B approximating the Hessian and the gradient vector b, trust region radius Δ.

Output: optimal solution d^*

if *B is positive definite, and* $\|d\| = \left\|-B^{-1}b\right\| < \Delta$ **then**

  return $d^* := -B^{-1}b$

else

  Compute the minimum eigenvalue λ_1 of B and corresponding eigenvectors q_j.

  if $q_j^T b = 0$ *for all* $j : \lambda_j = \lambda_1$ *and* $\left\|(B - \lambda_1 I)^\dagger)b\right\| \leq \Delta$ **then**

    $$d^* := -(B - \lambda_1 I)^\dagger b + \tau q_1$$

    where τ is computed from

    $$\tau^2 = \Delta^2 - \left\|(B - \lambda_1 I)^\dagger b\right\|^2$$

    return d^*

  else

    Use Newton root-finding method or bisection method to find $\lambda^* \in [\max\{0, \lambda_l\}, \lambda_u]$ such that

    $$\left\|(B + \lambda^* I)^{-1})b\right\| = \Delta$$

    Compute

    $$d^* := -(B + \lambda^* I)^{-1} b$$

    return d^*;

  end

end

9.2.1 Implementation

For a practical and efficient way to implement the Newton root-finding method, please see Algorithm 4.3 on Page 87 in Numerical Optimization (by Jorge Nocedal and Stephen Wright).

In Algorithm 3, $(B - \lambda_1 I)^\dagger)$ is the (Moore-Penrose) pseudo-inverse of $(B - \lambda_1 I)$. For the definition and how to compute the pseudo-inverse, see the link Pseudo-inverse Python or Pseudo-inverse MATLAB

9.3 Convergence Analysis

Define the sublevel set S as

$$S = \{x \mid f(x) \leq f(x_0)\}$$

and define a neighborhood of this set by

$$S(R_0) = \{x \mid \|x - y\| < R_0 \text{ for some } y \in S\}$$

where R_0 is a positive constant.

To guarantee the convergence of trust region method, we don't need to solve the trust region subproblem exactly, specially, we require that

$$m(0) - m(d) \geq c_1(m(0) - m(d^*)) \tag{9.5a}$$

$$\|d\| \leq \gamma\Delta \tag{9.5b}$$

where d^* is an exact solution of the trust region subproblem, d is an approximate solution of the trust region subproblem, $c_1 \in (0, 1]$ and $\gamma > 0$

THEOREM 9.4. Let $\|B_k\| \leq \beta$ for sine constant β, f is bounded below on the level set S , ∇f is Lipschitz continous in $S(R_0)$ for some $R_0 > 0$ and f is C^2-class function in the sublevel set S. Suppose that $B^k = \nabla^2 f(x_k)$ for all k, and that the approximate solution of the trust region subproblem d^k satisfies (9.5) for some fixed $\gamma > 0$. Then

$$\lim_{k\to\infty} \|g_k\| = 0.$$

If in addition, the sublevel set S is compact, then the sequence $\{x^k\}$ generated by the trust region method has a limit point x^* in S at which the second order necessary conditions for local optimal solution hold.

Proof. We omit the proof, for a detaled proof, see More and Sorensen, Computing a trust region step, SIAM Journal on Scientific and Statistical Computing (1983). □

10 Theory of Constrained Optimization

A general formulation for constrained optimization problem is

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & c_i(x) = 0, i \in \mathcal{E} \\ & c_i(x) \leq 0, i \in \mathcal{I}. \end{array} \tag{10.1}$$

DEFINITION 10.1. (Active Set)

The active set at any feasible x is defined as follows

$$\mathcal{A}(x) = \mathcal{E} \cup \{i \in \mathcal{I} \mid c_i(x) = 0\}$$

10.1 Examples

Consider the following two-variable problem

$$\begin{array}{ll} \min & x + y \\ \text{s.t.} & x^2 + y^2 - 2 = 0 \end{array} \tag{10.2}$$

The optimal solution $x^* = (-1, -1)^T$.

At the solution x^*, there is no direction to move that stays feasible while decreasing the objective function f, which gives the following condition

$$\nabla f(x^*) = -\lambda_1^* \nabla c_1(x^*)$$

for some nonnegative λ_1^*.

By introducing the Lagrangian function

$$\mathcal{L}(x, \lambda_1) = f(x) + \lambda_1 c_1(x)$$

Noting that

$$\nabla_x \mathcal{L}(x^*, \lambda_1^*) = \nabla f(x^*) + \lambda_1^* \nabla c_1(x^*)$$

The above condition is equivalent to

$$\nabla_x \mathcal{L}(x^*, \lambda_1^*) = 0$$

This is called *stationary condition* of the Lagrangian function for constrained optimization problem. The scalar quantity λ_1 is called a *Lagrange multiplier* for the constraint $c_1(x) = 0$

The stationary condition is necessary (under some assumption which we will see later) but not sufficient. It is derived from the fact that at the local optimal solution x^*, there is no direction to move that stays feasible while decreasing the objective function f,

Consider the following problem with one inequality

$$\begin{array}{ll} \min & f(x, y) \\ \text{s.t.} & x^2 + y^2 - 2 \leq 0 \end{array} \tag{10.3}$$

Case 1: $f(x, y) = x^2 + y^2$. The Optimal solution $x^* = (0, 0)^T$is in the interior. which gives necessary condition

$$\nabla f(x) = 0$$

Case 2: $f(x,y) = x + y$. The optimal solution $x^* = (-1,-1)^T$ is on the boundary which gives necessary condition

$$\nabla f(x) + \lambda_1 \nabla c_1(x) = 0, \text{for some } \lambda_1 \geq 0$$

We can check in both case I and II, there exists some $\lambda_1^* \geq 0$ such that

$$\nabla_x \mathcal{L}(x^*, \lambda_1^*) = \nabla f(x^*) + \lambda_1^* \nabla c_1(x^*) = 0$$

and

$$\lambda_1^* c_1(x^*) = 0$$

In case I, λ_1^* is just 0. In case II, we have $\lambda_1^* = \frac{1}{2}$.

Condition $\lambda_1^* c_1(x^*) = 0$ is known as *complementarity condition* it implies

- (a) The Lagrange multiplier λ_1 can be zero when the corresponding inequality constraint c_1 is inactive.
- (b) If the inequality constraint c_1 is active, the corresponding Lagrange multiplier can be nonnegative.
- (c) If c_i is an equality constraint, the corresponding Lagrange multiplier can be negative, positive or zero.

10.2 First Order Necessary Conditions (KKT) for Constrained Optimization

DEFINITION 10.2. (Lagrangian Function)

The *Lagrangian function* of the general constraint optimization problem 10.1 is defined as following

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i \in \mathcal{E} \cap \mathcal{I}} \lambda_i c_i(x)$$

DEFINITION 10.3. (Karush-Kuhn-Tucker Conditions (KKT Condition))

Given a point x and Lagrangian multiplier vector λ, the KKT condition for x and λ is the following:

$$\nabla_x \mathcal{L}(x, \lambda) = 0. \qquad \text{(stationary)} \tag{10.4a}$$
$$\lambda_i c_i(x) = 0, \quad \text{for all } i \in \mathcal{E} \cup \mathcal{I} \qquad \text{(complementarity)} \tag{10.4b}$$
$$\lambda_i \geq 0, \quad \text{for all } i \in \mathcal{I} \tag{10.4c}$$
$$c_i(x) = 0, \quad \text{for all } i \in \mathcal{E} \tag{10.4d}$$
$$c_i(x) \leq 0, \quad \text{for all } i \in \mathcal{I} \tag{10.4e}$$

Consider the following two examples

$$\begin{aligned} \min \quad & x \\ \text{s.t.} \quad & -x + 3 \leq 0 \end{aligned} \tag{10.5}$$

$$\begin{aligned} \min \quad & x \\ \text{s.t.} \quad & (-x+3)^3 \leq 0 \end{aligned} \tag{10.6}$$

It is obvious that these two problems have the same feasible region and objective function, therefore their optimal solution should be the same, i.e., the optimal solution is $x^* = 3$.

However, one can check the optimal solution x^* satisfies KKT conditions in Problem (10.5) but fails to satisfy KKT condition in Problem (10.6).

The reason is that in Problem (10.5), the normal vector of the feasible region at the optimal point x^* is $d = -1$ (or a positive multiple of d) and the gradient of the constraint at x^* is also -1. However, in Problem (10.6), the gradient of the constraint at x^* is 0. The gradient does not capture the essential geometric features at the optimal point

10.3 Tangent Cone and Constraint Qualifications

To address the above issue, we introduce the notion of tangent cone and linear feasible direction set.

DEFINITION 10.4. (Tangent Vector)

The vector d is said to be a tangent vector to x if there are feasible sequence $\{z_k\}$ approaching x and a sequence of positive scalars $\{t_k\}$ with $\{t_k\} \to 0$ such that

$$\lim_{k\to\infty} \frac{z_k - x}{t_k} = d$$

The set of all such d is called the tangent cone and is denoted as $T_\Omega(x)$

DEFINITION 10.5. (Linearized Feasible Direction)

Given a feasible x and the active constraint set $\mathcal{A}(x)$, the set of linearized feasible direction $\mathcal{F}(x)$is

$$\mathcal{F}(x) = \left\{ d \;\middle|\; \begin{array}{ll} d^T \nabla c_i(x) = 0 & \text{for all } i \in \mathcal{E} \\ d^T \nabla c_i(x) \le 0 & \text{for all } i \in \mathcal{A}(x) \cap \mathcal{I} \end{array} \right\}$$

THEOREM 10.6. (Geometric Necessary Optimal Condition)

Let x^* be a local optimal solution of Problem 10.1. If $f(x)$ and $c_i(x), i \in \mathcal{E} \cap \mathcal{I}$ are differentiable at x^*. Then

$$d^T \nabla f(x^*) \ge 0, \ \forall d \in T_\Omega(x^*)$$

Proof. We prove it by contradiction.

Suppose there exists d such that $d \in T_\Omega(x^*)$ and $d^T \nabla f(x^*) < 0$. Then there exists $\{t_k\}_{k=1}^\infty$ and $\{d_k\}_{k=1}^\infty$ such that $x^* + t_k d_k \in \Omega$ where $t_k \to 0$ and $d_k \to d$.

Since $\nabla f(x^*)^T d < 0$, for sufficiently large k we have

$$\begin{aligned} f(x^* + t_k d_k) &= f(x^*) + t_k \nabla f(x^*)^T d_k + o(t_k) \\ f(x^* + t_k d_k) &= f(x^*) + t_k \nabla f(x^*)^T d + t_k \nabla f(x^*)^T (d_k - d) + o(t_k) \\ &= f(x^*) + t_k \nabla f(x^*)^T d + o(t_k) \\ &< f(x^*). \end{aligned}$$

This is a contradiction to the local optimality of x^*. □

10.3.1 Constraint qualifications

Constraint qualifications are conditions that make $T_\Omega(x) = \mathcal{F}(x)$.

DEFINITION 10.7. (Linear Independence Constraint Qualification (LICQ))

Linear independence constraint qualification holds if the set of active constraint gradient

$$\{\nabla c_i(x),\ i \in \mathcal{A}(x)\}$$

is linearly independent

Lemma 10.8. Let x^* be a feasible point, then the following two statements are true:

1. $T_\Omega(x) \subset \mathcal{F}(x^*)$
2. If LICQ holds, then they are equal.

Proof. 1. Let $d \in T_\omega(x)$, then by the definition of Tangent cone, we have

$$\lim_{k\to\infty} \frac{z_k - x}{t_k} \to d, \lim_{k\to\infty} t_k = 0, z_k \in \Omega$$

which is equivalent to

$$z_k = x + t_k d + o(t_k)$$

For $i \in \mathcal{E}$, we have $c_i(x) = 0$, therefore by taylor expansion of $c_i(z_k)$ at x, we have

$$\begin{aligned} 0 &= c_i(z_k) \\ &= c_i(x) + \nabla c_i(x)^T(t_k d + o(t_k)) + o(t_k d + o(t_k)) \\ &= \nabla c_i(x)^T t_k d + o(t_k) \end{aligned}$$

Divide by t_k, we have

$$\nabla c_i(x)^T d + o(1) = 0$$

Let $k \to \infty$, then $o(1) \to 0$, therefore $\nabla c_i(x)^T d = 0$ for $i \in \mathcal{E}$.

For $i \in \mathcal{A}(x) \cap \mathcal{I}$, we have

$$\begin{aligned} 0 &\geq c_i(z_k) \\ &= c_i(x) + \nabla c_i(x)^T(t_k d + o(t_k)) + o(t_k d + o(t_k)) \\ &= \nabla c_i(x)^T t_k d + o(t_k) \end{aligned}$$

Divide by t_k, we have

$$\nabla c_i(x)^T d + o(1) \leq 0$$

Let $k \to \infty$, then $o(1) \to 0$, therefore $\nabla c_i(x)^T d \leq 0$ for $i \in \mathcal{A}(x) \cap \mathcal{I}$.

2. To prove the second item, we need the *implicit function theorem* (Theorem 10.9) from multivariate calculus.

   First we construct matrix of the gradient of the active constraints as

$$A(x^*) = [\nabla c_i(x^*)]_{i\in\mathcal{A}(x^*)}$$

Since LICQ holds, we have that the matrix $A(x^*)$ has full row rank. Let Y be a matrix whose columns form a basis for the null space of $A(x^*)$, then we have $A(x^*)Y = 0$ and $\begin{bmatrix} A(x^*) \\ Y^T \end{bmatrix}$ is a nonsingular square matrix.

Let $d \in \mathcal{F}(x^*)$, and let $\{t_k\}_{k=0}^{\infty}$ be a sequence such that $\lim_{k\to\infty} t_k = 0$. Define a system of equations $R : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ by

$$R(z,t) = \begin{bmatrix} c(z) - tA(x^*)d \\ \\ Y^T(z - x^* - td) \end{bmatrix} = \begin{bmatrix} 0 \\ \\ 0 \end{bmatrix}.$$

At $t = 0$ and $z = x^*$, the Jacobian matrix of R at this point is

$$\nabla_z R(x^*, 0) = \begin{bmatrix} A(x^*) \\ Y^T \end{bmatrix} \tag{10.7}$$

which is full rank by the construction of Y. Hence, according to the implicit function theorem, there exists a continuously differentiable function g such that $z_k = g(t_k)$ is a solution of equation (10.7) for all values t_k sufficiently small (k sufficiently large), therefore by the definition of $R(z,t)$ we have

$$\begin{array}{rcll} c_i(z_k) & = & t_k \nabla c_i(x^*)^T d & \text{for all} \quad i \in \mathcal{E} \\ c_i(z_k) & = & t_k \nabla c_i(x^*)^T d & \text{for all} \quad i \in \mathcal{A}(x^*) \cap \mathcal{I} \end{array}$$

By the definition of $\mathcal{F}(x^*)$, we have $t_k \nabla c_i(x^*)^T d = 0$ for all $i \in \mathcal{E}$ and $t_k \nabla c_i(x^*)^T d \leq 0$ for all $i \in \mathcal{A}(x^*) \cap \mathcal{I}$. Therefore $c_i(z_k) = 0$ for all $i \in \mathcal{E}$ and $c_i(z_k) \leq 0$ for all $i \in \mathcal{A}(x^*) \cap \mathcal{I}$.

Also we know $c_i(x^*) < 0$ for all $i \in \mathcal{I} \backslash \mathcal{A}(x^*)$. Since the implicit function $z_k = g(t_k)$ is continuously differentiable, for t_k small enough, $||z_k - x^*||$ will be small enough such that $c_i(z^k) < 0$ for all $i \in \mathcal{I} \backslash \mathcal{A}(x^*)$ as well.

So we have proved that for k large enough, the sequence $\{z_k\}$ is a sequence of feasible points. The rest is to prove $d = \lim_{k\to\infty} \dfrac{z_k - x^*}{t_k}$.

Using the fact that $R(z_k, t_k) = 0$ for all k together with Taylor's theorem, we have

$$\begin{array}{rcl} R(z_k, t_k) & = & \begin{bmatrix} c(z_k) - t_k A(x^*)d \\ \\ Y^T(z^k - x^* - t_k d) \end{bmatrix} \\ & = & \begin{bmatrix} A(x^*)(z_k - x^*) + o(||z_k - x^*||) - t_k A(x^*)d \\ \\ Y^T(z_k - x^* - t_k d) \end{bmatrix} \\ & = & \begin{bmatrix} A(x^*) \\ \\ Y^T \end{bmatrix} (z_k - x^* - t_k d) + o(||z_k - x^*||) \\ & = & 0 \end{array}$$

By Multiplying this expression by $\begin{bmatrix} A(x^*) \\ Y^T \end{bmatrix}^{-1}$ and dividing by t_k, we obtain

$$\frac{z_k - x^*}{t_k} = d + \frac{o(||z_k - x^*||)}{t_k}$$

from which it follows that $\lim_{k\to\infty} \frac{z_k - x^*}{t_k} = d$. (why the limit can not be infinity or zero?) Hence we have $d \in T_\Omega(x^*)$ for any $d \in \mathcal{F}(x^*)$, the proof is complete.

□

Note in example (10.6), the tangent cone $T_{\Omega(x^*)}$ at $x^* = 3$ is $\{d : d \geq 0\}$, however $\mathcal{F}(x^*) = \{d \mid d \in \mathbb{R}\}$ at $x^* = 3$, so $T_\Omega(x^*)$ is strictly contained in $\mathcal{F}(x^*)$

THEOREM 10.9. (Implicit Function Theorem)

Let $f : \mathbb{R}^{n+m} \to \mathbb{R}^m$ be a continuously differentiable (C^1) function, and let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ be the unknowns of f. Let $a \in \mathbb{R}^n, b \in \mathbb{R}^m$ be a point such that $f(a, b) = \vec{0}$ where $\vec{0}$ is a vector of zeros. If the Jacobian matrix of f in terms of y is a full rank matrix at (a, b), i.e., if

$$\frac{\partial f}{\partial y}(a,b) = \begin{bmatrix} \frac{\partial f_1}{\partial y_1}(a,b) & \dots & \frac{\partial f_1}{\partial y_m}(a,b) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial y_1}(a,b) & \dots & \frac{\partial f_m}{\partial y_m}(a,b) \end{bmatrix}$$

is invertible.Then there exists an open set $U \in \mathbb{R}^n$ containing a and on open set $V \in \mathbb{R}^m$ containing b such that for each $x_k \in U$, there exists a unique $y_k \in V$ with $f(x_k, y_k) = \vec{0}$. The (implicit) function $g : \mathbb{R}^n \to \mathbb{R}^m$ defined by $g(x_k) = y_k$ for all $x_k \in U$ is also a continuously differentiable function over U.

Remark 10.10. Note in example (10.6), the tangent cone $T_\Omega(x^*)$ at $x^* = 3$ is $\{d : d \geq 0\}$, however $\mathcal{F}(x^*) = \{d \mid d \in \mathbb{R}\}$ at $x^* = 3$, so $T_\Omega(x^*)$ is strictly contained in $\mathcal{F}(x^*)$

Lemma 10.11. (Farkas Lemma)

Let $a_i, i = 1, ..., p$, $b_i, i = 1, ..., q$ be two group of vectors in $\mathbb{R}^n$ and $c \in \mathbb{R}^n$. Then the following statement is true:

There does NOT exist any d which satisfies the following conditions

$$d^T a_i = 0, i = 1, ..., p \tag{10.8a}$$
$$d^T b_i \geq 0, i = 1, ..., q \tag{10.8b}$$
$$d^T c < 0 \tag{10.8c}$$

if and only if there exists λ_i, $i = 1, ..., p$ and $\mu_i \geq 0$, $i = 1, ..., q$ such that

$$c = \sum_{i=1}^{p} \lambda_i a_i + \sum_{i=1}^{q} \mu_i b_i.$$

Proof. If there exists λ_i and $\mu_i \geq 0$ such that

$$c = \sum_{i=1}^{p} \lambda_i a_i + \sum_{i=1}^{q} \mu_i b_i.$$

Then for any d satisfying (10.8a) and (10.8b) we have

$$d^T c = \sum_{i=1}^{p} \lambda_i d^T a_i + \sum_{i=1}^{1} \mu_i d^T b_i \geq 0$$

Therefore the solution set of equations (10.8) is empty.

For the other direction, assume there does NOT exist λ_i and $\mu_i \geq 0$ such that

$$c = \sum_{i=1}^{p} \lambda_i a_i + \sum_{i=1}^{q} \mu_i b_i.$$

Define

$$S = \{z \mid z = \sum_{i=1}^{p} \lambda_i a_i + \sum_{j=1}^{q} \mu_i b_i,\ \lambda_i \in \mathbb{R},\ \mu_i \geq 0\}$$

then we have $c \notin S$ by assumption. It is easy to see that S is a closed convex set. Therefore by Hyperplane Seperation Theorem, there exists a hyperplane $d^T x = \alpha$ that separates c from S, i.e.,

$$\forall z \in S, \quad d^T c < \alpha < d^T z$$

Since $0 \in S$, we have

$$\alpha < d^T 0 = 0,$$

which means $d^T c \leq 0$. On the other hand, for any b_i, $i = 1, ..., q$ we have

$$\forall t \geq 0, \quad t b_i \in S.$$

Therefore

$$\forall t > 0, \quad t d^T b_i > \alpha$$

Let $t \to +\infty$, we have

$$d^T b_i \geq 0.$$

Similarly, for any $a_i, i = 1, ..., p$ we have

$$\forall t \in \mathbb{R}, \quad t a_i \in S,$$

therefore

$$\forall t \in \mathbb{R}, \quad t d^T a_i > \alpha$$

Let $t \to +\infty$ and $t \to -\infty$, we have

$$d^T a_i = 0.$$

Hence d is a solution of equations (10.8). □

THEOREM 10.12. (KKT Conditions)

Suppose x^* is a local minimizer of the constrained optimization problem (10.1) and $f(x)$, $c_i(x), i \in \mathcal{E} \cup \mathcal{I}$ are differentiable at x^*. If

$$T_\Omega(x^*) = \mathcal{F}(x^*).$$

Then there exists Lagrangian multipliers $\lambda_i^*, i \in \mathcal{E} \cup \mathcal{I}$ such that (x^*, λ_i^*) satisfy the KKT conditions (10.4).

Proof. By the geometric necessary condition, and the fact $T_\Omega(x^*) = \mathcal{F}(x^*)$, we have

$$d^T \nabla f(x^*) \geq 0, \quad \forall d \in \mathcal{F}(x^*)$$

which is equivalent to saying that the following set

$$\left\{ d \;\middle|\; \begin{array}{lcl} d^T \nabla f(x^*) & < & 0 \\ d^T \nabla c_i(x^*) & = & 0,\ i \in \mathcal{E} \\ d^T \nabla c_i(x^*) & \leq & 0,\ i \in \mathcal{A}(x^*) \cap \mathcal{I} \end{array} \right\}$$

is empty. Therefore by Farkas Lemma the following condition holds

$$\nabla f(x^*) + \sum_{i \in \mathcal{E}} \lambda_i^* \nabla c_i(x^*) + \sum_{i \in \mathcal{A}(x^*) \cap \mathcal{I}} \lambda_i^* \nabla c_i(x^*) = 0$$

for some $\lambda_i^* \in \mathbb{R}, i \in \mathcal{E}$ and $\lambda_i^* \geq 0, i \in \mathcal{A}(x^*) \cap \mathcal{I}$.

If in addition we define $\lambda_i^* = 0, i \in \mathcal{I} \backslash \mathcal{A}(x^*)$, then

$$\nabla f(x^*) + \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i^* \nabla c_i(x^*) = 0,$$

which is just the stationary condition in the KKT conidtions. Also for any $i \in \mathcal{I}$, we have

$$\lambda_i^* c_i(x^*) = 0$$

Therefore the complementarity condition holds. Hence the KKT conditions hold. □

10.4 Second Order Optimal Conditions for Constrained Optimization

DEFINITION 10.13. (Critical Cone)

Given the linearized feasible set $\mathcal{F}(x^*)$ and Lagrange multiplier vector λ^* satisfying the KKT conditions. We define the *critical cone* $\mathcal{C}(x^*, \lambda^*)$ as follows:

$$\mathcal{C}(x^*, \lambda^*) = \{w \in \mathcal{F}(x^*) \mid \nabla c_i(x^*)^T w = 0, \text{ for all } i \in \mathcal{A}(x^*) \cap \mathcal{I} \text{ with } \lambda_i > 0\}$$

Equivalently,

$$w \in \mathcal{C}(x^*, \lambda^*) \iff \begin{cases} \nabla c_i(x^*)^T w = 0, & \text{for all } i \in \mathcal{E} \\ \nabla c_i(x^*)^T w = 0, & \text{for all } i \in \mathcal{A}(x^*) \cap \mathcal{I} \text{ with } \lambda_i^* > 0 \\ \nabla c_i(x^*)^T w \leq 0, & \text{for all } i \in \mathcal{A}(x^*) \cap \mathcal{I} \text{ with } \lambda_i^* = 0 \end{cases}$$

THEOREM 10.14. (Second Order Necessary Condition)

Let x^* be a local minimizer of problem (10.1), and $T_\Omega(x^*) = \mathcal{F}(x^*)$. Let λ^* be the corresponding Lagrangian multiplier vector, and x^*, λ^* satisfies the KKT conditions. Then

$$\forall d \in \mathcal{C}(x^*, \lambda^*), \quad d^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) d \geq 0,$$

Proof. The proof is omitted. □

THEOREM 10.15. (Second Order Sufficient Condition)

Let x^* be a feasible point of problem (10.1). If there exists Lagrangian multipliers λ^* such that x^*, λ^* satisfies the KKT conditions, and

$$\forall d \in \mathcal{C}(x^*, \lambda^*), d \neq 0, \quad d^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) d > 0.$$

Then x^* is a strict local minimizer of problem (10.1). If the critical cone contains only the zero vector, then x^* is also a strict local minimizer.

Proof. We only prove the case where the critical cone only contains zero vector. The rest of the proof is omitted which can be found in Theorem 12.6 in the book Numerical Optimization by Nocedal and Wright.

We prove the result by showing that every feasible sequence z^k approaching x^* has $f(z_k) > f(x^*)$ for all k sufficiently large. (This would imply x^* is a strict local minimizer, otherwise you can find smaller and smaller neighborhoods $N_t(x^*) \cap \Omega$ of x^* such that x^* is not a strict minimizer, so you can find a y^t in $N_t(x^*) \cap \Omega$ such that $f(y^t) \leq f(x^*)$. Therefore $\{y^t\}$ is a feasible sequence approaching x^* but $f(y^t) \leq f(x^*)$, a contradition)

Suppose this is not the case, then there exits a feasible sequence $\{z^k\}$ approaching x^* with

$$f(z_k) \leq f(x^*), \quad \text{for all } k. \tag{10.9}$$

By taking a subsequence if necessary, we can assume there is a limiting direction d such that

$$\lim_{k \to \infty} \frac{z_k - x^*}{\|z_k - x^*\|} = d, \quad \text{(Bolzano-Weierstrass Theorem)} \tag{10.10}$$

Clearly $d \in T_\Omega(x^*)$ and $\|d\| = 1$ so $d \neq 0$. Since the tangent cone is always contained in the linearized feasibe set, so $d \in \mathcal{F}(x^*)$. But the critical cone $\mathcal{C}(x^*, \lambda^*)$ only contains the zero vector by the assumption, we know $d \neq \mathcal{C}(x^*, \lambda^*)$. So the critical cone is strictly contained in the linearized feasible set, therefore we can identify some index $j \in \mathcal{A}(x^*) \cap \mathcal{I}$ such that the strict positivity condition

$$\lambda_j^* \nabla c_j(x^*)^T d < 0$$

is satisfied. (If $\mathcal{A}(x^*) \cap \mathcal{I} = \emptyset$, then we would have $\mathcal{C}(x^*, \lambda^*) = \mathcal{F}(x^*)$). For the remaining index $i \in \mathcal{A}(x^*)$, we have

$$\lambda_i^* \nabla c_i(x^*)^T d \leq 0.$$

From Taylor's theorem and $z_k - x^* = d\,\|z_k - x^*\| + o(\|z_k - x^*\|)$ by (10.10), we have for this particular index j that

$$\begin{aligned} \lambda_j^* c_j(z_k) &= \lambda_j^* c_j(x^*) + \lambda_j^* \nabla c_j(x^*)^T (z_k - x^*) + o(\|(\| z_k - x^*)) \\ &= \|z_k - x^*\|\, \lambda_j^* \nabla c_j(x^*)^T d + o(\|z_k - x^*\|). \end{aligned}$$

Therefore

$$\begin{aligned} \mathcal{L}(z_k, \lambda^*) &= f(z_k) + \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* c_i(z_k) \\ &\leq f(z_k) + \lambda_j^* c_j(z_k) \\ &\leq f(z_k) + \|z_k - x^*\|\, \lambda_j^* \nabla c_j(x^*)^T d + o(\|z_k - x^*\|). \end{aligned} \tag{10.11}$$

By Taylor's theorem, we have

$$\begin{aligned} \mathcal{L}(z_k, \lambda^*) &= \mathcal{L}(x^*, \lambda^*) + (z_k - x^*)^T \nabla_x \mathcal{L}(x^*, \lambda^*) + o(\|z_k - x^*\|) \\ &= f(x^*) + o(\|z_k - x^*\|) \quad \text{(stationary and complementarity conditions)} \end{aligned} \tag{10.12}$$

Combining (10.11) with (10.12), we obtain

$$f(z_k) \geq f(x^*) - \|z_k - x^*\| \lambda_j^* \nabla c_j(x^*)^T d + o(\|z_k - x^*\|)$$

Since $\lambda_j^* \nabla c_j(x^*)^T d < 0$, we have $f(z_k) > f(x^*)$ for k sufficiently large, which is a contradiction to (10.9). Therefore the case when critical cone contains only zero is proved.

□

10.4.1 Example

Example. Consider the following constraint optimization problem

$$\begin{array}{ll} \min & x^2 + y^2 \\ \text{s.t.} & \frac{x^2}{4} + y^2 - 1 = 0 \end{array}$$

The Lagrangian function is

$$\mathcal{L}(x, y, \lambda) = x^2 + y^2 + \lambda(\frac{x^2}{4} + y^2 - 1)$$

The linearized feasible set is

$$\mathcal{F}(x, y) = \{(d_1, d_2) \mid \frac{x}{2} d_1 + 2y d_2 = 0\}$$

Since there exists only 1 constraint and the gradient of the constraint is nonzero, LICQ holds and $\mathcal{F}(x) = \mathcal{C}(x, \lambda)$. The Lagrangian function is

$$\mathcal{L}(x, y, \lambda) = x^2 + y^2 + \lambda(\frac{x^2}{4} + y^2 - 1)$$

Hence the stationary condition gives

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x} &= 2x + \frac{x}{2}\lambda = 0 \\ \frac{\partial \mathcal{L}}{\partial y} &= 2y + 2\lambda y = 0 \end{aligned}$$

If $\lambda = -4$, then $y = 0$ and $x = 2, -2$. If $\lambda = -1$, then $x = 0$ and $y = 1, -1$. So in total there are 4 KKT points $z = (x, y, \lambda)$

$$(2, 0, -4), \quad (-2, 0, -4), \quad (0, 1, -1), \quad (0, -1, -1)$$

We consider the first KKT point and the third one. Let $z_1 = (2, 0, -4).=$, then

$$\nabla_{xx}^2 \mathcal{L}(z_1) = \begin{bmatrix} 0 & 0 \\ 0 & -6 \end{bmatrix}, \quad \mathcal{C}(z_1) = \{(d_1, d_2) \mid d_1 = 0\}$$

Let $d = (0, 1)$ then

$$d^T \nabla_{xx}^2 \mathcal{L}(z_1) d = -6 < 0$$

Therefore z_1 is not local minimizer (local maximizer). Similarly, let $z_3 = (0, 1, -1)$, then

$$\nabla_{xx}^2 \mathcal{L}(z_3) = \begin{bmatrix} \frac{3}{2} & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathcal{C}(z_1) = \{(d_1, d_2) \mid d_2 = 0\}$$

Let $d = (d_1, 0)$ and $d_1 \neq 0$, then

$$d^T \nabla_{xx}^2 \mathcal{L}(z_3) d = \frac{3}{2} d_1^2 > 0$$

Therefore z_3 is a strict local minimizer.

10.4.2 Finishing the proof for the trust region subproblem

We now finishing the necessary part of Theorem 9.1

Proof. Recall the trust region subproblem is the following:

$$\begin{array}{ll} \min_{d\in\mathbb{R}^n} & c + b^T d + \frac{1}{2} d^T B d = m(d) \\ \text{s.t.} & \|d\| \leq \Delta \end{array}$$

First, if the optimal solution d^* of the trust region subproblem is not on the boundary, then B has to be positive semidefinite (second order necessary condition for unconstrained optimization) and $Bd^* = -b$(stationary condition for unconstrained optimization). Therefore the pair $(d^*, 0)$ satisfies conditions (9.2)

Now let's assume $\|d^*\| = \Delta$. We can assume the constraint is the following $\|d\|^2 = \Delta^2$.

It is easy to see that LICQ constraint qualification holds at $d = d^*$ since $\nabla(d^T d - \Delta^2) = 2d$ which is linearly independent (only 1 vector). So KKT conditions hold at $d = d^*$. Therefore there exists $\lambda^* \geq 0$ such that

$$\begin{aligned} (B + \lambda^* I)d^* &= -b \\ \lambda^*(\|d^*\| - \Delta) &= 0 \end{aligned}$$

It remains to check $B + \lambda^* I \succeq 0$. Since d^* is a global minimizer, we have $m(d) \geq m(d^*) + \frac{\lambda^*}{2}(\|d*\|^2 - \|d\|^2)$ for any d such that $\|d\| = \Delta$. Substitute the expression $b = -(B + \lambda^* I)d^*$ into this expression, and after some rearrangement we obtain

$$\frac{1}{2}(d - d^*)^T (B + \lambda^* I)(d - d^*) \geq 0 \tag{10.13}$$

Since the set of directions

$$\left\{ w \;\middle|\; w = \pm \frac{d - d^*}{\|d - d^*\|}, \text{ for some } d \text{ with } d = \Delta \right\},$$

is dense on the unit sphere, (10.13) suffices to prove $B + \lambda^* I \succeq 0$. □

10.5 Duality Theory

Given a general constrained NLP

$$(P) \qquad \begin{array}{lll} p^* := & \min & f(x) \\ & \text{s.t.} & c_i(x) = 0, i \in \mathcal{E} \\ & & c_i(x) \leq 0, i \in \mathcal{I}. \end{array} \tag{10.14}$$

and the corresponding Lagrangian

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i\in\mathcal{E}} \mu_i c_i(x) + \sum_{i\in\mathcal{I}} \lambda_i c_i(x)$$

where $\mu_i \in \mathbb{R}, \lambda_i \geq 0$.

We call this problem as the *primal problem.*

We define the Lagrangian function as follows:

$$g(\lambda, \mu) = \inf_{x\in\text{dom}(f)} (\mathcal{L}(x, \lambda, \mu))$$

We claim $g(\lambda, \mu)$ is a concave function. In fact,

$$-g(\lambda, \mu) = -\inf_{x\in\text{dom}(f)} (\mathcal{L}(x, \lambda, \mu)) = \sup_{x\in\text{dom}(f)} (-\mathcal{L}(x, \lambda, \mu))$$

Since $\mathcal{L}(x, \lambda, \mu)$ is a linear function in terms of λ and μ, therefore is convex. We know pointwise supremum of a family of convex functions is still convex by Proposition 3.17. So $-g(\lambda, \mu)$ is convex, therefore $g(\lambda, \mu)$ is concave.

The *dual* problem of the primal problem is defined as

$$\begin{array}{lllll} & d^* := & \max & g(\lambda, \mu) & \\ (D) & & \text{s.t.} & \mu_i \in \mathbb{R}, i \in \mathcal{E} & (10.15) \\ & & & \lambda_i \geq 0, i \in \mathcal{I}. & \end{array}$$

Lemma 10.16. (Weak Duality)

The primal optimal value p^* is greater or equal to the dual optimal value d^*. $(p^* \geq d^*)$

We say *strong duality* holds if the primal optimal value is equal to the dual optimal value. $(p^* = d^*)$. The *duality gap* is defined as $p^* - d^*$.

Strong duality connects the global optimal solution with the KKT condition.

THEOREM 10.17. Assume the domain of f is restricted to be an open set. If strong duality holds, then any par of the primal and the dual optimal solutions must be a KKT point.

Proof. Let x^* be a primal optimal and (λ^*, μ^*) be a dual optimal point. This means that

$$\begin{aligned} f(x^*) &= g(\lambda^*, \mu^*) \quad \text{(strong duality)} \\ &= \inf_{x \in \text{dom}(f)} (f(x) + \sum_{i \in \mathcal{I}} \lambda_i^* c_i(x) + \sum_{i \in \mathcal{E}} \mu_i^* c_i(x^*)) \\ &\leq f(x^*) + \sum_{i \in \mathcal{I}} \lambda_i^* c_i(x) + \sum_{i \in \mathcal{E}} \mu_i^* c_i(x^*) \\ &\leq f(x^*) \quad (\lambda_i^* \geq 0, c_i(x^*) \leq 0, \forall i \in \mathcal{I}) \end{aligned}$$

Since $f(x^*) = f(x^*)$, we conclude that the two inequalities in this chain must hold with equality. Therefore we have the complementarity condition of the KKT condition.

$$\sum_{i \in \mathcal{I}} \lambda_i^* c_i(x^*) = 0$$

Also since

$$\inf_{x \in \text{dom}(f)} (f(x) + \sum_{i \in \mathcal{I}} \lambda_i^* c_i(x) + \sum_{i \in \mathcal{E}} \mu_i^* c_i(x^*)) = f(x^*) + \sum_{i \in \mathcal{I}} \lambda_i^* c_i(x) + \sum_{i \in \mathcal{E}} \mu_i^* c_i(x^*)$$

we conclude x^* minimize $\mathcal{L}(x, \lambda^*, \mu^*)$ over x in $\text{dom}(f)$. Since $\text{dom}(f)$ is restricted to be an open set, x^* must be a stationary point of $\mathcal{L}(x, \lambda^*, \mu^*)$, therefore, we have the stationary condition of the KKT condition, i.e.,

$$\nabla f(x^*) + \sum_{i \in \mathcal{I}} \lambda_i^* \nabla c_i(x^*) + \sum_{i \in \mathcal{E}} \mu_i^* \nabla c_i(x^*) = 0$$

□

10.6 Convex Optimization Problem

Again consider the general nonlinear optimization problem:

$$\begin{aligned} p^* := \quad & \min \quad f(x) \\ & \text{s.t.} \quad c_i(x) = 0, i \in \mathcal{E} \\ & \qquad c_i(x) \leq 0, i \in \mathcal{I}. \end{aligned} \tag{10.16}$$

If the objective function $f(x)$ is a convex function in the domain $\text{dom}(f)$. The inequality constraint functions are all convex functions, and the equality constraints are all affine functions. Then it is called *convex optimization problem*

For convex problems, the KKT conditions are also sufficient for global optimality.

THEOREM 10.18. Given a convex optimization problem, if (x^*, λ^*, μ^*) is a KKT point, then x^* is a global optimum of the primal problem, (λ^*, μ^*) is a global optimum of the dual problem, and the duality gap is zero.

Proof. Let x^*, λ^*, μ^* be KKT points that satisfy the KKT conditions:

$$\begin{aligned} c_i(x^*) &\leq 0, \quad i \in \mathcal{I} \\ c_i(x^*) &= 0, \quad i \in \mathcal{E} \\ \lambda_i^* &\geq 0, \quad i \in \mathcal{I} \\ \lambda_i^* c_i(x^*) &= 0, \quad i \in \mathcal{I} \\ \nabla f(x^*) + \sum_{i \in \mathcal{I}} \lambda_i^* \nabla c_i(x^*) + \sum_{i \in \mathcal{E}} \mu_i^* \nabla c_i(x^*) &= 0 \end{aligned}$$

The first two equations state that x^* is primal feasible. Since $\lambda_i^* \geq 0$, $\mathcal{L}(x, \lambda^*, \mu^*)$ is convex in x. The last KKT condition states that $x = x^*$ is a stationary point of $\mathcal{L}(x, \lambda^*, \mu^*)$. Therefore x^* minimizes $\mathcal{L}(x, \lambda^*, \mu^*)$ over x. From this we conclude that

$$\begin{aligned} g(\lambda^*, \mu^*) &= \mathcal{L}(x^*, \lambda^*, \mu^*) \\ &= f(x^*) + \sum_{i \in \mathcal{I}} \lambda_i^* c_i(x) + \sum_{i \in \mathcal{E}} \mu_i^* c_i(x^*) \\ &= f(x^*) \end{aligned}$$

This shows that x^* and (λ^*, μ^*) has zero duality gap and therefore are primal and dual optimal. □

For convex problems, we have an easy-to -check constraint qualifications, which is called the *Slater condition*.

DEFINITION 10.19. (Slater Condition)

Let Ω be the feasible region of a convex optimization problem. If there exists some $x^* \in \Omega$ such that

$$c_i(x^*) < 0, \quad \forall i \in \mathcal{I}$$

Then we say *Slater condition* holds.

THEOREM 10.20. If Slater condition holds for convex optimization problems, then $T_\Omega(x) = \mathcal{F}(x)$ for all $x \in \Omega$

Proof. Proof is omitted. (Theorem 12.5 in the book Numerical Optimization by Nocedal and Wright.) □

THEOREM 10.21. For convex optimization problem, if Slater condition holds, then strong duality holds.

Proof. We assume the optimal value of primal problem is attained at some point x^*, then Slater condition implies x^* is a KKT point by Theorem 10.20. Also for convex problem, KKT point is a global optimum and strong duality holds.

In the case of the optimal value is not attained, this is also true, the proof is omitted. □

10.7 Different formulations of the primal yields different dual

The dual problem is not unique, in fact, for the same optimization problem (geometrically), one can write down the primal problems differently, therefore the dual may have different formulations, and the duality gap can be different.

Examples:

$$\begin{array}{ll} \min & x^3 + y^3 = f(x,y) \\ \text{s.t.} & -x - y \le -1 \\ & x, y \ge 0 \end{array} \tag{10.17}$$

where the domain of the objective function is the whole plane.

The optimal value of the primal problem is $p^* = \frac{1}{4}$ at $x = \frac{1}{2}, y = \frac{1}{2}$.

The Lagrangian dual function is the following:

$$g(\lambda) = \inf_{x,y \in \mathbb{R}} x^3 + y^3 + \lambda_1(-x - y + 1) - \lambda_2 x - \lambda_3 y,$$

which is equal to $-\infty$ no matter what λ is. Therefore, the dual problem

$$\max g(\lambda), \quad s.t. \quad \lambda \ge 0$$

has optimal $d^* = -\infty$. Therefore strong duality doesn't hold. This due to the fact that the objective function is not convex.

Equivalently, the primal problem can be written as follows:

$$\begin{array}{ll} \min_{x,y \in \operatorname{dom} f} & x^3 + y^3 = f(x,y) \\ \text{s.t.} & -x - y \le -1 \end{array} \tag{10.18}$$

where the domain of f is the non-negative orthant of x and y ($x \ge 0, y \ge 0$).

The Lagrangian dual function is the following:

$$g(\lambda) = \inf_{x \ge 0, y \ge 0} x^3 + y^3 + \lambda(-x - y + 1)$$

Since $\lambda \geq 0$, the minimum is attained at the stationary point, i.e., $3x^2 - \lambda = 0$ and $3y^2 - \lambda = 0$. By substituting $x = y = \sqrt{\frac{\lambda}{3}}$ into $g(\lambda)$ we have

$$g(\lambda) = \frac{2\lambda}{3}\sqrt{\frac{\lambda}{3}} - 2\lambda\sqrt{\frac{\lambda}{3}} + \lambda = -\frac{4\lambda}{3}\sqrt{\frac{\lambda}{3}} + \lambda$$

and $\max\{g(\lambda), \lambda \geq 0\} = \frac{1}{4}$ at $\lambda = \frac{3}{4}$. Therefore $p^* = d^* = \frac{1}{4}$, strong duality holds. This is because the objective function is convex at domain $x \geq 0, y \geq 0$, therefore this is a convex optimization problem. It is easy to check Slater condition holds, hence Theorem 10.21 applies here.

11 Algorithms for constrained optimization problem

11.1 Quadratic Penalty method for constrained optimization

Consider the following constrained optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & c_i(x) = 0. \quad i \in \mathcal{E} \end{array} \tag{11.1}$$

DEFINITION 11.1. (Quadratic Penalty for Equality Constraint Optimization)

Define the quadratic penalty function

$$P_{\mathcal{E}}(x, \sigma) = f(x) + \frac{1}{2}\sigma \sum_{i \in \mathcal{E}} c_i^2(x)$$

The parameter σ is called the penalty coefficients.

Algorithm 4: Quadratic Penalty Method

Input: $\sigma_0 > 0$, x^0, $\rho > 1$
while *Stopping criteria not satisfied* **do**
  Solve $x^{k+1} = \arg\min_x P_E(x, \sigma_k)$ using x^k as an initial point
  $\sigma_{k+1} \leftarrow \rho\sigma_k$
  $k \leftarrow k + 1$
end

THEOREM 11.2. Let x^{k+1} be the global minimizer of $P_E(x, \sigma_k)$, let $\sigma_k \to \infty$. Then any accumulation point of $\{x_k\}$ is a global minimizer of the constrained optimization problem (11.1)

Proof. Let x^* be a global minimizer of problem (11.1). Then

$$P_E(x^{k+1}, \sigma_l) \leq P_E(x^*, \sigma_k)$$

which is equivalent to

$$f(x^{k+1}) + \frac{\sigma_k}{2} \sum_{i \in \mathcal{E}} c_i^2(x^{k+1}) \leq f(x^*) + \frac{\sigma}{2} \sum_{i \in \mathcal{E}} c_i(x^*) = f(x^*)$$

After rearrangement, we have

$$\sum_{i \in \mathcal{E}} c_i^2(x^{k+1}) \leq \frac{2}{\sigma_k}(f(x^*) - f(x^{k+1}))$$

Let $\hat{x}$ be any accumulation point of $\{x^k\}$. Let $k \to \infty$, then $\sigma_k \to \infty$ which implies

$$\sum_{i \in \mathcal{E}} c_i^2(\hat{x}) = 0$$

Therefore, $\hat{x}$ is a feasible point. Also since $f(x^{k+1}) \leq f(x^*)$ we have $f(\hat{x}) \leq f(x^*)$. Therefore $\hat{x}$ is also a global minimizer. □

Consider the following inequality constrained optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & c_i(x) \leq 0. \quad i \in \mathcal{I} \end{array} \tag{11.2}$$

DEFINITION 11.3. (Quadratic Penalty for Inequality Constraint Optimization)

Define the quadratic penalty function

$$P_{\mathcal{I}}(x, \sigma) = f(x) + \frac{1}{2}\sigma \sum_{i \in \mathcal{I}} \tilde{c}_i^2(x)$$

where $\tilde{c}_i(x) = \max\{c_i(x), 0\}$

Note that $h(t) = (\max\{t, 0\})^2$ is differentiable with regard to t, therefore the gradient of $P_I(x, \sigma)$ exists. So gradient descent method can be applied here.

Consider the following general equality and inequality constrained optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & c_i(x) = 0. \quad i \in \mathcal{E} \\ & c_i(x) \leq 0. \quad i \in \mathcal{I} \end{array} \tag{11.3}$$

DEFINITION 11.4. (Quadratic Penalty for Equality and Inequality Constraint Optimization)

Define the quadratic penalty function

$$P_{\mathcal{I}}(x, \sigma) = f(x) + \frac{1}{2}(\sigma \sum_{i \in \mathcal{I}} \tilde{c}_i^2(x) + \sum_{i \in \mathcal{E}} c_i^2(x))$$

where $\tilde{c}_i(x) = \max\{c_i(x), 0\}$

11.2 Application to LASSO problem

LASSO problem is the following problem

$$\min_x \frac{1}{2} \|Ax - b\|^2 + \mu \|x\|_1$$

The goal of LASSO is to solve the following Basis Pursuit (BP) problem

$$\begin{array}{ll} \min & \|x\|_1 \\ \text{s.t.} & Ax = b \end{array}$$

We use a quadratic penalty term on the constraint to obtain

$$\min_x \|x\|_1 + \frac{\sigma}{2} \|Ax - b\|^2$$

Therefore when $\sigma = \frac{1}{\mu}$, LASSO problem is equivalent to the quadratic penalty subproblem of (BP) problem.

11.2.1 Example of unstable solutions

Example.

$$\begin{array}{ll} \min & (x-1)^2 + (y-1)^2 \\ \text{s.t.} & x + y = 4 \end{array}$$

Drawbacks of quadratic penalty method

1. As σ grows larger, it becomes much harder to solve the penalty function minimization problem.
2. Solutions becomes very unstable from ill-conditioning of the problem, resulting poor convergence as we increase σ.

11.3 Augmented Lagrangian Method for Constraint Optimization

The augmented Lagrangian method reduce the possibility of ill-conditioning by introducing an explicit estimate of the Lagrangian multiplier.

Define

$$\mathcal{L}_A(x, \mu, \rho) = f(x) - \sum_{i \in \mathcal{E}} \mu_i c_i(x) + \frac{\sigma}{2} \sum_{i \in \mathcal{E}} c_i^2(x)$$

The idea is to alternatively update x and μ_i as we increase σ, so in the end, it converges to a stationary point which is also feasible.

At step k by differentiating the augmented Lagrangian w.r.t x we get

$$\nabla_x \mathcal{L}_A(x_k, \mu^k, \sigma_k) = \nabla f(x_k) - \sum_{i \in \mathcal{E}} [\mu_i^k - \sigma_k c_i(x_k)] \nabla c_i(x^k)$$

which suggests a formula for updating μ^i

$$\mu_i^{k+1} = \mu_i^k - \sigma_k c_i(x^{k+1})$$

To compute x^{k+1}, we compute

$$x^{k+1} = \arg\min_x \mathcal{L}_A(x, \mu^k, \sigma_k)$$

where the initial point is x^k.

To update σ_k, simply compute

$$\sigma_{k+1} = \rho \sigma_k$$

where $\rho > 1$

Algorithm 5: Augmented Lagrangian method

Input: $\sigma_0 > 0$, starting point x^0, μ_0, and factor $\rho > 1$, $k = 0$.
while *Stopping criteria not satisfied* **do**
  Solve $x^{k+1} = \arg\min_x \mathcal{L}_A(x, \mu^k, \sigma_k)$ approximately using x^k as an initial point
  Set $\mu_i^{k+1} = \mu_i^k - \sigma_k c_i(x^{k+1})$
  Set $\sigma_{k+1} = \rho \sigma_k$
  $k \leftarrow k + 1$
end

The convergence of Augmented Lagrangian method can be assured without increasing σ indefinitely. Ill conditioning therefore is less of a problem than the quadratic penalty method. We have the following theorem.

THEOREM 11.5. Let x^* ve a local minimizer of (11.1) at which LICQ holds, and the second order sufficient condition in Theorem 10.15 are satisfied for $\mu = \mu^*$. Then there exists a threshold $\bar{\sigma}$ such that for all $\sigma \geq \bar{\sigma}$, x^* is a strict local minimizer of $\mathcal{L}_A(x, \mu^*, \sigma)$

Proof. □

11.4 Interior Point Method for Conic Optimization Problems

11.4.1 Interior Point Method for Linear Programming

Consider the following primal and dual linear program

$$(P) \quad \begin{array}{ll} \min & c^T x \\ \text{s.t.} & Ax = b \\ & x \geq 0 \end{array}$$

$$(D) \quad \begin{array}{ll} \max & b^T y \\ \text{s.t.} & c - A^T y = s \\ & s \geq 0 \end{array}$$

with KKT condition (if and only if for optimality of LP)

$$\begin{aligned} Ax &= b \\ A^T y + s &= c \\ x_i s_i &= 0, \quad i = 1, ..., n \\ x \geq 0, s &\geq 0 \end{aligned}$$

The (primal) interior point method assures the iterated points stay in the interior of the feasible region ($x \geq 0$) of the primal problem by introducing a barrier function as a penalty function

$$\phi(x) = -\sum_i \log(x_i)$$

Algorithm 6: Primal Interior Method for LP

Input: $\sigma_0 > 0$, starting point $x^0 > 0$ such that $Ax^0 = b$, and factor $0 < \rho < 1$, $k = 0$.
while *Stopping criteria not satisfied* **do**
  Solve $x^{k+1} = \arg\min_x \{c^T x + \sigma^k \phi(x) : Ax = b\}$ approximately using x^k as an initial point
  Set $\sigma_{k+1} = \rho \sigma_k$
  $k \leftarrow k + 1$
end

The KKT condition of problem $\min_x \{c^T x + \sigma^k \phi(x) : Ax = b\}$ is following

$$\begin{aligned} Ax &= b \\ A^T y + s &= c \\ x_i s_i &= \sigma^k, \quad i = 1, ..., n \\ x > 0, s &> 0 \end{aligned}$$

Here $x > 0$ due the log function is defined over $x > 0$ and $s > 0$ due to $\sigma^k > 0$

Therefore as σ^k goes to zero, the KKT condition of the penalty problem becomes the KKT condition of the original linear programming problem. Hence x^k is converging to the optimal solution of the LP problem.

If we use the quadratic approximation of $\phi(x)$ at $x = x^k$ where x^k is a feasible point from previous iteration, we have

$$\phi(x) \approx \phi(x^k) + (x - x^k)^T \nabla \phi(x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 \phi(x^k)(x - x^k)$$

and $A(x^k) = b$.

The KKT condition for the approximate constraint quadratic optimization problem is

$$\begin{bmatrix} \sigma^k \nabla^2 \phi(x^k) & A^T \\ A & 0 \end{bmatrix} \cdot \begin{bmatrix} x - x^k \\ y \end{bmatrix} = \begin{bmatrix} -c - \sigma^k \nabla \phi(x^k) \\ 0 \end{bmatrix}$$

where

$$\nabla\phi(x^k) = -\begin{bmatrix} \frac{1}{x_1^k} \\ \vdots \\ \frac{1}{x_n^k} \end{bmatrix}, \quad \nabla^2\phi(x^k) = \begin{bmatrix} \frac{1}{(x_1^k)^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{(x_n^k)^2} \end{bmatrix}$$

This is just a linear system which can be solved efficiently by Gaussian elimination. In fact, if we cancel s in the KKT conditions, i.e.,

$$\begin{aligned} Ax &= b \\ A^T y + [\tfrac{\sigma_k}{x_1}, \cdots, \tfrac{\sigma_k}{x_n}]^T &= c, \end{aligned}$$

then apply first order Taylor expansion at x^k, we obtain

$$\begin{aligned} Ax^k + A(x - x^k) &= b \\ A^T y + \sigma^k(\nabla\phi(x^k) - \nabla^2\phi(x^k)(x - x^k) &\approx c \end{aligned}$$

which is the same linear system as before and this is just one Newton root-finding step , therefore after a few Newton root-finding steps, we can get a good approximate solution of the penalty subproblem.

11.4.2 From Interior to Vertex: Purification

When an interior solution is close enough to the optimal solution, we can switch to a purification procedure to get an exact solution of a linear program whose objective value is not greater than the objective value of the interior solution.

Algorithm 7: Purify an Interior Solution to a Vertex Solution

Input: $A \in \mathbb{R}^{m\times n}$ such that $\text{rank}(A) = m$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$. An interior solution $x^* \in \mathbb{R}^n$
Output: x is a vertex solution. Let $x = x^*$.
for $k = 1, ..., n - m$ **do**
  Compute $0 \neq d \in \text{Null}\,(A)$ such that $c^T d \leq 0$. (make sure objective value is non-increasing)
  Find the maximum $t \geq 0$ such that $x + td \geq 0$
  Update $x := x + td$ (at least one more component of x becomes zero)
  Find i such that $x_i = 0$ (i should be different from the previous ones) and update

$$A := \begin{bmatrix} A \\ e_i^T \end{bmatrix} (e_i \text{ is the } i\text{th unit vector})$$

end

11.4.3 Complexity of Linear Programming and Practical Implementation

Let L be the bit size of the input data from a LP problem. Let x_1 and x_2 be two vertices of the feasible region, if $c^T x_1 \neq c^T x_2$, then $|c^T x_1 - c^T x_2| > 2^{-2L}$. Therefore if we come within $2^{-O(L)}$ of the optimal value, then after purification, we get an exact optimal solution. In general, the complexity of linear programming is $O(n^{3.5}L^2)$ on a Turing machine (Karmarkar, A New Polynomial-Time Algorithm for Linear Programming, Combinatorica 4, 1984). The first polynomial algorithm for linear programming is discovered by Khachiyan using ellipsoid method in 1979, but it is not practical.

In practice, you can run an interior point method, and purify your solution to a vertex solution for every t steps of iterations. Then check if the vertex is optimal, if it is optimal, then stop, otherwise continue the interior point method.

11.4.4 General Conic Programming

The idea of interior point method for LP can be generated to more general conic programming problems

Recall K is a convex cone if K is convex, nonempty and $x \in K, \lambda \geq 0 \implies \lambda x \in K$.

DEFINITION 11.6. (Conic Program)

A conic program is the following optimization problem

$$\begin{array}{ll} \min & c^T x \\ \text{s.t.} & Ax = b \\ & x \in K \end{array}$$

where K is a closed convex cone.

Three most common used convex cones are:

$$\mathbb{R}^n_+ = \{x \in \mathbb{R}^n, x \geq 0\}$$

$$C_2^{n+1} = \{(y, x) \in \mathbb{R} \times \mathbb{R}^n : y \geq ||x||_2\}$$

$$\mathcal{S}^n_+ = \{X \in \mathbb{R}^{n \times n} : X \succeq 0\}$$

The standard barrier functions used for the three cones are

- $\mathbb{R}^n_+ : \phi(x) = -\sum_i \log x_i$
- $C_2^{n+1} : \phi(y, x) = -\log(y^2 - \|x\|_2^2)$
- $\mathcal{S}^n_+ : \phi(X) = -\log(\det(X))$

A framework for a general conic program is the following:

Algorithm 8: Primal Interior Method for Conic Programming

Input: $\sigma_0 > 0$, starting point $x^0 \in \text{int}(K)$ (interior of a closed convex cone K) such that $Ax^0 = b$, and factor $0 < \rho < 1$, $k = 0$.

while *Stopping criteria not satisfied* **do**

  Solve $x^{k+1} = \arg\min_x \{c^T x + \sigma^k \phi(x) : Ax = b\}$ approximately using x^k as an initial point

  $\sigma_{k+1} \leftarrow \rho \sigma_k$

  $k \leftarrow k + 1$

end

.

12 Introduction to Neural network

The material for this section is from the paper *Deep Learning: An Introduction for Applied Mathematicians* by Catherine Higham and Desmond Higham. The supplementary material and matlab files for this paper can be downloaded here.

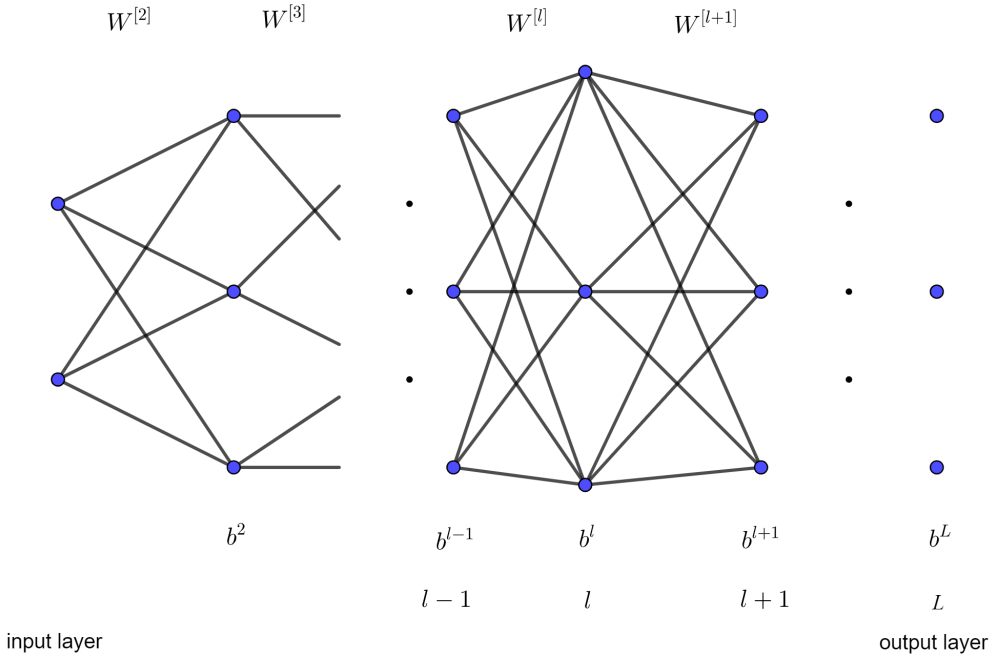


Figure 18: Neural network

A neuron network consists of L layers of neurons. The first layer is the input, the last layer is the output. There are n_l neurons at layer l.

The neuron k at layer $l-1$ are connected to neuron j at layer l by a weight $W_{kj}^{[l]}$. So at layer l, the associated weight matrix is an $n_l \times n_{l-1}$ matrix.

For neurons at layer l, we associate them with a vector $b^{[l]}$ which is called the *bias*. So $b_i^{[l]}$ is the bias for neuron i at layer l.

We use the sigmoid function to simulate the output of a neuron.

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

The sigmoid function approaches 1 as $x \to +\infty$. It approaches 0 as $x \to -\infty$. So it mimics the behavior of a neuron in the brain-firing (giving output close to one) if the input is large enough, and remaining inactive (giving output close to zero) otherwise.

The derivative of the sigmoid function has a simple form:

$$\sigma'(x) = \sigma(x)(1-\sigma(x))$$

For $z \in \mathbb{R}^m$, $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ is defined by applying the sigmoid function in the obvious componentwise manner, so that

$$(\sigma(z))_i = \sigma(z_i)$$

Let $a^{[l]}$ denote the output or the *activation* from neuron j at layer l. Then we have

$$a^{[1]} = x \in \mathbb{R}^{n_1}$$

which is the input from a data point x. and

$$a^{[l]} = \sigma(W^{[l]}a^{[l-1]} + b^{[l]}) \in \mathbb{R}^{n_l}, \text{ for } l = 2, 3, ..., L$$

where n_l is the number of neurons at Layer l, $W^{[l]}$ is a matrix of $n_{l-1} \times n_l$ and $W^{[l]}_{kj}$ is the weight from neuron k at layer $l-1$ to neuron j at layer l, $b^l \in \mathbb{R}^{n_l}$ is the bias vector at Layer l. L is the total number of layers.

The goal of a neural network is to minimize the cost function w.r.t to W and b.

$$\text{Cost } = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}||y(x^{\{i\}}) - a^{[L]}(x^{\{i\}})||$$

where $y(x^{\{i\}})$ the label of data point $x^{\{i\}}$ and N is the number of training data points.

Since the cost is the summation of the cost from every data point, we may drop the dependence on $x^{\{i\}}$ and simply write

$$C = \frac{1}{2}||y - a^{[L]}||^2$$

for one data point x and its label y

Let $z^{[l]}$ be the weighted input vector for neurons at layer l, then

$$z^{[l]} = W^{[l]}a^{[l-1]} + b^{[l]} \in \mathbb{R}^{n_l}$$

The output or activation from neurons at layer l is

$$a^{[l]} = \sigma(z^{[l]}), \text{ for } l = 2, 3, ..., L$$

For computational purpose, we define $\delta^{[l]} \in \mathbb{R}^{n_l}$ as

$$\delta^{[l]} = \frac{\partial C}{\partial z^{[l]}_j} \quad \text{for } 1 \leq j \leq n_l, \quad 2 \leq l \leq L$$

12.1 Back propagation

Lemma 12.1. We have

- $\delta^{[L]} = \sigma'(z^{[L]}) \circ (a^{[L]} - y)$
- $\delta^{[l]} = \sigma'(z^{[l]}) \circ (W^{[l+1]})^T\delta^{[l+1]}$
- $\dfrac{\partial C}{\partial b^{[l]}_j} = \delta^{[l]}_j$
- $\dfrac{\partial C}{\partial w^{[l]}_{jk}} = \delta^{[l]}_j a^{[l-1]}_k$

Proof. Here $\circ$ is the Hadamard product (componentwise multiplication). See Lemma 1 in the paper for the proof. The proof is a straightforward application of chain rule. □

The output $a^{[L]}$ can be computed from a forward pass through the network, by computing $a^{[1]}, z^2a^{[2]}, z^3, a^{[3]}, ..., a^L$ in order. Then $\delta^{[L]}$ is immediately available, Then $\delta^{[L-1]}, \delta^{[L-2]}, ..., \delta^{[2]}$ can be computed in a backward pass. Then we have access to all the partial derivatives. Computing gradients in this way is known as *back propagation*.

12.2 Stochastic gradient descent

After we compute the gradient by back propagation, we can run gradient descent method, i.e.,

$$W^{[l]} \leftarrow W^{[l]} - \eta \frac{\partial C}{\partial W^{[l]}}$$

and

$$b^{[l]} \leftarrow b^{[l]} - \eta \frac{\partial C}{\partial b^{[l]}}$$

Here η is the step size. In machine learning it is also called *learning rate*. Usually η is chosen as a constant.

In practice, there are a lot of data points from the training set. Therefore it is not efficient to compute the gradient for all the data points. Instead, people often choose a set of k points randomly from the training data set and do a few steps of gradient descent, then choose k points randomly again and do a few steps of gradient descent. Repeating this process until certain stopping criteria are met. This is called *stochastic gradient descent* method.

13 Course review for final

- Psd and pd matrix, matrix and vector norms, pseudo inverse
- Convexity, strong convexity
- Coercive functions
- Unconstrained quadratic optimization (first and second order optimal conditions)
- Least square problem
- Line search algorithm (Armoji condition, Wolfe condition, back tracking) Gradient descent method, Newton's method
- Trust region method (easy case and hard case)
- Constrained optimization (KKT conditions, second order optimal conditions, constraint qualifications)
- Duality theory, convex optimization problems, Slater condition
- Constrained optimization algorithms (penalty method, log barrier penalty method, interior method for conic programming)

Index