高级工程数学个性化教材

教师	序号	姓名	学号
沈超敏	76	吴振昊	51275901108

目录

序章 数学史

数学的词源：

- 数学一词起源于希腊语“μαθηματικός (Mathematikós)”，意为“学问的基础”，其更早的词根“μάθημα (máthema)”意为“科学、知识或学问”。
- 数学的专业化使用可以追溯到毕达哥拉斯学派，这一学派首次明确将“数学”与数的研究及逻辑推理联系起来。

关键数学家与贡献：

- **泰勒斯 (Thales)**: 提出了著名的“半圆内切角是直角”的几何理论，为后来的几何学奠定了基础。
- **毕达哥拉斯 (Pythagoras)**: 在数学领域具有开创性贡献，特别是对数论和比例的研究。
- **欧几里得 (Euclid)**: 以《几何原本》奠定了公理化方法，将几何系统化为一个逻辑体系。
- **阿基米德 (Archimedes)**: 以数学的方式研究物理学问题，被称为古代数学的巅峰人物，对微积分的萌芽发展起到了重要作用。
- 微积分的创立者是**牛顿 (Isaac Newton)和莱布尼茨 (Gottfried Wilhelm Leibniz)**，他们独立地提出了这一伟大的数学工具。

数学史上的里程碑：

- 微积分的萌芽可以追溯到阿基米德的研究，他通过分割法接近曲线下的面积，间接为微积分的诞生奠定了思想基础。
- 在此后的 1500 多年间，数学从古希腊的几何学逐步扩展到代数和解析几何，为近代数学的快速发展打下了基础。
- 牛顿和莱布尼茨的微积分推动了数学的现代化发展，使数学在物理、天文等领域的应用取得重大突破。

数学发展的视角：

- 学生高中毕业时的数学水平相当于 400 年前的水平。
- 学完高等数学的学生可以达到 150 年前的数学水平。
- 学完本课程后，数学水平提升至约 50 年前的水平，同时掌握部分前沿内容。

名言引用：

- 普希金认为，数学是“跟随伟大人物的思想，是一门最引人入胜的科学”。

第一章 预备知识：数学证明方法与基本记号

本章导言：

本章作为教材的开篇，旨在为读者建立坚实的数学基础。我们将从最基本的数学证明方法入手，介绍常用的逻辑运算符，并详细讲解直接证明法、反证法、逆否命题证明法和数学归纳法等重要的证明技巧。此外，本章还将规范化数学符号的使用，包括标量、向量、矩阵和集

合的表示方法，这些都是后续章节学习的基石。通过本章的学习，读者将掌握严谨的数学思维方式，并具备使用规范数学语言进行表达的能力。

1.1 数学证明方法

1.1.1 命题与逻辑运算

- **命题**：明确陈述的、可以判断真假的陈述句，例如 "2 是偶数" (真命题) 或 "3 是偶数" (假命题)。
- **逻辑运算符:**
  - **与 (and)**：命题 A and B，当 A 和 B 都为真时，结果为真；否则为假。
  - **或 (or)**：命题 A or B，当 A 或 B 至少有一个为真时，结果为真；当 A 和 B 都为假时，结果为假。
  - **非 (not)**：命题 not A，当 A 为真时，结果为假；当 A 为假时，结果为真。
- **真值表**：使用表格形式清晰地展示逻辑运算的结果。

A	B	A and B	A or B
True	True	True	True
True	False	False	True
False	True	False	True
False	False	False	False

A	not A
True	False
False	True

- **德摩根定律 (DeMorgan's Law):** not (A and B) 等价于 (not A) or (not B)

1.1.2 蕴含关系与等价关系

- **蕴含 (implies)**： 命题 A implies B，当 A 为真时，B 必然为真。常用符号： $A \Rightarrow B$
  - 等价表达: (not A) or B
  - 理解方式：将 A 分为“成立”和“不成立”两种情况进行讨论
  - 蕴含的几种常用表达方式：
    - If A then B
    - A only if B (B 成立时 A 才成立，反之 A 不成立则 B 也不成立)
    - A is sufficient for B （A 是 B 的充分条件）
    - B is necessary for A (B 是 A 的必要条件)
- **等价 (equivalent)**：命题 A is equivalent to B，当 A 为真时，B 也为真；当 A 为假时，B 也为假。 常用符号：$A \Leftrightarrow B$
  - 等价表达： $(A \Rightarrow B)$ and $(B \Rightarrow A)$
  - 也可以表达为 A if and only if B
  - 等价关系的证明通常需要证明双向的蕴含关系。

1.1.3 数学证明的基本方法

- **直接证明法 (The direct method)**：从已知条件出发，通过逻辑推理，逐步推导出结论。
- **逆否命题证明法 (Proof by contraposition)**：要证明 $A \Rightarrow B$，可以转化为证明 (not B) $\Rightarrow$ (not A)。
- **反证法 (Proof by contradiction)**：先假设结论不成立，然后通过逻辑推理，导出与已知条件或公理矛盾的结果，从而证明结论是成立的。
- **数学归纳法 (Principle of induction)**：用于证明与自然数相关的命题。包含以下两个

步骤：

1. **基本情况 (Base Case)**：证明当 $n = 1$ 时，命题成立。
2. **归纳步骤 (Inductive Step)**：假设当 $n = k$ 时，命题成立；然后证明当 $n = k + 1$ 时，命题也成立。
   - **注意事项**：在使用数学归纳法时，一定要注意基本情况的验证，否则可能导致错误的结论。

1.2 基本记号

1.2.1 标量、向量与矩阵

- **标量 (Scalar):** 用小写字母表示，例如：x, y, a, b。
- **向量 (Vector):** 用小写粗体字母表示，例如：**x**, **y**, **a**, **b**。一个列向量 x 可以表示为 $x = (x1, x2, \cdots, xn)^T$，其中 x1, x2, ⋯, xn 是向量的分量。
- **矩阵 (Matrix):** 用大写粗体字母表示，例如：**A**, **B**。一个 m × n 的矩阵 **A** 可以表示为：

  A = | a_{11} a_{12} ... a_{1n} |
  | a_{21} a_{22} ... a_{2n} |
  | ... ... |
  | a_{m1} a_{m2} ... a_{mn} |
- **转置 (Transpose):** 向量或矩阵的转置用上标 T 表示。例如，列向量 **a** 的转置是行向量 $\mathbf{a}^T$ = [a1, a2, ..., an]。矩阵 **A** 的转置是将矩阵的行和列互换。

1.2.2 集合及其表示

- **集合 (Set):** 用大写花体字母表示，例如：X, Y。
- **集合元素 (Element):** 用小写字母表示，例如：x, y。
- **集合表示方法：**
  - 列举法: X = {x1, x2, ⋯, xn}
  - 描述法: X = {x | x 满足某种性质}

本章小结

本章介绍了数学证明的基础知识和基本记号，这是进行高级工程数学学习的必要准备。通过对逻辑运算符、证明方法和数学符号的理解和掌握，读者将为后续章节的学习打下坚实的基础。

第二章 向量空间与矩阵

本章导言：

本章将深入探讨向量空间与矩阵的概念，这是高级工程数学中极为重要的基石。我们将介绍向量的定义、向量的基本运算，以及向量空间的概念和性质。随后，我们将详细讲解矩阵的定义、矩阵运算以及矩阵的秩等重要概念，并介绍线性方程组及其解的存在性和求解方法。此外，本章还将引入内积、范数等概念，它们是衡量向量和矩阵大小的重要工具，为后续的分析和计算打下基础。通过本章的学习，读者将掌握线性代数的基本理论和方法，为解决实际工程问题提供数学工具。

2.1 向量与矩阵

2.1.1 向量的定义与表示

- **列向量 (Column vector):** n 个数的有序排列，表示为一个列的形式。
    - 例如：$\mathbf{a} = [a_1, a_2, \cdots, a_n]^T$
    - a_i 表示向量 **a** 的第 i 个分量。
- **行向量 (Row vector):** n 个数的有序排列，表示为一个行的形式。
    - 例如：$[a_1, a_2, \cdots, a_n]$
- **向量的相等:** 两个向量的对应元素均相等时，两向量相等。
- **转置 (Transpose):** 将列向量转化为行向量，或将行向量转化为列向量。例如：$\mathbf{a}^T = [a_1, a_2, \cdots, a_n]$

2.1.2 向量的基本运算

- **向量加法:** 两个向量对应分量相加。
    - **交换律 (Commutative law):** $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$
    - **结合律 (Associative law):** $(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$
    - **零向量 (Zero vector):** 存在一个零向量 $\mathbf{0} = [0, 0, \cdots, 0]^T$，使得 $\mathbf{a} + \mathbf{0} = \mathbf{0} + \mathbf{a} = \mathbf{a}$
- **向量的减法:** $\mathbf{a} - \mathbf{b} = \mathbf{a} + (-\mathbf{b})$，其中 $-\mathbf{b} = \mathbf{0} - \mathbf{b}$。
    - **向量的差 (Difference):** 向量 **a** 与 **b** 的差是 $[a_1 - b_1, a_2 - b_2, \cdots, a_n - b_n]^T$。
    - **性质：**
        - $(- \mathbf{b}) = \mathbf{b}$
        - $-(\mathbf{a} - \mathbf{b}) = \mathbf{b} - \mathbf{a}$
- **数乘 (Scalar multiplication):** 一个向量乘以一个标量。
    - 若 $\mathbf{a} \in R^n$，标量 $\alpha \in R$，则 $\alpha\mathbf{a} = [\alpha a_1, \alpha a_2, \cdots, \alpha a_n]^T$
    - **分配律 (Distributive law):**
        - $\alpha(\mathbf{a} + \mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}$
        - $(\alpha + \beta)\mathbf{a} = \alpha\mathbf{a} + \beta\mathbf{a}$
    - **结合律 (Associative law):** $\alpha(\beta\mathbf{a}) = (\alpha\beta)\mathbf{a}$
    - **单位标量:** 存在单位标量 1 使得 $1\mathbf{a} = \mathbf{a}$
    - **向量 0:** 对于任意标量 α，有 $\alpha\mathbf{0} = \mathbf{0}$
    - **零标量:** 对于任意向量 **a**，有 $0\mathbf{a} = \mathbf{0}$
- **数乘性质**:
    - $\alpha\mathbf{a} = \mathbf{0} \Leftrightarrow \alpha = 0$ 或 $\mathbf{a} = \mathbf{0}$
    - 此性质可以推导出：如果 $\alpha\mathbf{a} = \mathbf{0}$ 且 $\mathbf{a} \neq \mathbf{0}$，则 $\alpha = 0$.

2.1.3 向量的线性相关性与基

- **线性组合 (Linear combination):** 向量的线性组合是由向量乘以标量并相加得到的。例如：$\alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \ldots + \alpha_n\mathbf{a}_n$
- **线性无关 (Linearly independent):** 一组向量，如果它们的任何非零线性组合都不等于零向量，则称这些向量线性无关。
    - 换句话说，如果 $\alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \ldots + \alpha_n\mathbf{a}_n = \mathbf{0}$ 仅当 $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$ 成立，则向量 $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$ 线性无关。
- **线性相关 (Linearly dependent):** 如果一组向量不是线性无关的，则称它们是线性相关的，即存在一组不全为零的标量 $\alpha_1, \alpha_2, \cdots, \alpha_n$，使得 $\alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \ldots + \alpha_n\mathbf{a}_n = \mathbf{0}$
- **向量空间的基 (Basis):** 对于一个向量空间 V 的子空间，如果一组线性无关的向量

$\{\mathbf{a}_1, \cdots, \mathbf{a}_k\}$ 能够张成这个子空间，那么称这组向量为该子空间的一个基。也就是说，子空间内的任意向量都能用该基的线性组合表示。

- **向量空间的维数 (Dimension):** 一个向量空间的所有基包含的向量个数都相同，这个数就称为向量空间的维数，记为 dim V.
- **基的唯一性表示**: 对于 V 中的任何向量 a，都可以被唯一地表示为 $a = a_1 a_1 + \ldots + a_k a_k$
- **坐标 (Coordinates):** 在一组基下，向量 a 的表示式 $a = a_1 a_1 + \ldots + a_k a_k$ 中，系数 $a_1, \ldots, a_k$被称为向量 a 在该基下的坐标。
- **标准基 (Natural basis):** R^n 的标准基是一组特殊的基，由向量 $e_1 = [1, 0, \ldots, 0]^T$，$e_2 = [0, 1, \ldots, 0]^T, \ldots, e_n = [0, 0, \ldots, 1]^T$ 构成。

2.1.4 矩阵的定义与表示

- **矩阵 (Matrix):** 由 m 行 n 列的数字组成的一个矩形阵列。一个 m × n 的矩阵可以表示为:
  - $\mathbf{A} = [a_{ij}]$，其中 $i = 1, 2, \ldots, m$， $j = 1, 2, \ldots, n$
- **矩阵的转置 (Transpose):** 将矩阵的行和列互换。 例如， $A = [a_{ij}]$，则 $A^T = [a_{ji}]$
- **矩阵的相等**: 两个矩阵的行数和列数都相等，且对应元素都相等时，两矩阵相等。

2.2 矩阵的秩

2.2.1 矩阵的秩的定义

- 对于矩阵 A，其列向量构成的集合中，线性无关的列向量的最大数目称为矩阵 A 的秩，记为 rank A。
- 矩阵 A 的秩也是矩阵 A 列向量所张成的子空间的维度。
- 矩阵 A 的第 k 列记为 a_k，则 $A = [a_1, \cdots, a_n]$

2.2.2 矩阵秩的性质

- **秩的不变性 (Invariance of rank)**: 矩阵 A 的秩在下列操作下保持不变:
  1. 将矩阵 A 的列向量乘以非零标量。
  2. 交换矩阵 A 的列向量。
  3. 将矩阵 A 的一列加上其他列的线性组合。
- 矩阵 A 中，线性无关向量的个数与排列顺序无关。

2.2.3 行列式 (Determinant)

- **行列式的定义**: 对于方阵 A，可以定义一个标量，称为行列式，记为 det A 或 |A|。
- **行列式的性质**:
  - 矩阵的行列式是矩阵每列的线性函数
    - $\det[a_1, \cdots, a_{k-1}, \alpha \mathbf{a}_k^{(1)} + \beta \mathbf{a}_k^{(2)}, a_{k+1}, \ldots, a_n] = \alpha \det[a_1, \cdots, \mathbf{a}_k^{(1)}, a_{k+1}, \ldots, a_n] + \beta \det[a_1, \cdots, \mathbf{a}_k^{(2)}, a_{k+1}, \ldots, a_n]$
  - 如果矩阵 A 中存在两列向量相等，即 $\mathbf{a}_k = \mathbf{a}_{k+1}$ ， 则 $\det A = 0$
  - 单位矩阵 I_n 的行列式等于 1，即 $\det I_n = 1$
- **其他性质:**
  - $\det[a_1, \ldots, a_{k-1}, a_k + \alpha a_j, a_{k+1}, \ldots, a_j, \ldots, a_n] = \det[a_1, \ldots, a_n]$
- **p 阶子式**: 一个 m x n 的矩阵 A 的 p 阶子式，其中 p<min{m,n}，指通过删除矩阵 m-p 行和 n-p 列后，得到的 p x p 方阵的行列式值
- **定理**: 如果一个 m x n (m > n) 的矩阵 A 有一个非零的 n 阶子式，则矩阵 A 的列是线性无关的，即 rank A = n。

2.2.4 非奇异矩阵

- **非奇异矩阵 (Nonsingular matrix):** 对于一个 $n \times n$ 的方阵 A，如果存在一个 $n \times n$ 的矩阵 B，使得 $AB = BA = In$，则称 A 是非奇异的，B 为 A 的逆矩阵，记为 $B = A^{-1}$

2.3 线性方程组

2.3.1 线性方程组的表示

- **线性方程组:** 包含若干个线性方程的方程组。
  - 例如：

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1$$
$$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2$$
$$\ldots$$
$$a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m$$

  - 可以表示为：$\mathbf{Ax = b}$，其中 A 是系数矩阵，$\mathbf{x}$ 是未知向量，$\mathbf{b}$ 是常数向量。
  - 也可以表示为： $x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \ldots + x_n\mathbf{a}_n = \mathbf{b}$，其中 $\mathbf{a}_1, \cdots, \mathbf{a}_n$ 是矩阵 A 的列向量。
- **增广矩阵 (Augmented matrix):** $[\mathbf{A}, \mathbf{b}] = [\mathbf{a}_1, \cdots, \mathbf{a}_n, \mathbf{b}]$。

2.3.2 线性方程组解的存在性

- **定理 2.1:** 线性方程组 $\mathbf{Ax = b}$ 有解的充分必要条件是 $\text{rank}\,\mathbf{A} = \text{rank}\,[\mathbf{A}, \mathbf{b}]$。
  - **证明:**
    - **充分性 (=>)**：若 $Ax = b$ 有解，则 $\mathbf{b}$ 可以表示为 $\mathbf{A}$ 的列向量的线性组合，从而 $\mathbf{b} \in \text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_n)$.因此， $\text{rank}\,\mathbf{A} = \dim \text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_n) = \dim \text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_n, \mathbf{b}) = \text{rank}\,[\mathbf{A}, \mathbf{b}]$
    - **必要性 (<=>)**：若 $\text{rank}\,\mathbf{A} = \text{rank}\,[\mathbf{A}, \mathbf{b}] = r$，则 $\mathbf{A}$ 的前 r 列线性无关。 由于 $\text{rank}\,[\mathbf{A}, \mathbf{b}] = r$，则 $\mathbf{b}$ 可以由 $\mathbf{A}$ 的前 r 列表示，因此，存在向量 $\mathbf{x}$，使得 $\mathbf{Ax} = \mathbf{b}$

2.3.3 线性方程组的求解

- **定理 2.2:** 对于方程组 $\mathbf{Ax = b}$，其中 $A \in R^{m\times n}$，$\text{rank}\,A = m$，则方程组的解可以通过给定 $n - m$ 个变量的任意值，然后求解剩余的 m 个变量得到。
  - 证明：
    - 可将方程组表示为 $x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \ldots + x_m\mathbf{a}_m = \mathbf{b} - x_{m+1}\mathbf{a}_{m+1} - \ldots - x_n\mathbf{a}_n$
    - 将 $x_{m+1}, \ldots, x_n$ 设为任意值 $d_{m+1}, \ldots, d_n$
    - 构造矩阵 $B = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$， 由于 $\text{rank}\,A = m$，因此 B 是可逆矩阵，此时方程组为 $Bx = \mathbf{b} - d_{m+1}\mathbf{a}_{m+1} - \ldots - d_n\mathbf{a}_n$
    - 解得 $x = B^{-1}(\mathbf{b} - d_{m+1}\mathbf{a}_{m+1} - \ldots - d_n\mathbf{a}_n)$

2.4 内积与范数

2.4.1 绝对值

- **定义:** $|a|$ 表示实数 a 的绝对值
- **性质:**
  1. $|a| = |-a|$
  2. $-|a| \leqslant a \leqslant |a|$
  3. $|a + b| \leqslant |a| + |b|$
  4. $||a| - |b|| \leqslant |a - b| \leqslant |a| + |b|$

5. $|ab| = |a||b|$
6. $|a| < c$ and $|b| < d$ imply that $|a + b| < c + d$
7. $|a| < b$ is equivalent to $-b < a < b$ (i.e., $a < b$ and $-a < b$)。此不等式对于 "<=" 也适用

2.4.2 欧几里得内积

- **欧几里得内积 (Euclidean inner product)** 对于两个 R^n 中的向量 x 和 y 定义为:
  - $(x, y) = \sum_i x_i y_i = \mathbf{x}^T\mathbf{y}$
- **内积的性质:**
  1. **正定性 (Positivity):** $(x, x) > 0$, 当且仅当 $x = 0$ 时 $(x, x) = 0$
  2. **对称性 (Symmetry):** $(x, y) = (y, x)$
  3. **加法性 (Additivity):** $(x + y, z) = (x, z) + (y, z)$
  4. **齐性 (Homogeneity):** $(rx, y) = r(x, y)$, 其中 $r \in R$
  - 第二向量满足加法性和齐性:
  - $(x, y + z) = (x, y) + (x, z)$
  - $(x, ry) = r(x, y)$, 其中 $r \in R$
- **正交 (Orthogonal):** 如果 $(x, y) = 0$, 则称向量 x 和 y 正交。

2.4.3 欧几里得范数

- **欧几里得范数 (Euclidean norm)** 向量 x 的欧几里得范数定义为:
  - $||x|| = \sqrt{<x, x>} = \sqrt{(x^T x)}$
- **范数的性质:**
  1. **正定性 (Positivity):** $||x|| > 0$, 当且仅当 $x = 0$ 时 $||x|| = 0$
  2. **齐性 (Homogeneity):** $||rx|| = |r|\cdot ||x||$, 其中 $r \in R$
  3. **三角不等式 (Triangle inequality):** $||x + y|| \leqslant ||x|| + ||y||$

2.4.4 柯西-施瓦茨不等式 (Cauchy-Schwarz Inequality)

- 对于 R^n 中的任意两个向量 x 和 y, 有: $|(x, y)| \leqslant ||x|| \cdot ||y||$。等号成立当且仅当 $x = \alpha y$, 其中 $\alpha \in R$。

2.4.5 p-范数 (p-norm)

- 对于向量 $x \in R^n$, 定义 p-范数为:
- $||x||_p = (|x_1|^p + \dots + |x_n|^p)^{1/p}$, 当 $1 \leqslant p < \infty$
- $||x||_p = \max\{|x_1|, \dots, |x_n|\}$, 当 $p = \infty$
- **p 范数可用于描述连续函数**

2.4.6 复向量空间的内积

- 对于复数向量空间 C^n, 内积定义为 $(x, y) = \sum_i x_i \overline{y_i}$
- 复数空间的内积性质:
  1. **正定性**: $(x, x) > 0$, 当且仅当 $x = 0$ 时 $(x, x) = 0$
  2. **对称性**: $(x, y) = (y, x)$ 的共轭
  3. **加法性**: $(x + y, z) = (x, z) + (y, z)$
  4. **齐性 (齐性):** $(rx, y) = r(x, y)$ for every $r \in C$
  - 对于第二个变量满足: $(x, r_1 y + r_2 z) = \overline{r_1}(x, y) + \overline{r_2}(x, z)$

本章小结

本章深入探讨了向量空间和矩阵的基础知识，包括向量的运算、矩阵的表示和秩、线性方程组的解、以及内积和范数等重要概念。这些概念和方法为进一步学习高级工程数学奠定了基础，并为解决实际工程问题提供了数学工具。

第三章 线性变换、特征值与特征向量及正交投影

本章导言：

本章将继续深入探讨线性代数的核心概念，主要围绕线性变换、特征值与特征向量以及正交投影展开讨论。线性变换是向量空间之间保持线性关系的映射，它可以通过矩阵来表示。特征值和特征向量则揭示了线性变换的内在结构，它们描述了在变换下保持方向不变的向量及其对应的缩放因子。最后，正交投影则是一种特殊的线性变换，它将向量投影到子空间上，保持垂直性，是解决线性方程组、优化问题和信号处理等领域的重要工具。通过本章的学习，读者将对线性变换的几何意义、特征分解的概念以及投影变换的性质有深刻的理解，并能够灵活运用这些工具解决实际问题。

3.1 线性变换

3.1.1 线性变换的定义

- **线性变换 (Linear transformation):** 从向量空间 R^n 到向量空间 R^m 的一个映射 $L: R^n \to R^m$，如果满足以下两个条件，则称为线性变换：

  1. **齐次性 (Homogeneity):** $L(\alpha\mathbf{x}) = \alpha L(\mathbf{x})$，对于任意 $\mathbf{x} \in R^n$，任意标量 $\alpha \in R$ 成立。
  2. **可加性 (Additivity):** $L(\mathbf{x}_1 + \mathbf{x}_2) = L(\mathbf{x}_1) + L(\mathbf{x}_2)$，对于任意 $\mathbf{x}_1, \mathbf{x}_2 \in R^n$ 成立。

3.1.2 线性变换的矩阵表示

- 如果确定了 R^n 和 R^m 的基，那么线性变换 L 可以用一个矩阵表示。
- 若 $\mathbf{x} \in R^n$， $\mathbf{x}'$ 为 $\mathbf{x}$ 在 R^n 的给定基下的表示； $\mathbf{y} = L(\mathbf{x})$， $\mathbf{y}'$为 $\mathbf{y}$ 在 R^m 的给定基下的表示。则存在一个矩阵 $A \in R^{m \times n}$，使得 $y' = Ax'$。此时称 A 为线性变换 L 的矩阵表示。
- 特别地，当假设 R^n 和 R^m 的基均为标准基时， 则有 $L(\mathbf{x}) = \mathbf{Ax}$

3.1.3 变换矩阵与基的关系

- 设 $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\}$ 和 $\{\mathbf{e}'_1, \mathbf{e}'_2, ..., \mathbf{e}'_n\}$ 是 R^n 的两组基，定义矩阵：
  - $\mathbf{T} = [\mathbf{e}'_1, \mathbf{e}'_2, ..., \mathbf{e}'_n]^{-1} [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n]$
- 称 $\mathbf{T}$ 为从基 $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\}$ 到基 $\{\mathbf{e}'_1, \mathbf{e}'_2, ..., \mathbf{e}'_n\}$ 的变换矩阵。
- 显然有： $[\mathbf{e}'_1, \mathbf{e}'_2, ..., \mathbf{e}'_n]\, \mathbf{T} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n]$
- $\mathbf{T}$ 的第 i 列就是向量 e_i 在基 $\{\mathbf{e}'_1, \mathbf{e}'_2, ..., \mathbf{e}'_n\}$ 下的坐标。
- 对于任意向量 $\mathbf{v}$，设 $\mathbf{x}$ 为其在 $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ 下的坐标， $\mathbf{x}'$ 为其在 $\{\mathbf{e}'_1, ..., \mathbf{e}'_n\}$ 下的坐

标，则有： $\mathbf{x} = \mathbf{T}\mathbf{x}'$。

- **过渡矩阵 (Transition matrix):** 若（$\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots, \boldsymbol{\eta}_n$) = ($\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \cdots, \boldsymbol{\varepsilon}_n$) C， 则称矩阵 C 为由基 $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \cdots, \boldsymbol{\varepsilon}_n$ 到基 $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots, \boldsymbol{\eta}_n$ 的过渡矩阵。

3.2 特征值与特征向量

3.2.1 特征值与特征向量的定义

- **特征值 (Eigenvalue):** 对于 $n \times n$ 的实数方阵 A，如果存在一个标量 λ (可以是复数) 和一个非零向量 **v**，使得 $\mathbf{Av} = \lambda\mathbf{v}$， 则称 λ 为矩阵 A 的一个特征值。
- **特征向量 (Eigenvector):** 满足 $\mathbf{Av} = \lambda\mathbf{v}$ 的非零向量 **v** 为矩阵 A 的属于特征值 λ 的一个特征向量。

3.2.2 特征方程

- 要使 λ 为 A 的特征值， 必须并且只须矩阵 λI - A 为奇异矩阵， 即 $\det[\lambda I - A] = 0$，其中 I 是 $n \times n$ 的单位矩阵。
- $\det[\lambda I - A] = 0$ 是关于 λ 的一个 n 次多项式方程，称为 **特征方程 (characteristic equation)**。$\det[\lambda I - A]$称为矩阵 A 的 **特征多项式 (characteristic polynomial)** 。
- n 阶特征方程有 n 个复数根（可能相同）。
- 若有 n 个不同的特征根，则有 n 个线性无关的特征向量。

3.2.3 实对称矩阵的特征值与特征向量

- **定理 3.2：** 实对称矩阵($A=A^T$)的所有特征值都是实的。
  - **证明:** 若 $\mathbf{Ax} = \lambda\mathbf{x}, \mathbf{x} \neq 0$，则有（$\mathbf{Ax}, \mathbf{x}$）= $\lambda<\mathbf{x}, \mathbf{x}>$。另一方面,（$\mathbf{Ax}, \mathbf{x}$）= $(\mathbf{x}, \mathbf{A}^T\mathbf{x}) = (\mathbf{x}, \mathbf{Ax}) = \bar{\lambda}(\mathbf{x}, \mathbf{x})$ 。 由于（$\mathbf{x}, \mathbf{x}$）是实数并且>0，因此 $\lambda = \bar{\lambda}$，即 λ 是实的。
- **定理 3.3：** 任意 $n \times n$ 实对称矩阵具有 n 个相互正交的特征向量.
- **证明**: 此处只证 n 个特征值不同的情形。
  - 设 $\mathbf{Av}_1 = \lambda_1\mathbf{v}_1, \mathbf{Av}_2 = \lambda_2\mathbf{v}_2$，其中 $\lambda_1 \neq \lambda_2$，则有： $<\mathbf{Av}_1, \mathbf{v}_2> = \lambda_1<\mathbf{v}_1, \mathbf{v}_2>$。
  - 由于 $A = A^T$，则 $<\mathbf{Av}_1, \mathbf{v}_2> = <\mathbf{v}_1, A^T\mathbf{v}_2> = <\mathbf{v}_1, \mathbf{Av}_2> = \lambda_2<\mathbf{v}_1, \mathbf{v}_2>$
  - 因此， $\lambda_1<\mathbf{v}_1, \mathbf{v}_2> = \lambda_2<\mathbf{v}_1, \mathbf{v}_2>$ ，由于 $\lambda_1 \neq \lambda_2$，则 $<\mathbf{v}_1, \mathbf{v}_2> = 0$。 * 当特征值 $\lambda_1 \neq \lambda_2$时， 对应的特征向量是正交的。

3.3 正交投影

3.3.1 子空间和正交补

- **子空间 (Subspace):** 向量空间 R^n 的一个子集 V，如果满足以下条件，则称为 R^n 的一个子空间：
  - 若 $\mathbf{x}_1, \mathbf{x}_2 \in V$， 则 $\alpha\mathbf{x}_1 + \beta\mathbf{x}_2 \in V$，对于任意 $\alpha, \beta \in R$ 成立
- **子空间的维度 (Dimension):** 一个子空间 V 的维度等于子空间中线性无关向量的最大个数。
- **正交补 (Orthogonal complement):** 对于 R^n 的一个子空间 V，其正交补记为 $V^\perp$，$V^\perp$ 由所有与 V 中所有向量正交的向量构成。即 $V^\perp = \{\mathbf{x}: \mathbf{v}^T\mathbf{x} = 0 \text{ for all } \mathbf{v} \in V\}$
  - V 和 $V^\perp$ 张成 R^n， 即 $\forall\mathbf{x} \in R^n$ 可唯一表示为 $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$，其中 $\mathbf{x}_1 \in V$，$\mathbf{x}_2 \in V^\perp$。称为正交分解。
- **正交投影 (Orthogonal projection):** $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 分别是向量 **x** 在子空间 V 和 $V^\perp$ 上

的正交投影。

- $R^n = V \oplus V^{\perp}$ (直和)

3.3.2 正交投影的定义

- **正交投影矩阵 (Orthogonal projector):** 线性变换 P 是一个到子空间 $V = R(P)$ 的正交投影矩阵 如果对于任意向量 $\mathbf{x} \in R^n$，有 $P\mathbf{x} \in V$。
- **定理 3.5：** 矩阵 P 是正交投影矩阵 当且仅当 $P^2 = P = P^T$
  - **证明:**
    - 利用 $x = Px + (x - Px)$
    - $R(P)^{\perp} = N(P^T)$
    - => 如果 P 是正交投影， 则 $R(I-P) \subseteq R(P)^{\perp} = N(P^T)$. 所以 $P^T(I-P) = O$, $P^T = P^TP$。即 $P=P^T=P^2$ * <= 如果 $P=P^T=P^2$，对于任意 $\mathbf{x}$, $(P\mathbf{y})^T(I - P)\mathbf{x} = \mathbf{y}^T P^T (I - P)\mathbf{x} = \mathbf{y}^T P(I - P)\mathbf{x} = 0$ 。 因此，$(I-P)\mathbf{x} \in R(P)^{\perp}$.

本章小结

本章介绍了线性变换、特征值与特征向量以及正交投影等概念，这些概念不仅是线性代数的核心内容，也是其他数学领域和工程应用的重要基础。理解这些概念将有助于读者更好地分析和解决实际问题。

第四章 几何概念：线段、超平面与凸集

本章导言:

本章将介绍一些基本的几何概念，这些概念在优化问题、机器学习等领域有着广泛的应用。我们将首先定义线段的概念，这是构建更复杂几何结构的基石。然后，我们将深入探讨超平面和线性簇的概念，这它们是高维空间中线性关系的重要体现，也是许多优化问题的约束条件。最后，我们将介绍凸集及其相关性质，凸集是一类重要的几何对象，在优化理论中具有重要的地位，因为局部最优解通常也是全局最优解。理解这些几何概念，将有助于读者从几何角度认识和分析实际问题，从而更好地理解和应用相关的数学工具。

4.1 线段

- **线段的定义:** 在 R^n 空间中，连接两个点 $\mathbf{x}$ 和 $\mathbf{y}$ 的线段，是位于连接这两点的直线上的点的集合。
  - 如果点 $\mathbf{z}$ 在 $\mathbf{x}$ 和 $\mathbf{y}$ 之间的线段上，则 $\mathbf{z} - \mathbf{y} = \alpha(\mathbf{x} - \mathbf{y})$，其中 $\alpha \in [0, 1]$。
  - 也可以写为 $\mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{y}$，其中 $\alpha \in [0, 1]$。
  - 线段可以用集合表示为：$\{ \alpha\mathbf{x} + (1 - \alpha)\mathbf{y} : \alpha \in [0,1] \}$

4.2 超平面与线性簇

4.2.1 超平面的定义

- **超平面 (Hyperplane):** 在 R^n 空间中，超平面是由满足线性方程的所有点组成的集

合。

- ○ 方程形式： $u_1x_1 + u_2x_2 + ... + u_nx_n = v$，其中 $u_1, u_2,..., u_n$ 为实数，至少一个非零，$\mathbf{x} = [x_1, x_2, ..., x_n]^T \in R^n$，v 为实数。
- ○ 向量形式： $\{ \mathbf{x} \in R^n: \mathbf{u}^T\mathbf{x} = v\}$

- **法向量 (Normal):** 向量 **u** 称为超平面的法向量，它与超平面内任意两个向量的差，也即超平面内任意向量正交。
- 如果 **a** 是超平面上的任意一点，则 $\mathbf{u}^T(\mathbf{x} - \mathbf{a}) = 0$，也可以理解为 **u** 和 **x - a** 是相互正交的。
  - ○ **超平面的法向量**： 对于超平面 $\mathbf{u}^T \mathbf{x} = v$, **u** 是法向量，代表了超平面方向，而 v 控制了超平面与原点的距离。
- **半空间 (Half-space):** 超平面将空间分为两个半空间。
  - ○ **正半空间 (Positive half-space):** 由满足不等式 $u_1x_1 + ... + u_nx_n \geqslant v$ 的点组成，表示为 $H_+ = \{\mathbf{x} \in R^n: \mathbf{u}^T\mathbf{x} \geqslant v\}$
  - ○ **负半空间 (Negative half-space):** 由满足不等式 $u_1x_1 + ... + u_nx_n < v$ 的点组成，表示为 $H_- = \{\mathbf{x} \in R^n: \mathbf{u}^T\mathbf{x} < v\}$

4.2.2 线性簇

- **线性簇 (Linear variety):** 由满足线性方程组 **Ax = b** 的所有点构成的集合。表示为：
  - ○ $\{ \mathbf{x} \in R^n: \mathbf{Ax = b} \}$
  - ○ 其中 $A \in R^{m\times n}$，$b \in R^m$。
- **线性簇的维度 (Dimension):** 如果 $\dim N(A) = r$，我们称该线性簇的维度为 r。
- 当 $\mathbf{b} = 0$ 时，线性簇是一个 **子空间**
- 当 $A = O$ 时，线性簇是 R^n。
- 维度小于 n 的线性簇可以被看作有限个超平面的交集。

4.3 凸集

4.3.1 凸集的定义

- **凸组合 (Convex combination):** 两个点 **u** 和 **v** 的凸组合是 $w = \alpha\mathbf{u} + (1 - \alpha)\mathbf{v}$，其中 $\alpha \in [0,1]$。
- **凸集 (Convex set):** 如果对于任意 $\mathbf{u}, \mathbf{v} \in \Theta$，连接 **u** 和 **v** 的线段上的所有点都属于 Θ，则称集合 Θ 是凸集。即若 $\alpha\mathbf{u} + (1-\alpha)\mathbf{v} \in \Theta$ 对于 $\forall \mathbf{u}, \mathbf{v} \in \Theta$，$\forall\alpha \in [0,1]$ 都成立，则称 Θ 为凸集。

4.3.2 凸集的性质

- **定理 4.1：** 凸集具有以下性质：
  1. 如果 Θ 是凸集，且 β 是实数，则 $\beta\Theta = \{ \mathbf{x}: \mathbf{x} = \beta\mathbf{v}, \mathbf{v} \in \Theta \}$也是凸集。
  2. 如果 Θ_1 和 Θ_2 是凸集，则 $\Theta_1 + \Theta_2 = \{ \mathbf{x}: \mathbf{x} = \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_1 \in \Theta_1, \mathbf{v}_2 \in \Theta_2 \}$ 也是凸集。
  3. 任何凸集的集合的交集仍然是凸集。

4.3.3 凸集的例子

- 空集是凸集。
- 只包含一个点的集合是凸集。
- 直线或线段是凸集。
- 子空间是凸集。
- 超平面是凸集。
- 线性簇是凸集。
- 半空间是凸集。
- R^n 是凸集。

4.3.4 凸集的极点

- **极点 (Extreme point):** 凸集 Θ 中的一个点 **x** 如果无法被表示为 Θ 中其他两个不同点的凸组合，则称 **x** 为极点。
  - 例如：圆的边界上的点是极点， 凸多边形的顶点是极点。

4.4 邻域

- **邻域 (Neighborhood):** 点 **x** ∈ R^n 的邻域是指以 **x** 为中心，半径为 ε 的球形区域。
  - 数学表示：{**y** ∈ R^n : ||**y** - **x**|| < ε}，其中 ε 为正实数。
  - 在二维平面 R^2 上，一个邻域是以 **x** 为圆心的圆盘。
  - 在三维空间 R^3 中，一个邻域是以 **x** 为球心的球体。
- **内点 (Interior point):** 如果一个点 **x** ∈ S ， 并且集合 S 包含 **x** 的一个邻域，则称 **x** 为 S 的一个内点。换句话说，S 中的每一个点都是其内点，就称 S 是开集 (open set)。
  - 开集不包含任何边界点。
- **边界点 (Boundary point):** 一个点 **x** 是集合 S 的边界点，如果 **x** 的每个邻域都包含 S 中的点，也包含 S 外的点。
  - 边界点可能属于 S，也可能不属于 S。
  - 所有边界点的集合构成了集合 S 的边界。
- **闭集 (Closed set):** 如果一个集合包含其所有边界点，则称该集合是闭集。
  - 一个集合是闭集当且仅当其补集是开集。
- **有界集 (Bounded set):** 如果一个集合能够包含在一个有限半径的球内，则称该集合是有界的。
- **紧集 (Compact set):** 如果一个集合是闭集且是有界的，则该集合称为紧集。
  - 紧集在优化问题中非常重要，因为可以保证最值点的存在，例如： 连续函数在紧集上必定有最大值和最小值 (Weierstrass 定理)。
    - **定理 4.2 维尔斯特拉斯定理 (Theorem of Weierstrass):** 对于一个连续函数 $f: \Omega \to R$，其中 $\Omega \subset R^n$ 是一个紧集，则必定存在一个点 $x_0 \in \Omega$，使得 $f(x_0) \leq f(x)$ 对所有 $x \in \Omega$ 成立，即函数 f 在 Ω 上可以取得最小值。

4.5 多面体与多胞体

- **凸多胞体 (Convex polytope):** 可以表示成有限个半空间的交集，称该集合为凸多胞体。
- **多面体 (Polyhedron):** 有界的凸多胞体称为多面体。

本章小结

本章介绍了线段、超平面、凸集以及邻域等基本的几何概念。这些概念为后续讨论优化问题和机器学习算法奠定了基础。理解这些几何结构能够帮助读者从几何的角度理解数学，从而为解决复杂问题提供新的视角。

第五章 微积分基础：序列、极限与微分

本章导言：

本章将介绍微积分的基本概念，为后续的优化算法和理论提供必要的工具。我们将从序列及其极限的概念入手，定义单调序列、有界序列和收敛序列。然后，我们将详细讲解函数的可微性，包括导数矩阵的概念和求导法则。最后，我们将引入水平集和梯度，以及泰勒级数等重要概念，这些工具在优化问题的分析和求解中扮演着关键角色。通过本章的学习，读者将掌握微积分的基本原理和运算方法，为深入学习优化理论奠定坚实的数学基础。

5.1 序列与极限

5.1.1 实数序列的定义与分类

- **实数序列 (Sequence of real numbers):** 定义域为自然数集 N (1, 2, 3, ⋯)，值域包含在实数集 R 的函数。可表示为 $\{x_1, x_2, x_3, \ldots\}$ 或者 $\{x_k\}$ 或 $\{x_k\}_{k=1}^{\infty}$
- **递增序列 (Increasing sequence):** 对于所有 k，有 $x_k < x_{k+1}$。
- **非递减序列 (Nondecreasing sequence):** 对于所有 k，有 $x_k \leqslant x_{k+1}$。 * **递减序列 (Decreasing sequence):** 对于所有 k，有 $x_k > x_{k+1}$ * **非递增序列 (Nonincreasing sequence):** 对于所有 k，有 $x_k \geqslant x_{k+1}$ * **单调序列 (Monotone sequences):** 非递增或非递减的序列统称为单调序列。

5.1.2 实数序列的极限

- **极限 (Limit):** 实数序列 $\{x_k\}$ 的极限是一个数 x，*如果对于任意 $\varepsilon > 0$，存在一个整数 K(可能取决于 ε)，使得当 $k > K$ 时，有 $|x_k - x| < \varepsilon$*。 * 可以表示为：$x^* = \lim_{k\to\infty} x_k$ 或者 $x_k \to x^*$

5.1.3 高维空间序列及其极限

- **高维空间序列**: R^n 中的序列可以用 $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\}$ 或者 $\{\mathbf{x}^{(k)}\}$ 表示，其中 $\mathbf{x}^{(k)} \in R^n$。
- **高维空间序列的极限:** 对于 R^n 中的序列 $\{\mathbf{x}^{(k)}\}$，如果对于任意 $\varepsilon > 0$，存在一个整数 K(可能取决于 ε)，使得当 $k > K$ 时，有 $\|\mathbf{x}^{(k)} - \mathbf{x}^*\| < \varepsilon$，则称 $\mathbf{x}^*$ 为序列 $\{\mathbf{x}^{(k)}\}$ 的极限。
- 可以表示为: $\mathbf{x}^* = \lim_{k\to\infty} \mathbf{x}^{(k)}$ 或者 $\mathbf{x}^{(k)} \to \mathbf{x}^*$

5.1.4 收敛序列的性质

- **唯一性定理 (Theorem 5.1):** 如果一个序列收敛，它的极限是唯一的。

- **有界性定理 (Theorem 5.2):** 每一个收敛序列都是有界的。
- **子序列收敛性定理 (Theorem 5.4):** 如果一个序列收敛到某个极限，那么该序列的任何子序列也都收敛到相同的极限。
- **单调有界收敛定理 (Theorem 5.3):** R 中的每个单调有界序列都是收敛的。

5.1.5 连续函数与极限

- **函数在一点的连续性:** 如果对于任意收敛到 x_0 的序列$\{x^{(k)}\}$，都有 $\lim_{k\to\infty} f(x^{(k)}) = f(x_0)$，则稱函数 f 在 x_0 处是连续的。
- 如果函数 f 在 x_0 处连续，我们可以使用 $\lim_{x\to x_0} f(x) = f(x_0)$来表示。

5.1.6 矩阵序列的极限

- **矩阵序列的极限:** 如果一个 m × n 矩阵的序列 $\{\mathbf{A}_k\}$ 的所有元素组成的实数序列都收敛到对应位置的矩阵 **A** 的元素， 则称 矩阵序列 $\{\mathbf{A}_k\}$ 收敛到矩阵 **A**。 * 数学表示为: $\lim_{k\to\infty} \|\mathbf{A} - \mathbf{A}_k\| = 0$
- **引理 5.1:** 对于任意 $A \in R^{n\times n}$, $\lim_{k\to\infty} \mathbf{A}^k = 0$ 当且仅当 A 的特征值的绝对值都小于 1。

5.2 可微性

5.2.1 仿射函数

- **仿射函数 (Affine function):** 从 R^n 到 R^m 的一个函数 $\mathbf{A}: R^n \to R^m$，如果存在一个线性变换 $L: R^n \to R^m$ 以及一个向量 $\mathbf{y} \in R^m$，使得对于所有 $\mathbf{x} \in R^n$，有 $\mathbf{A}(\mathbf{x}) = L(\mathbf{x}) + \mathbf{y}$， 则称 **A** 为仿射函数。

5.2.2 可微性的定义

- 若要用仿射函数在 x_0 附近逼近函数 $f: R^n \to R^m$，则需要保证:
  1. $\mathbf{A}(\mathbf{x_0}) = f(\mathbf{x_0})$
  2. 当 **x** 接近 $\mathbf{x_0}$ 时，$\mathbf{A}(\mathbf{x})$ 趋近 $f(\mathbf{x})$ 的速度要快于 **x** 趋近 $\mathbf{x_0}$ 的速度。* 数学上，就是让误差 $\| f(\mathbf{x}) - \mathbf{A}(\mathbf{x})\|$ 与 $\|\mathbf{x}-\mathbf{x_0}\|$ 的比值在 **x** 趋近 $\mathbf{x_0}$ 时趋于 0 。
  - 即 $\lim_{x\to x_0} \| f(x) - A(x)\|/\|x - x_0\| = 0$

5.3 导数矩阵

5.3.1 导数矩阵的定义

- 对于一个可微函数 $f: R^n \to R^m$，在给定点 x_0，其导数可以用一个 m × n 的矩阵来表示，记作 $Df(x_0)$，称为**导数矩阵 (Derivative matrix)**。
- 将 $f : R^n \to R^m$ 看作 $f = [f_1, f_2, \ldots, fm]^T$，则导数矩阵 $Df(x_0)$ 的第 j 列是 $\mathbf{L}\,\mathbf{e}_i$，其中 $\mathbf{e}_i$ 为 R^n 的标准基的第 i 列。也就是函数在 x_0 处沿着第 j 个坐标轴方向的偏导数向量。
  - 数学表示:
    - $Df(x_0)$ 的第 j 列 $= \lim_{t\to 0} [f(x_0 + t\mathbf{e}_j) - f(x_0)]/t$。
    - $Df(x_0) = [\partial f/\partial x_1, \partial f/\partial x_2, \ldots, \partial f/\partial x_n]$

- **梯度的定义:** 当 $f: R^n \to R$ 时，梯度记为 $\nabla f(x)$，是导数矩阵 $Df(x)$的转置。 * $\nabla f(x) = [\partial f/\partial x_1, \partial f/\partial x_2, \ldots, \partial f/\partial x_n]^T$

5.3.2 导数矩阵与函数的关系

- 一个可微函数 $f: R^n \to R^m$ 在 x_0 点的最佳仿射逼近为： $A(x) = f(x_0) + Df(x_0)(x - x_0)$
- 一个函数在某点可微，该点的导数矩阵是唯一的。
- 导数矩阵 $Df(x_0)$ 的列是向量偏导数。* 向量 $\partial f/\partial x_j(x_0)$ 是在点 x_0 处，沿着第 j 个坐标轴方向的切向量。

5.3.3 海森矩阵

- **二阶可微函数：** 如果函数 $f: R^n \to R$ 的梯度 ∇f 是可微的，则称 f 为二阶可微的。
- **海森矩阵 (Hessian Matrix):** 二阶可微函数 f 在 x 处的二阶导数用一个 $n \times n$ 的矩阵表示，称为海森矩阵，记为 $D^2f(x)$ 或 $F(x)$ * $D^2 f(x) = [\partial^2 f/(\partial x_i \partial x_j)]$。 * ($\partial^2 f/\partial x_i \partial x_j$ 是 $\partial f/\partial x_j$ 关于 x_i 的偏导数)
- 如果函数 f 在 x 处二阶连续可微，则海森矩阵是对称的，满足 Clairaut 定理（也叫 Schwarz 定理），即 $\partial^2 f/\partial x_i \partial x_j = \partial^2 f/\partial x_j \partial x_i$
- 当二阶偏导数不连续时，海森矩阵不一定是对称的。

5.4 求导法则

5.4.1 链式法则（Chain rule）

- 对于函数 $f: R \to R^n$ 和 $g: R^n \to R$，复合函数 $h(t) = g(f(t))$ 的导数 $h'(t) = \nabla g(f(t)) \cdot f'(t)$
- 更一般地，$h'(t) = Dg(f(t))\, Df(t)$

5.4.2 乘积法则

- 对于可微函数 $f: R^n \to R^m$ 和 $g: R^n \to R^m$， 定义 $h: R^n \to R$ 为 $h(\mathbf{x}) = f(\mathbf{x})^T g(\mathbf{x})$，则 $Dh(x) = f(x)^T Dg(x) + g(x)^T Df(x)$。

5.4.3 常用导数公式

- $D(y^T Ax) = y^T A$，其中 $y \in R^m$， $A \in R^{m\times n}$
- $D(x^T Ax) = x^T (A + A^T)$，如果 $m = n$
- $D(x^T y) = y^T$，其中 $y \in R^n$， 如果 y 与 x 无关
- $D(x^T Qx) = 2x^T Q$，如果 Q 是对称矩阵
- $D(x^T x) = 2x^T$

5.5 水平集与梯度

5.5.1 水平集

- **水平集 (Level set):** 函数 $f: R^n \rightarrow R$ 的水平集是指函数值等于某个常数的点的集合。
  - 即 $S = \{ \mathbf{x} : f(\mathbf{x}) = c\}$。
  - 当 $f : R^2 \rightarrow R$ 时，水平集 S 通常是一条曲线。
  - 当 $f : R^3 \rightarrow R$ 时，水平集 S 通常是一个曲面。

5.5.2 梯度与水平集的关系

- 如果存在一条曲线 y 位于 水平集 S 上，且参数化表示为函数 $g : R \rightarrow R^n$， 那么 $f(g(t))=c$。若 $g(t_0) = x_0$，且 $g'(t_0) = v$， 则切向量 v 应该与梯度 $\nabla f(x_0)$ 正交。
- **定理：** 梯度 $\nabla f(x_0)$ 正交于过点 x_0 的水平集的切线方向。 * 梯度 $\nabla f(\mathbf{x})$ 指向函数值增长最快的方向。
- 梯度 $-\nabla f(\mathbf{x})$ 指向函数值下降最快的方向，即最速下降方向。

5.6 泰勒级数

- **泰勒定理 (Taylor's Theorem):** 如果函数 $f: R \rightarrow R$ 在区间 [a, b] 上 m 次连续可微，则： * $f(b) = f(a) + f'(a)(b-a)/1! + f''(a)(b-a)^2/2! + \ldots + f^{(m-1)}(a)(b-a)^{m-1}/(m-1)! + Rm$ * 余项 $Rm = f^{(m)}(a+\theta h)\, h^m (1-\theta)^{m-1} / (m-1)!$，或 $Rm = f^{(m)}(a+\theta' h)\, h^m / m!$ * 其中 $h = b - a$, $\theta, \theta' \in (0,1)$ 。
- 泰勒定理也可以推广到多元函数 * 对于多元函数 f，其在 x_0 处的泰勒展开式为:
  - $f(x) = f(x_0) + \nabla f(x_0)^T (x-x_0) + (x-x_0)^T D^2 f(x_0) (x-x_0) /2! + o(||x-x_0||^2)$
- **中值定理:** 如果 $f: R^n \rightarrow R^m$ 在开集 Ω 内可微，那么对于 Ω 中任意两点 x 和 y，存在矩阵 **M**， 使得 $f(x) - f(y) = \mathbf{M}(x-y)$
- **M** 的每一行，均为 Df 在联结 x 和 y 路径上的点的导数。

本章小结

本章介绍了微积分的基本概念，包括序列及其极限，函数的可微性、导数矩阵、梯度、水平集和泰勒展开等。这些概念是优化算法的基础，也是理解许多高级数学概念的基石。

第六章 约束优化问题的最优性条件

本章导言：

本章将深入探讨约束优化问题的最优性条件，这是优化理论的核心内容。我们将首先介绍约束优化问题的基本概念和分类，然后重点讨论一阶必要条件（FONC）和二阶必要条件（SONC），以及二阶充分条件（SOSC）。这些条件为判断一个点是否为局部最优解提供了理论依据。在此基础上，我们将详细介绍拉格朗日乘子法，这是一种将约束优化问题转化为无约束优化问题的重要方法。通过本章的学习，读者将掌握判断约束优化问题最优解的条件和方法，并能够运用拉格朗日乘子法解决实际的优化问题。

6.1 约束优化问题概述

6.1.1 约束优化问题的定义

- **约束优化问题 (Constrained optimization problem):** 在一个给定的集合 Ω (称为可行域) 中找到一个点 $\mathbf{x}^*$，使得目标函数 $f(\mathbf{x})$ 取得最小值或最大值。

- 数学表示：
- minimize f(x)
- subject to x ∈ Ω

其中 $f: R^n \to R$ 是目标函数，$\Omega \subseteq R^n$ 是可行域。

- **可行域 (Feasible region):** 满足所有约束条件的点的集合。
- **局部极小点 (Local minimizer):** 如果存在 **x*** 的一个邻域 N，使得对于所有 **x** ∈ N ∩ Ω 且 **x** ≠ **x***，都有 f(**x**) ≥ f(**x***)，则称 **x*** 为一个局部极小点。
- **严格局部极小点 (Strict local minimizer):** 如果存在 **x*** 的一个邻域 N，使得对于所有 **x** ∈ N ∩ Ω 且 **x** ≠ **x***，都有 f(**x**) > f(**x***)，则称 **x*** 为一个严格局部极小点。
- **全局极小点 (Global minimizer):** 如果对于所有 **x** ∈ Ω，都有 f(**x**) ≥ f(**x***)，则称 **x*** 为一个全局极小点。

6.1.2 约束优化问题的分类

- **等式约束 (Equality constraints):** 约束条件为等式形式，例如 $h_i(\mathbf{x}) = 0$。
- **不等式约束 (Inequality constraints):** 约束条件为不等式形式，例如 $g_i(\mathbf{x}) \leq 0$。
- **线性约束 (Linear constraints):** 约束函数为线性函数。
- **非线性约束 (Nonlinear constraints):** 约束函数为非线性函数。

6.2 最优性条件

6.2.1 一阶必要条件 (First-Order Necessary Condition, FONC)

- **可行方向 (Feasible direction):** 在可行域 Ω 中的点 **x** 处，如果存在一个向量 **d** 和一个正数 α_0，使得对于所有 $\alpha \in [0, \alpha_0]$，都有 $\mathbf{x} + \alpha\mathbf{d} \in \Omega$，则称 **d** 为在 **x** 处的一个可行方向。
- **定理 6.1 (FONC):** 如果 **x*** 是约束优化问题的一个局部极小点，且 **d** 是 **x*** 处的一个可行方向，则 $\nabla f(\mathbf{x}^*)^T\mathbf{d} \geq 0$。
  - **证明:** 类似于无约束情况，利用泰勒展开和反证法。
    - 定义 $\varphi(\alpha) = f(\mathbf{x}^* + \alpha\mathbf{d})$，其中 **d** 是可行方向。
    - 将 $\varphi(\alpha)$ 在 $\alpha = 0$ 处进行泰勒展开：$f(\mathbf{x}^* + \alpha\mathbf{d}) = f(\mathbf{x}^*) + \alpha\nabla f(\mathbf{x}^*)^T\mathbf{d} + o(\alpha)$
    - 如果 $\nabla f(\mathbf{x}^*)^T\mathbf{d} < 0$，则对于足够小的 $\alpha > 0$，有 $f(\mathbf{x}^* + \alpha\mathbf{d}) < f(\mathbf{x}^*)$，与 **x*** 是局部极小点矛盾。
- **推论:** 如果 **x*** 是可行域内部的一个局部极小点，则 $\nabla f(\mathbf{x}^*) = 0$。
  - 因为此时任意方向都是可行方向，可以取 $\mathbf{d} = -\nabla f(\mathbf{x}^*)$。

6.2.2 二阶必要条件 (Second-Order Necessary Condition, SONC)

- **定理 6.2 (SONC):** 如果 **x*** 是约束优化问题的一个局部极小点，**d** 是 **x*** 处的一个可行方向，且 $\nabla f(\mathbf{x}^*)^T\mathbf{d} = 0$，则 $\mathbf{d}^T\nabla^2 f(\mathbf{x}^*)\mathbf{d} \geq 0$。
  - **证明:** 同样利用泰勒展开和反证法。
    - 将 $\varphi(\alpha)$ 在 $\alpha = 0$ 处进行二阶泰勒展开：$f(\mathbf{x}^* + \alpha\mathbf{d}) = f(\mathbf{x}^*) + \alpha\nabla f(\mathbf{x}^*)^T\mathbf{d} + (\alpha^2/2)\mathbf{d}^T\nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2)$
    - 由于 $\nabla f(\mathbf{x}^*)^T\mathbf{d} = 0$，如果 $\mathbf{d}^T\nabla^2 f(\mathbf{x}^*)\mathbf{d} < 0$，则对于足够小的 $\alpha > 0$，有 $f(\mathbf{x}^* + \alpha\mathbf{d}) < f(\mathbf{x}^*)$，与 **x*** 是局部极小点矛盾。
- **推论:** 如果 **x*** 是可行域内部的一个局部极小点，则 $\nabla f(\mathbf{x}^*) = 0$ 且 $\nabla^2 f(\mathbf{x}^*)$ 是半正定矩阵。

6.2.3 二阶充分条件 (Second-Order Sufficient Condition, SOSC)

- **定理 6.3 (SOSC):** 设 **x*** 是可行域 Ω 的一个内点，如果 $\nabla f(\mathbf{x}^*) = 0$ 且 $\nabla^2 f(\mathbf{x}^*)$ 是正定矩阵，则 **x*** 是一个严格局部极小点。
  - **证明:** 利用泰勒展开和正定矩阵的性质。
  - 将 f(x) 在 x* 处进行泰勒展开： $f(x) = f(x^*) + \nabla f(x^*)^T(x-x^*) + (1/2)(x-x^*)^T\nabla^2 f(x^*)(x-x^*) + o(||x-x^*||^2)$
  - 由于$\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$ 正定，则存在 $\lambda_{min} > 0$ 使得 $(x-x^*)^T\nabla^2 f(x^*)(x-x^*) > \lambda_{min}||x-x^*||^2$
  - 于是 $f(x) - f(x^*) = (1/2)(x-x^*)^T\nabla^2 f(x^*)(x-x^*) + o(||x-x^*||^2) \geqslant (\lambda_{min}/2)||x-x^*||^2 + o(||x-x^*||^2) > 0$
  - 因此 $f(x) > f(x^*)$
- **注意:** 对于边界点，SOSC 需要更复杂的形式。

6.3 拉格朗日乘子法 (Lagrange Multipliers)

6.3.1 等式约束优化问题

- 考虑如下等式约束优化问题:
- minimize f(x)
- subject to $h_i(x) = 0, i = 1, 2, ..., m$

其中 $f: R^n \rightarrow R$, $h_i: R^n \rightarrow R$。

- **拉格朗日函数 (Lagrangian):** $L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \Sigma_i \lambda_i h_i(\mathbf{x})$，其中 $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, ..., \lambda_m]^T$ 称为拉格朗日乘子。
- **定理 6.4 (拉格朗日乘子法):** 如果 **x*** 是上述等式约束优化问题的一个局部极小点，并且 $h_i(\mathbf{x})$ 在 **x*** 处线性无关 (即 $\nabla h_i(\mathbf{x}^*)$ 线性无关), 则存在一组拉格朗日乘子 **λ***，使得:
  - $\nabla_x L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \nabla f(\mathbf{x}^*) + \Sigma_i \lambda_i \nabla h_i(x^*) = 0$
  - $\nabla_\lambda L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = h_i(\mathbf{x}^*) = 0, i = 1, 2, ..., m$
  - **几何解释:** 在最优点 **x*** 处，目标函数的梯度 $\nabla f(\mathbf{x}^*)$ 可以表示为约束函数梯度 $\nabla h_i(\mathbf{x}^*)$ 的线性组合，即目标函数的等值线与约束曲面相切。

6.3.2 拉格朗日乘子的意义

- 拉格朗日乘子 λ_i^* 表示约束条件 $h_i(\mathbf{x}) = 0$ 发生微小变化时，目标函数最优值的变化率。
  - 假设 $h_i(\mathbf{x}) = 0$ 变为 $h_i(\mathbf{x}) = \varepsilon_i$，则最优值的变化近似为 $\Sigma_i \lambda_i^* \varepsilon_i$。

6.3.3 不等式约束优化问题 (简要介绍)

- 对于不等式约束 $g_i(\mathbf{x}) \leqslant 0$，可以引入松弛变量将其转化为等式约束，然后应用拉格朗日乘子法。
- **KKT 条件 (Karush-Kuhn-Tucker conditions):** 推广的拉格朗日乘子法，用于处理不等式约束。

本章小结

本章介绍了约束优化问题的最优性条件，包括一阶必要条件、二阶必要条件和二阶充分条件，以及将约束优化问题转化为无约束优化问题的拉格朗日乘子法。这些理论和方法为解决实际的约束优化问题提供了有力的工具。

第七章 不等式约束优化问题的最优性条件：KKT 条件

本章导言:

本章将重点介绍处理不等式约束优化问题的关键工具——卡罗需-库恩-塔克条件 (Karush-Kuhn-Tucker Conditions)，简称 KKT 条件。KKT 条件是一阶必要条件的推广，它不仅适用于等式约束，也适用于不等式约束。我们将首先回顾对偶性质引出 KKT 条件的重要性，然后详细介绍 KKT 条件的具体内容及其推导过程，包括互补松弛条件、原始可行性和对偶可行性等关键概念。最后，我们将通过例题讲解 KKT 条件的具体应用，并简要讨论 KKT 条件的局限性和实际应用中的求解方法。通过本章的学习，读者将能够理解并掌握 KKT 条件，并将其应用于解决实际的不等式约束优化问题。

7.1 对偶性质与 KKT 条件的引入

7.1.1 原问题与对偶问题

- **原问题 (Primal problem):**
- minimize $f(x)$
- subject to: $h_i(x) = 0$, $i = 1, \ldots, m$
- $g_i(x) \leqslant 0$, $i = 1, \ldots, p$
- $x \in R^n$
- **拉格朗日函数 (Lagrangian function):**
    - $L(x, \lambda, \mu) = f(x) + \Sigma_i \lambda_i h_i(x) + \Sigma_j \mu_j g_j(x)$
    - 其中 λ_i 为等式约束的拉格朗日乘子，μ_j 为不等式约束的拉格朗日乘子。
- **对偶函数 (Dual function):**
    - $d(\lambda, \mu) = \inf_x L(x, \lambda, \mu)$
- **对偶问题 (Dual problem):**
- maximize $d(\lambda, \mu)$
- subject to: $\mu \geqslant 0$

7.1.2 弱对偶性与强对偶性

- **弱对偶性 (Weak duality):** 对于任意可行解 x 和对偶可行解 (λ, μ)，都有 $d(\lambda, \mu) \leqslant f(x)$。即对偶问题的最优解是对偶问题的最优解是对原问题最优解的下界。
    - 证明:
    - 令 $A(x) = \max_{\lambda,\mu:\mu\geqslant 0} L(x, \lambda, \mu)$
    - $A(x) = \max_{\lambda,\mu:\mu\geqslant 0} L(x, \lambda, \mu) \geqslant L(x, \lambda, \mu) \geqslant \min_x L(x, \lambda, \mu) = d(\lambda, \mu)$
    - $A(x) \geqslant \min_x A(x) \geqslant \max_{\lambda,\mu:\mu\geqslant 0} d(\lambda, \mu) \geqslant d(\lambda, \mu)$
- **强对偶性 (Strong duality):** 在满足一定条件下 (例如 Slater 条件)，原问题的最优解等于对偶问题的最优解，即 $p^* = d^*$。
- **Slater 条件:** 存在一个 x 使得所有不等式约束严格成立，即 $g_i(x) < 0$。
- 通过引入对偶问题，我们可以得到关于原问题最优解的一个下界。这促使我们思考：在什么条件下，对偶问题的最优解能够精确地给出原问题的最优解呢？

7.2 KKT 条件

7.2.1 KKT 条件的陈述

- 对于一般约束优化问题:
- min $f(x)$
- subject to $h_i(x) = 0, i = 1, \ldots m$
- $g_i(x) \leq 0, i = 1, \ldots p$
- **定理 7.1 (KKT 条件):** 如果 x^* 是上述约束优化问题的一个局部最优解, 且在 x^* 处满足一定的正则性条件 (例如线性无关约束规范 LICQ) , 则存在拉格朗日乘子 λ^* 和 μ^*, 使得以下条件成立:
  1. **稳定性条件 (Stationarity):** $\nabla f(x) + \Sigma_i \lambda_i \nabla h_i(x) + \Sigma_j \mu_j \nabla g_j(x^*) = 0$
  2. **互补松弛条件 (Complementary slackness):** $\mu_j g_j(x) = 0, j = 1, \ldots, p$
  3. **原问题可行性 (Primal feasibility):**
     - $h_i(x^*) = 0, i = 1, \ldots, m$
     - $g_j(x^*) \leq 0, j = 1, \ldots, p$
  4. **对偶可行性 (Dual feasibility):** $\mu_j^* \geq 0, j = 1, \ldots, p$

7.2.2 KKT 条件的解释

- **稳定性条件:** 在最优点处, 目标函数的负梯度可以表示为所有起作用约束 (包括等式约束和不等式约束) 的梯度的线性组合, 且线性组合系数非负。
- **互补松弛条件:** 对于每个不等式约束, 要么拉格朗日乘子 μ_j^* 为 0 (约束不起作用), 要么约束条件取等号 $g_j(x^*) = 0$ (约束起作用)。
- **原问题可行性:** 解必须满足所有约束条件。
- **对偶可行性:** 不等式约束对应的拉格朗日乘子非负。

7.2.3 KKT 条件的推导 (简要说明)

- KKT 条件可以看作是拉格朗日乘子法在不等式约束情况下的推广。其推导过程较为复杂, 主要思想是通过构造一个辅助函数, 利用可行方向和泰勒展开等工具, 推导出最优解必须满足的条件。

7.3 KKT 条件的应用

7.3.1 求解步骤

1. 写出约束优化问题的拉格朗日函数。
2. 列出 KKT 条件 (稳定性条件、互补松弛条件、原始可行性、对偶可行性)。
3. 求解 KKT 条件得到的方程组和不等式组, 得到候选的最优解和对应的拉格朗日乘子。
4. 验证正则性条件 (例如 LICQ), 并根据二阶条件或其他方法进一步判断候选解是否为局部最优解。

7.3.2 例题讲解

- **例题 1 (参考 PDF 第 9-10 页):**
- min x^2
- subject to $1 \leq x \leq 2$
  - 解题过程:
    1. 将约束条件改写为标准形式: $-x + 1 \leq 0$ 和 $x - 2 \leq 0$
    2. 构造拉格朗日函数: $L(x, \lambda_1, \lambda_2) = x^2 + \lambda_1(-x + 1) + \lambda_2(x - 2)$
    3. 列出 KKT 条件:
       - $2x - \lambda_1 + \lambda_2 = 0$
       - $\lambda_1(-x + 1) = 0$

        - $\lambda_2(x - 2) = 0$
        - $-x + 1 \leqslant 0$
        - $x - 2 \leqslant 0$
        - $\lambda_1 \geqslant 0$
        - $\lambda_2 \geqslant 0$
    4. 求解 KKT 条件:
        - 分析 λ_1 和 λ_2 的取值情况，得到 $x^* = 1, \lambda_1^* = 2, \lambda_2^* = 0$
    5. 验证解的有效性: $x^* = 1$ 是该问题的全局最优解。

- **例题 2 (参考 PDF 第 11-16 页):**
- $\min f(x_1, x_2) = x_1^2 + 2x_2^2 - 4x_1 - 4x_2$
- subject to: $g_1(x_1, x_2) = x_1 + x_2 - 3 \leqslant 0$
- $g_2(x_1, x_2) = 5 - x_1 - 2x_2 \leqslant 0$
    - 解题过程:
        1. 构造拉格朗日函数: $L(x_1, x_2, \lambda_1, \lambda_2) = x_1^2 + 2x_2^2 - 4x_1 - 4x_2 + \lambda_1(x_1 + x_2 - 3) + \lambda_2(5 - x_1 - 2x_2)$
        2. 列出 KKT 条件:
            - $2x_1 - 4 + \lambda_1 - \lambda_2 = 0$
            - $4x_2 - 4 + \lambda_1 - 2\lambda_2 = 0$
            - $\lambda_1 g_1(x_1, x_2) = 0$
            - $\lambda_2 g_2(x_1, x_2) = 0$
            - $g_1(x_1, x_2) \leqslant 0$
            - $g_2(x_1, x_2) \leqslant 0$
            - $\lambda_1 \geqslant 0$
            - $\lambda_2 \geqslant 0$
        3. 求解 KKT 条件:
            - 分别讨论 λ_1 和 λ_2 的四种取值组合 (0, 0), (0, +), (+, 0), (+, +)，并求解相应的方程组。
            - 对于每种组合，检查求得的解是否满足 KKT 条件。
            - 最终得到满足所有 KKT 条件的解: $x_1^* = 1, x_2^* = 2, \lambda_1^* = 8, \lambda_2^* = 6$。
        4. 验证解的有效性: 可以验证该解满足二阶充分条件，因此是局部最优解 (由于目标函数是凸函数，也是全局最优解)。

7.4 KKT 条件的局限性与实际应用

7.4.1 KKT 条件的局限性

- KKT 条件是必要条件，而不是充分条件。满足 KKT 条件的点可能是局部最优解、鞍点或全局最优解，需要进一步判断。
- KKT 条件的求解通常需要解一个非线性方程组和不等式组，计算复杂度较高。
- KKT 条件要求目标函数和约束函数可微。
- 对于某些非凸优化问题，即使满足强对偶性，KKT 条件也不一定能找到全局最优解。

7.4.2 实际应用

- KKT 条件是许多优化算法的基础，例如内点法、序列二次规划 (SQP) 等。
- 在实际应用中，通常需要结合数值方法和启发式方法来求解 KKT 条件。
- 可以使用一些现成的优化软件 (例如 MATLAB 的 fmincon 函数) 来求解 KKT 条

件。

- 对于大规模问题，可以考虑使用分解协调等方法来降低计算复杂度。
- 一些计算工具和程序可以帮助求解 KKT 条件，比如 PDF 文件中提到的利用 SymPy 库进行符号求解(参考 PDF 文件第 19 页) 和利用梯度下降法进行数值求解(参考 PDF 文件第 20 页)

本章小结

本章详细介绍了 KKT 条件，这是解决不等式约束优化问题的重要工具。我们从对偶性质出发，引出了 KKT 条件的重要性，并详细阐述了 KKT 条件的内容、解释和应用。通过例题讲解，读者可以更好地理解 KKT 条件的具体应用步骤。最后，我们也讨论了 KKT 条件的局限性和实际应用中的一些处理方法。

第八章 牛顿法 (Newton's Method)

本章导言：

本章将介绍一种经典的求解无约束优化问题的迭代算法——牛顿法。牛顿法利用目标函数的二阶导数信息，通过构造二次近似模型来逼近最优点，具有较快的收敛速度，特别是对于二次函数，牛顿法可以一步到位找到最优解。我们将首先介绍牛顿法的基本思想和迭代公式，然后分析其收敛性质，包括收敛速度和收敛条件。接着，我们将讨论牛顿法的优缺点，并介绍一些改进的牛顿法，例如阻尼牛顿法和 Levenberg-Marquardt 方法。此外，我们还将比较牛顿法和梯度下降法，尤其会通过具体的例子以及收敛速度的对比，来体现牛顿法的优势。最后，我们将通过习题来巩固所学内容。通过本章的学习，读者将能够深入理解牛顿法的原理，掌握其应用方法，并了解其优缺点和改进方向。

8.1 牛顿法的基本思想与迭代公式

8.1.1 牛顿法的基本思想

- 牛顿法的核心思想是在当前迭代点 x_k 处，用一个二次函数 q(x) 来近似目标函数 f(x)，然后求解二次函数的极小点作为下一个迭代点 x_{k+1}。这个二次函数是通过目标函数 f(x) 在 x_k 处的二阶泰勒展开得到的。
- 从几何上看，牛顿法是用一个抛物面来逼近目标函数的等值线，并用抛物面的顶点作为下一个迭代点。

8.1.2 牛顿法的迭代公式

- 考虑无约束优化问题：$\min f(x), x \in R^n$
- 将 f(x) 在当前迭代点 x_k 处进行二阶泰勒展开，得到二次近似函数：
  - $q(x) = f(x_k) + \nabla f(x_k)^T(x - x_k) + (1/2)(x - x_k)^T\nabla^2 f(x_k)(x - x_k)$
- 求解二次函数的极小点，即令 $\nabla q(x) = 0$，得到：
  - $\nabla f(x_k) + \nabla^2 f(x_k)(x - x_k) = 0$
- 如果 $\nabla^2 f(x_k)$ 可逆，则可以解得：
  - $x = x_k - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$
- 因此，牛顿法的迭代公式为：
  - $x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$
- 另一种表示形式：

1. 求解线性方程组： $\nabla^2 f(x_k)d_k = -\nabla f(x_k)$ 得到搜索方向 d_k
2. 更新迭代点： $x_{k+1} = x_k + d_k$

- 当 $n = 1$ 时，牛顿法的迭代公式简化为：
    - $x_{k+1} = x_k - f'(x_k) / f''(x_k)$
    - 几何解释：在一维情况下，牛顿法是用 $f(x)$ 在 x_k 处的切线的零点作为下一个迭代点。

8.2 牛顿法的收敛性分析

8.2.1 收敛速度

- **定理 8.1:** 设 $f \in C^3$，x^* 是一个局部极小点，且 $\nabla f(x^*) = 0$，$\nabla^2 f(x^*)$ 可逆。如果初始点 x_0 充分靠近 x^*，则牛顿法产生的序列 $\{x_k\}$ 收敛到 x^*，且收敛速度是二阶的，即：
    - $\|x_{k+1} - x^*\| \leq C\|x_k - x^*\|^2$，其中 C 是一个常数。
- **推导过程：** (参考 PDF 第 29-31 页)
    - 令 $F(x) = \nabla^2 f(x)$，根据泰勒展开、海森矩阵的性质以及 $F(x^*)$ 可逆性进行推导。
    - 关键步骤：
    - $x_{k+1} - x^* = x_k - x^* - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k) = [\nabla^2 f(x_k)]^{-1} [\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k)]$
    - $\nabla f(x^*) = \nabla f(x_k) + \nabla^2 f(x_k)(x^* - x_k) + O(\|x_k - x^*\|^2) = 0$
    - 利用 $\nabla f(x^*) = 0$，进行符号变换，得到： $\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k) = O(\|x_k - x^*\|^2)$
    - 最终得到 $\|x_{k+1} - x^*\| \leq C_1\|x_k - x^*\|^2$

8.2.2 收敛条件

- 初始点 x_0 必须充分靠近局部极小点 x^*。
- 目标函数 $f(x)$ 必须三阶连续可微。
- 在迭代过程中，海森矩阵 $\nabla^2 f(x_k)$ 必须可逆。

8.3 牛顿法的优缺点

8.3.1 优点

- **收敛速度快:** 在满足条件下，牛顿法具有二阶收敛速度，比梯度下降法（一阶收敛）快得多。尤其对于二次函数，牛顿法可以一步收敛到最优解。(参考 PDF 第 25 页图)
- **仿射不变性 (Affine invariant):** 牛顿法的收敛性不依赖于坐标系的选择。

8.3.2 缺点

- **局部收敛性:** 牛顿法只有当初始点充分靠近最优解时才能保证收敛。如果初始点远离最优解，牛顿法可能不收敛，甚至可能收敛到鞍点或极大点。(参考 PDF 第 26 页图)
- **计算量大:** 每次迭代都需要计算海森矩阵 $\nabla^2 f(x_k)$ 及其逆矩阵，对于大规模问题，计算量很大。
    - 计算海森矩阵的时间复杂度为 $O(n^2)$，求逆的复杂度为 $O(n^3)$。
- **海森矩阵奇异性问题:** 如果海森矩阵 $\nabla^2 f(x_k)$ 不可逆（奇异），则牛顿法无法进行。

8.4 牛顿法的改进

8.4.1 阻尼牛顿法 (Damped Newton's Method)

- 为了改善牛顿法的全局收敛性，可以在迭代公式中引入一个步长因子 α_k，即：
    - $x_{k+1} = x_k - \alpha_k[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$
- 步长因子 α_k 可以通过线性搜索 (line search) 来确定，例如：
    - $\alpha_k = \mathrm{argmin}_{\alpha\geq 0} f(x_k - \alpha[\nabla^2 f(x_k)]^{-1}\nabla f(x_k))$

- 阻尼牛顿法可以保证每次迭代都使目标函数值下降，从而提高算法的全局收敛性。

8.4.2 Levenberg-Marquardt 方法

- 针对海森矩阵奇异或不正定的情况，Levenberg-Marquardt 方法对牛顿法的迭代公式进行了修改：
  - $x_{k+1} = x_k - (\nabla^2 f(x_k) + \mu_k I)^{-1}\nabla f(x_k)$
  - 其中 μ_k 是一个正数，I 是单位矩阵。
- 当 $\mu_k = 0$ 时，Levenberg-Marquardt 方法退化为牛顿法。
- 当 $\mu_k \to +\infty$ 时，Levenberg-Marquardt 方法趋近于梯度下降法。
- 通过调整 μ_k 的大小，Levenberg-Marquardt 方法可以在牛顿法和梯度下降法之间进行切换，从而兼顾二者的优点。
- μ_k 的选择策略：
  - 选择 $\mu_k > 0$ 使得 $\nabla^2 f(x_k) + \mu_k I$ 正定。
  - 如果目标函数值下降，则减小 μ_k。
  - 如果目标函数值上升，则增大 μ_k。

8.5 牛顿法与梯度下降法的比较

特性	牛顿法	梯度下降法
收敛速度	二阶收敛	一阶收敛
计算量	较大（计算海森矩阵及其逆矩阵）	较小（仅计算梯度）
收敛性	局部收敛	全局收敛（步长选择合适的情况下）
海森矩阵要求	可逆	无
步长	通常为 1（经典牛顿法）或通过线搜索确定	需要选择合适的步长（例如，通过线搜索）
适用范围	初始点靠近最优解，且海森矩阵可逆的情况	更广泛，可用于海森矩阵不可逆或难以计算的情况

8.6 应用举例（结合前面章节）

- 利用牛顿法求解方程组：
- $f(x) = 0$，其中 $f: R^n \to R^n$
- 迭代公式： $x_{k+1} = x_k - J(x_k)^{-1}f(x_k)$，其中 $J(x_k)$ 为雅可比矩阵。
- 求解逻辑回归（Logistic Regression）的参数：
- 目标函数为对数似然函数的负数。
- 利用牛顿法迭代求解参数。

8.7 关于牛顿法中步长的一种理解方式

- 将 $f(x_{k+1})$ 在 x_k 附近进行泰勒展开： $f(x_{k+1}) \approx f(x_k) + f'(x_k)(x_{k+1} - x_k) + (1/2)f''(x_k)(x_{k+1} - x_k)^2$
- 为了简化计算，将最后一项的 $f''(x_k)$ 替换为 $1/t$，得到：$f(x_{k+1}) \approx f(x_k) + f'(x_k)(x_{k+1} - x_k) + (1/(2t))(x_{k+1} - x_k)^2$
- 此时，可以推导出 $x_{k+1} = x_k - tf'(x_k)$

习题

1. 证明对于二次函数，牛顿法可以一步到位找到最优解。
2. 使用牛顿法求解函数 $f(x) = x^3 - 2x + 2$ 的极小点，初始点分别为 $x_0 = 0$ 和 $x_0 = 1.5$，并观察迭代过程。(可参考教材第 26 页图)
3. 使用牛顿法求解方程组:
4. $x_1^2 + x_2^2 - 1 = 0$
5. $x_1 - x_2 = 0$

初始点为 $x_0 = (1, 0)^T$。

6. 考虑逻辑回归模型，目标函数为对数似然函数的负数，推导参数的牛顿法迭代公式。
7. (参考教材第 22 页) 考虑函数 $f(x_1, x_2) = x_1^4 + x_2^2$，从初始点 $x_0 = (1, 1)^T$ 开始，分别使用梯度下降法和牛顿法进行迭代，比较二者的收敛速度。
8. (参考教材第 36 页) 考虑函数 $y = x^{4/3}$，如何修改牛顿法使其能够有效求解该函数的极小点?

本章小结

本章详细介绍了牛顿法的原理、迭代公式、收敛性、优缺点以及改进方法。牛顿法是一种高效的求解无约束优化问题的迭代算法，特别适用于目标函数具有良好二次性质的情况。然而，牛顿法也存在局部收敛性和计算量大的问题，需要根据具体情况选择合适的算法或改进策略。

第九章 次梯度 (Sub-gradient)

本章导言:

在前面的章节中，我们学习的优化算法，如梯度下降法和牛顿法，都依赖于目标函数的可微性。然而，在实际问题中，我们经常会遇到不可微的函数，例如带有绝对值项的函数 (如 ℓ_1 范数)。为了处理这类不可微的凸优化问题，本章将引入次梯度的概念，它是梯度概念的推广。我们将首先定义次梯度和次微分，然后探讨次梯度的性质和计算规则。接着，我们将介绍基于次梯度的优化方法——次梯度方法，并分析其收敛性。最后，我们将通过一些例子，例如 Lasso 问题，来展示次梯度方法在求解实际问题中的应用。通过本章的学习，读者将能够理解次梯度的概念，掌握次梯度的计算方法，并能够运用次梯度方法解决实际的不可微凸优化问题。

9.1 次梯度与次微分

9.1.1 次梯度的定义

- **回顾梯度的性质:** 对于可微的凸函数 f，其在任意一点 x 处的梯度 $\nabla f(x)$ 满足以下不等式:
  - $f(y) \geq f(x) + \nabla f(x)^T(y - x), \forall y$
  - 几何意义：函数 f(x) 的图像始终位于其在点 x 处切线的上方，即线性近似总是低估了函数值。
- **次梯度 (Sub-gradient):** 对于凸函数 f (不一定可微)，其在点 x 处的次梯度是一个向量 g，满足以下不等式:
  - $f(y) \geq f(x) + g^T(y - x), \forall y$
  - 几何意义：函数 f(x) 的图像始终位于过点 (x, f(x)) 且斜率为 g 的超平面的上方。

- **注:** 即使对于非凸函数，也可以定义次梯度，但次梯度不一定存在。

9.1.2 次微分 (Sub-differential)

- **次微分的定义:** 函数 f 在点 x 处的所有次梯度的集合称为 f 在 x 处的次微分，记作 $\partial f(x)$。
- **性质:**
  - $\partial f(x)$ 是一个闭凸集 (即使 f 是非凸函数)。
  - 对于凸函数，$\partial f(x)$ 非空。
  - 如果 f 在 x 处可微，则 $\partial f(x) = \{\nabla f(x)\}$，即次微分只包含梯度。
  - 如果 $\partial f(x) = \{g\}$，即次微分只包含一个元素，则 f 在 x 处可微，且 $\nabla f(x) = g$。

9.2 次梯度的例子

9.2.1 一维绝对值函数

- $f(x) = |x|$
- 当 $x > 0$ 时，$\partial f(x) = \{1\}$
- 当 $x < 0$ 时，$\partial f(x) = \{-1\}$
- 当 $x = 0$ 时，$\partial f(x) = [-1, 1]$

9.2.2 ℓ_2 范数

- $f(x) = ||x||_2, x \in R^n$
- 当 $x \neq 0$ 时，$\partial f(x) = \{x / ||x||_2\}$
- 当 $x = 0$ 时，$\partial f(x) = \{z : ||z||_2 \leq 1\}$
- **说明**: 当 $x=0$ 时，函数 $f(x) = ||x||_2$ 不可微。此时，其在原点处的次微分是单位球内的所有向量。

9.2.3 ℓ_1 范数

- $f(x) = ||x||_1, x \in R^n$
- $\partial f(x) = \{g : g_i \in \text{sign}(x_i) \text{ if } x_i \neq 0, g_i \in [-1, 1] \text{ if } x_i = 0\}$

9.2.4 两个可微凸函数的最大值

- $f(x) = \max\{f_1(x), f_2(x)\}$，其中 $f_1(x)$ 和 $f_2(x)$ 是可微的凸函数。
- 当 $f_1(x) > f_2(x)$ 时，$\partial f(x) = \{\nabla f_1(x)\}$

- 当 $f_1(x) < f_2(x)$ 时，$\partial f(x) = \{\nabla f_2(x)\}$
- 当 $f_1(x) = f_2(x)$ 时，$\partial f(x) = \mathrm{conv}\{\nabla f_1(x), \nabla f_2(x)\}$，即 $\nabla f_1(x)$ 和 $\nabla f_2(x)$ 的凸包。

9.3 次梯度运算法则

9.3.1 数乘

- $\partial(af)(x) = a\partial f(x)$，其中 $a > 0$

9.3.2 加法

- $\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$

9.3.3 仿射变换

- 如果 $g(x) = f(Ax + b)$，则 $\partial g(x) = A^T\partial f(Ax + b)$

9.3.4 有限逐点最大值

- 如果 $f(x) = \max_{i=1,\dots,m} f_i(x)$，则 $\partial f(x) = \mathrm{conv}(\cup_{i:f_i(x)=f(x)} \partial f_i(x))$，即在 x 处取到最大值的所有 $f_i(x)$ 的次微分的并集的凸包。

9.4 次梯度方法

9.4.1 次梯度方法的迭代公式

- 考虑无约束凸优化问题：$\min f(x)$
- 次梯度方法的迭代公式为：
  - $x^{(k+1)} = x^{(k)} - t_k g^{(k)}$
  - 其中 $x^{(k)}$ 是第 k 次迭代的解，t_k 是步长，$g^{(k)}$ 是 $f(x)$ 在 $x^{(k)}$ 处的任意一个次梯度，即 $g^{(k)} \in \partial f(x^{(k)})$。
- **注意:** 次梯度方法与梯度下降法的区别在于，次梯度方法使用的是次梯度，而不是梯度。

9.4.2 步长选择

- **固定步长:** $t_k = t$
- **逐渐减小步长:** $t_k = \alpha / \sqrt{k}$，其中 α 是一个常数。
- **其他更复杂的步长选择策略。**

9.4.3 收敛性分析

- **定理 9.1:** 如果 $f(x)$ 是凸函数且满足 Lipschitz 连续条件（即存在常数 $L > 0$，使得 $\|g\| \leq L, \forall g \in \partial f(x)$），并且步长选择满足一定条件（例如 $t_k = \alpha / \sqrt{k}$），则次梯度方法收敛，即：
  - $\lim_{k\to\infty} f(x^{(k)}) = f$，*其中 f* 是 $f(x)$ 的最优值。
- **收敛速度:** 次梯度方法的收敛速度通常为 $O(1/\sqrt{k})$，比梯度下降法的收敛速度慢。

9.5 次梯度方法的应用：Lasso 问题

9.5.1 Lasso 问题的定义

- Lasso 问题可以表示为以下优化问题:
  - $\min_\beta (1/2)||y - X\beta||_2^2 + \lambda||\beta||_1$
  - 其中 $y \in R^n$ 是观测向量，$X \in R^{n\times p}$ 是设计矩阵，$\beta \in R^p$ 是待估计的参数向量，λ 是正则化参数。

9.5.2 Lasso 问题的最优性条件

- 利用次梯度 optimality condition，Lasso 问题的最优解 β* 满足以下条件:
  - $0 \in -X^T(y - X\beta) + \lambda\partial||\beta||_1$
  - 即 $X^T(y - X\beta) = \lambda v$，其中 $v \in \partial||\beta||_1$
- 根据 ℓ_1 范数的次微分性质，可以得到:
  - $X_i^T(y - X\beta) = \lambda \cdot sign(\beta_i)$ if $\beta_i* \neq 0$
  - $|X_i^T(y - X\beta)| \leq \lambda$ if $\beta_i = 0$

9.5.3 软阈值算子 (Soft-thresholding operator)

- 当 X = I 时，Lasso 问题可以得到显式解:
  - $\beta* = S_\lambda(y)$
  - 其中 $S_\lambda(y)$ 是软阈值算子，其定义为:
    - $[S_\lambda(y)]_i = (y_i - \lambda)_+$ if $y_i > \lambda$
    - $[S_\lambda(y)]_i = 0$ if $-\lambda \leq y_i \leq \lambda$
    - $[S_\lambda(y)]_i = (y_i + \lambda)_-$ if $y_i < -\lambda$
  - $(y)_+$ 表示取 y 的正部，$(y)_-$ 表示取 y 的负部。

9.6 次梯度方法的优缺点

9.6.1 优点

- 可以处理不可微的凸优化问题。
- 算法简单，易于实现。
- 对目标函数的性质要求较低，只需要凸性和 Lipschitz 连续性。

9.6.2 缺点

- 收敛速度较慢，通常为 $O(1/\sqrt{k})$。
- 步长选择对算法性能影响较大。
- 次梯度方法不一定是下降方法，目标函数值在迭代过程中可能会出现波动。

习题

1. 计算函数 $f(x) = \max\{x^2, 2x + 3\}$ 的次微分。
2. 证明次微分的加法法则。
3. 使用次梯度方法求解以下优化问题:
   - $\min |x - 1| + |x - 2|$
4. 考虑 Lasso 问题，证明软阈值算子 $S_\lambda(y)$ 满足 Lasso 问题的最优性条件。
5. (参考教材第 33 页) 编写程序实现软阈值算子，并求解简化的 Lasso 问题。

本章小结

本章介绍了次梯度的概念、性质和计算规则，以及基于次梯度的优化方法——次梯度方法。次梯度方法可以用于求解不可微的凸优化问题，具有广泛的应用。我们通过 Lasso 问题展

示了次梯度方法在实际问题中的应用。

第十章 迭代法与临近点算法（Iterative Methods and Proximal Algorithms）

本章导言:

本章将介绍求解线性方程组的迭代方法和一类重要的优化算法——临近点算法 (Proximal Algorithm)。我们将首先回顾求解线性方程组 $Ax = b$ 的经典迭代方法，例如 Jacobi 方法和 Gauss-Seidel 方法，并讨论它们的收敛性。随后，我们将引入临近点算法的概念，并详细讲解其与梯度下降法的联系和区别。临近点算法通过求解一个更简单的子问题来逼近原问题的解，特别适合于处理目标函数中包含不可微项的情况，例如包含 ℓ_1 范数的优化问题。我们将通过推导软阈值算子来具体说明临近点算法的应用，并介绍其在求解 Lasso 问题中的应用。最后，我们将介绍临近点梯度法 (Proximal Gradient Method) 及其快速版本 (FISTA)，并简要讨论其收敛性。通过本章的学习，读者将掌握求解线性方程组的迭代方法和临近点算法的基本原理，并能够运用这些方法解决实际的优化问题。

10.1 求解线性方程组的迭代方法

10.1.1 Jacobi 迭代法

- 考虑线性方程组 $Ax = b$，其中 $A \in R^{n\times n}$ 是非奇异矩阵，$b \in R^n$。
- 将 A 分解为 $A = D - L - U$，其中 D 是 A 的对角部分，-L 是 A 的严格下三角部分，-U 是 A 的严格上三角部分。
- Jacobi 迭代法的迭代公式为:
    - $x^{(k+1)} = D^{-1}(L + U)x^{(k)} + D^{-1}b$
    - 分量形式:
        - $x_i^{(k+1)} = (1/a_{ii})(b_i - \Sigma_{j\neq i} a_{ij}x_j^{(k)}), i = 1, 2, ..., n$
- **基本思想:** 每次迭代，用其他分量的当前值来更新一个分量。
- **例子:** (参考 PDF 第 23 页)
- $5x_1 - x_2 + 2x_3 = 12$
- $3x_1 + 8x_2 - 2x_3 = -25$
- $x_1 + x_2 + 4x_3 = 6$

可以改写为:

$x_1 = (12 + x_2 - 2x_3) / 5$

$x_2 = (-25 - 3x_1 + 2x_3) / 8$

$x_3 = (6 - x_1 - x_2) / 4$

然后进行迭代求解。

- **局限性:** Jacobi 迭代法不总是收敛的。例如，对于以下方程组 (参考 PDF 第 25 页):
- $x_1 + 7x_2 = 0$
- $3x_1 + x_2 = 1$

使用 Jacobi 迭代法会发散。

10.1.2 Gauss-Seidel 迭代法

- Gauss-Seidel 迭代法的迭代公式为:

- $x^{(k+1)} = (D - L)^{-1}Ux^{(k)} + (D - L)^{-1}b$
    - 分量形式：
        - $x_i^{(k+1)} = (1/a_{ii})(b_i - \Sigma_{j<i} a_{ij}x_j^{(k+1)} - \Sigma_{j>i} a_{ij}x_j^{(k)})$, $i = 1, 2, ..., n$
- **基本思想：** 每次迭代，用已经更新过的分量来更新其他分量。
- **例子：** (参考 PDF 第 24 页) 使用与 Jacobi 迭代法相同的例子，Gauss-Seidel 迭代公式为：
- $x_1^{(k+1)} = (12 + x_2^{(k)} - 2x_3^{(k)}) / 5$
- $x_2^{(k+1)} = (-25 - 3x_1^{(k+1)} + 2x_3^{(k)}) / 8$
- $x_3^{(k+1)} = (6 - x_1^{(k+1)} - x_2^{(k+1)}) / 4$
- **与 Jacobi 迭代法的比较：** Gauss-Seidel 迭代法通常比 Jacobi 迭代法收敛更快，但仍然不能保证收敛。

10.1.3 收敛性分析

- **定理 10.1：** 如果 A 是严格对角占优矩阵，则 Jacobi 迭代法和 Gauss-Seidel 迭代法都收敛。
- **严格对角占优矩阵：** 对于矩阵 A，如果满足 $|a_{ii}| > \Sigma_{j\neq i} |a_{ij}|$, $\forall i$，则称 A 是严格对角占优矩阵。

10.2 临近点算法 (Proximal Algorithm)

10.2.1 临近点算子 (Proximal Operator)

- 对于凸函数 $f: R^n \to R$，其临近点算子定义为：
    - $\text{prox}_{\lambda f}(v) = \text{argmin}_x (f(x) + (1/(2\lambda))||x - v||_2^2)$，其中 $\lambda > 0$ 是一个参数。
- **直观理解：** $\text{prox}_{\lambda f}(v)$ 是在函数 f(x) 的值和与点 v 的距离之间进行权衡后得到的点。参数 λ 控制权衡的力度。
- **几何解释：** (参考 PDF 第 26 页) $\text{prox}_{\lambda f}(v)$ 可以看作是在 v 附近寻找一个使 f(x) 尽可能小的点，同时又不会离 v 太远。

10.2.2 临近点算法的迭代公式

- 考虑无约束优化问题：$\min f(x) + g(x)$，其中 f(x) 是光滑凸函数，g(x) 是凸函数 (不一定光滑)。
- 临近点算法的迭代公式为：
    - $x^{(k+1)} = \text{prox}_{\lambda g}(x^{(k)} - \lambda\nabla f(x^{(k)}))$，其中 $\lambda > 0$ 是步长。
- **基本思想：** 每次迭代，先沿着 f(x) 的负梯度方向走一步，然后用 g(x) 的临近点算子将迭代点拉回到一个使 g(x) 较小的区域。

10.2.3 临近点算子的性质

- **投影算子：** 当 g(x) 是一个闭凸集 C 的示性函数 (indicator function) 时，即：
    - $g(x) = I_C(x) = \{ 0, \text{if } x \in C; +\infty, \text{if } x \notin C \}$
    - 临近点算子 $\text{prox}_{\lambda g}(v)$ 退化为投影算子，即 $\text{prox}_{\lambda g}(v) = P_C(v)$，其中 $P_C(v)$ 表示 v 在集合 C 上的投影 (参考 PDF 第 27 页)。
- **更一般的形式：** (参考 PDF 第 28 页)
    - $\min f(x) + \lambda g(x)$
    - 其中 f(x) 光滑，而 g(x) 不一定光滑，但其临近点算子 $\text{prox}_{\lambda g}(x)$ 容易计算。
    - 迭代公式：$x^{(k+1)} = \text{prox}_{\lambda_k g}(x^{(k)} - \lambda_k\nabla f(x^{(k)}))$，其中 $\lambda_k > 0$ 是步长。

10.2.4 临近点算子的计算

- 对于一些特殊的函数 g(x)，其临近点算子 $\text{prox}_{\lambda g}(x)$ 可以显式地计算出来。

- **例 1:** $g(x) = |x|, x \in R$ * $\text{prox}_{\lambda g}(b) = \{ b - \lambda, \text{if } b \geqslant \lambda; 0, \text{if } |b| \leqslant \lambda; b + \lambda, \text{if } b \leqslant -\lambda \}$ (参考 PDF 第 29 页) * 证明过程：
    - 当 $b \geqslant 0$ 时，目标函数为 $(1/(2\lambda))(x - b)^2 + x$，其极小点为 $\tilde{x} = b - \lambda$。
    - 如果 $\tilde{x} \geqslant 0$，即 $b \geqslant \lambda$，则 $\text{prox}_{\lambda g}(b) = b - \lambda$。
    - 如果 $\tilde{x} < 0$，即 $b < \lambda$，则 $\text{prox}_{\lambda g}(b) = 0$。
    - 当 $b < 0$ 时，同理可得 $\text{prox}_{\lambda g}(b) = b + \lambda$ (如果 $b \leqslant -\lambda$) 或 0 (如果 $|b| < \lambda$)。 * 可以合并写成：
    - $\text{prox}_{\lambda g}(b) = \{ \text{sgn}(b)(|b| - \lambda), \text{if } |b| > \lambda; 0, \text{otherwise} \}$

- **例 2:** $g(x) = ||x||_1, x \in R^n$ * $\text{prox}_{\lambda g}(b) = S_\lambda(b)$，其中 $S_\lambda(b)$ 是软阈值算子 (soft-thresholding operator)，其定义为：
    - $[S_\lambda(b)]_i = \{ b_i - \lambda, \text{if } b_i > \lambda; 0, \text{if } |b_i| \leqslant \lambda; b_i + \lambda, \text{if } b_i < -\lambda \}$ (参考 PDF 第 30 页)
    - 即逐分量应用一维情况下的软阈值算子。

10.3 临近点梯度法 (Proximal Gradient Method)

10.3.1 临近点梯度法的迭代公式

- 考虑优化问题：$\min F(x) = f(x) + g(x)$，其中 $f(x)$ 是光滑凸函数，$g(x)$ 是凸函数 (不一定光滑)。
- 临近点梯度法的迭代公式为：
- $x^{(k+1)} = \text{prox}_{t_k g}(x^{(k)} - t_k\nabla f(x^{(k)}))$，其中 $t_k > 0$ 是步长。
- **与梯度下降法的关系:** 当 $g(x) = 0$ 时，临近点梯度法退化为梯度下降法。
- **与迭代软阈值算法 (ISTA) 的关系:** 当 $g(x) = \lambda||x||_1$ 时，临近点梯度法退化为 ISTA 算法。

10.3.2 快速临近点梯度法 (FISTA - Fast Iterative Shrinkage-Thresholding Algorithm)

- FISTA 是对临近点梯度法的一种加速方法，其迭代公式为：
    - $x^{(k+1)} = \text{prox}_{t_k g}(y^{(k)} - t_k\nabla f(y^{(k)}))$
    - $y^{(k+1)} = x^{(k+1)} + ((k - 1)/(k + 2))(x^{(k+1)} - x^{(k)})$
    - 其中 $y^{(k)}$ 是一个辅助变量。
- **与临近点梯度法的比较:** FISTA 通过引入一个额外的动量项 (momentum term) 来加速收敛。

10.3.3 收敛性分析

- **定理 10.2:** 如果 $f(x)$ 是光滑凸函数，$\nabla f(x)$ 是 Lipschitz 连续的，$g(x)$ 是凸函数，则临近点梯度法和 FISTA 都收敛。
- **收敛速度:** 临近点梯度法的收敛速度为 $O(1/k)$，FISTA 的收敛速度为 $O(1/k^2)$。

10.4 应用：Lasso 问题

- Lasso 问题的目标函数可以写成 $f(x) + g(x)$ 的形式，其中：
    - $f(x) = (1/2)||Ax - b||_2^2$ (光滑凸函数)
    - $g(x) = \lambda||x||_1$ (凸函数，不可微)
- $\nabla f(x) = A^T(Ax - b)$
- $\text{prox}_{\lambda g}(x) = S_\lambda(x)$ (软阈值算子)
- 因此, Lasso 问题可以用临近点梯度法 (ISTA) 或 FISTA 求解 (参考 PDF 第 31 页)。

习题

1. 证明 Gauss-Seidel 迭代法的收敛性定理 (当 A 是严格对角占优矩阵时)。
2. 推导 ℓ_0 范数的临近点算子。
3. 使用 FISTA 算法求解 Lasso 问题，并与 ISTA 算法的收敛速度进行比较。
4. (参考教材第 2 页) 实现求解 Lasso 问题的程序，并输出类似第 3 页图的结果。
5. (参考教材第 2 页) 使用不同的初始值，测试牛顿法求解教材第 26 页给出的例子，并观察迭代结果。

本章小结

本章介绍了求解线性方程组的迭代方法和临近点算法。迭代方法为求解大型线性方程组提供了有效的途径。临近点算法是一类重要的优化算法，特别适合于处理目标函数中包含不可微项的情况。我们通过 Lasso 问题展示了临近点算法在实际问题中的应用，并介绍了 FISTA 算法来加速收敛。

第十一章 练习题

本章导言：

本章包含六道练习题，涵盖了前面章节的主要内容，包括：

- 向量空间的性质
- 线性方程组的求解
- 二次型的应用
- 概率图模型 (PGM)
- 最优性条件
- 拉格朗日乘子法与 KKT 条件

每道题都提供了详细的解答，希望读者能够通过这些练习题巩固所学知识，并提高解决实际问题的能力。

练习题 1：向量空间与线性方程组

设 $f: \mathbb{R}^n \to \mathbb{R}$ 为连续可微函数，写出相应的梯度和 Hessian 矩阵，并写出 Taylor 公式 (写至 2 阶)。

解答：

- **梯度 (Gradient):** $\nabla f(x)$ 是一个 n 维向量，其第 i 个分量为 $\partial f(x)/\partial x_i$。

  $$\nabla f(x) = \begin{bmatrix} \partial f(x)/\partial x_1 \\ \partial f(x)/\partial x_2 \\ \cdots \\ \partial f(x)/\partial x_n \end{bmatrix}$$

- **海森矩阵 (Hessian Matrix):** $\nabla^2 f(x)$ 是一个 $n \times n$ 的矩阵，其第 i 行第 j 列的元素

为 $\partial^2 f(x)/\partial x_i \partial x_j$。

$$\nabla^2 f(x) = \begin{bmatrix} \partial^2 f(x)/\partial x_1^2 & \partial^2 f(x)/\partial x_1\partial x_2 & \dots & \partial^2 f(x)/\partial x_1\partial x_n \\ \partial^2 f(x)/\partial x_2\partial x_1 & \partial^2 f(x)/\partial x_2^2 & \dots & \partial^2 f(x)/\partial x_2\partial x_n \\ \dots & \dots & \dots & \dots \\ \partial^2 f(x)/\partial x_n\partial x_1 & \partial^2 f(x)/\partial x_n\partial x_2 & \dots & \partial^2 f(x)/\partial x_n^2 \end{bmatrix}$$

- **二阶泰勒公式 (Taylor Expansion to the second order):** $f(x) \approx f(x_0) + \nabla f(x_0)^T(x - x_0) + (1/2)(x - x_0)^T\nabla^2 f(x_0)(x - x_0)$

练习题 2：线性方程组求解

设 A 为 $m \times n$ 矩阵：
(1) 假设 $m > n$，且 A^TA 正定，给出 $\min_x \|Ax - b\|_2^2$ 的公式。
(2) 假设 $m < n$，且 AA^T 可逆，给出以下问题的解的公式：
$\min. \|x\|_2^2 \ \text{s.t.}\ Ax = b$
解答：
(1) 当 $m > n$ 且 A^TA 正定 (此时 A 列满秩) 时，问题 $\min_x \|Ax - b\|_2^2$ 是一个最小二乘问题，其解可以通过求解正规方程得到。
 * 目标函数可以写成： $\|Ax - b\|_2^2 = (Ax - b)^T(Ax - b) = x^TA^TAx - 2b^TAx + b^Tb$
 * 对 x 求导并令其等于 0，得到正规方程： $A^TAx = A^Tb$
 * 由于 A^TA 正定，因此 A^TA 可逆，解得： $x = (A^TA)^{-1}A^Tb$
(2) 当 $m < n$ 且 AA^T 可逆 (此时 A 行满秩) 时，问题是一个带有等式约束的优化问题。
 * 构造拉格朗日函数： $L(x, \lambda) = \|x\|_2^2 + \lambda^T(Ax - b) = x^Tx + \lambda^T(Ax - b)$
 * 对 x 求导并令其等于 0，得到： $2x + A^T\lambda = 0$，即 $x = -(1/2)A^T\lambda$
 * 将 $x = -(1/2)A^T\lambda$ 代入约束条件 $Ax = b$，得到：$-(1/2)AA^T\lambda = b$
 * 由于 AA^T 可逆，解得：$\lambda = -2(AA^T)^{-1}b$
 * 将 λ 代回 x 的表达式，得到：$x = A^T(AA^T)^{-1}b$

练习题 3：二次型与曲线拟合

对曲线拟合问题，设有 N 个观测点 $x_1, \dots, x_n$，其观测值为 $t_1, \dots, t_n$。现用 M 次多项式拟合，记多项式的形式为：

$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$$

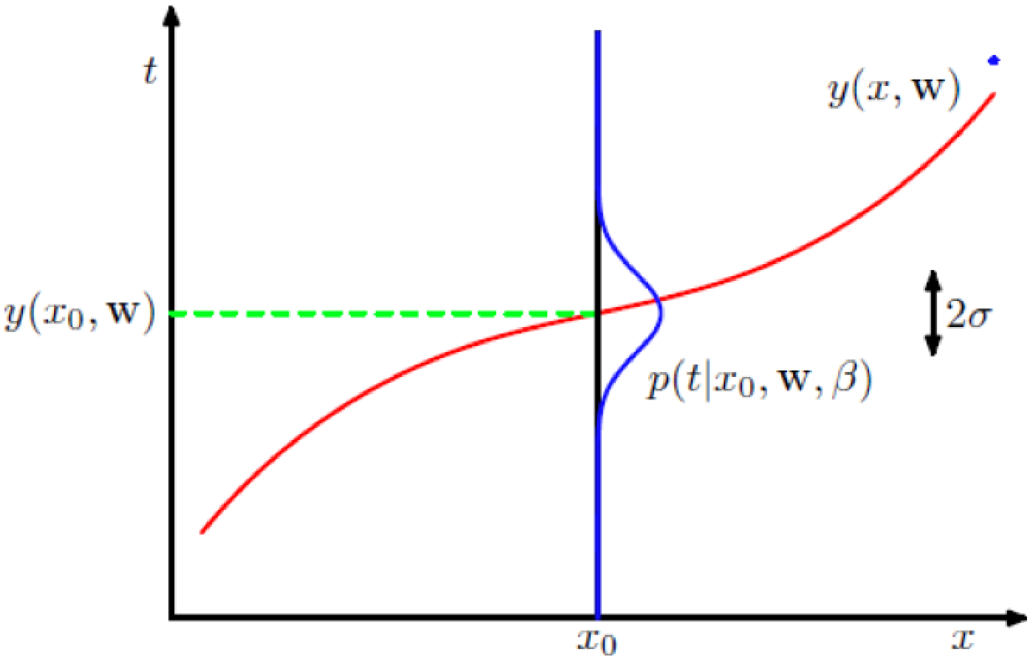
定义

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(y(x_i, \mathbf{w}) - t_i)^2$$

则多项式的系数为

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w})$$

现请用 MLE 方法推导出以上情形。附：用 Bayes 观点看待曲线拟合的示意图。

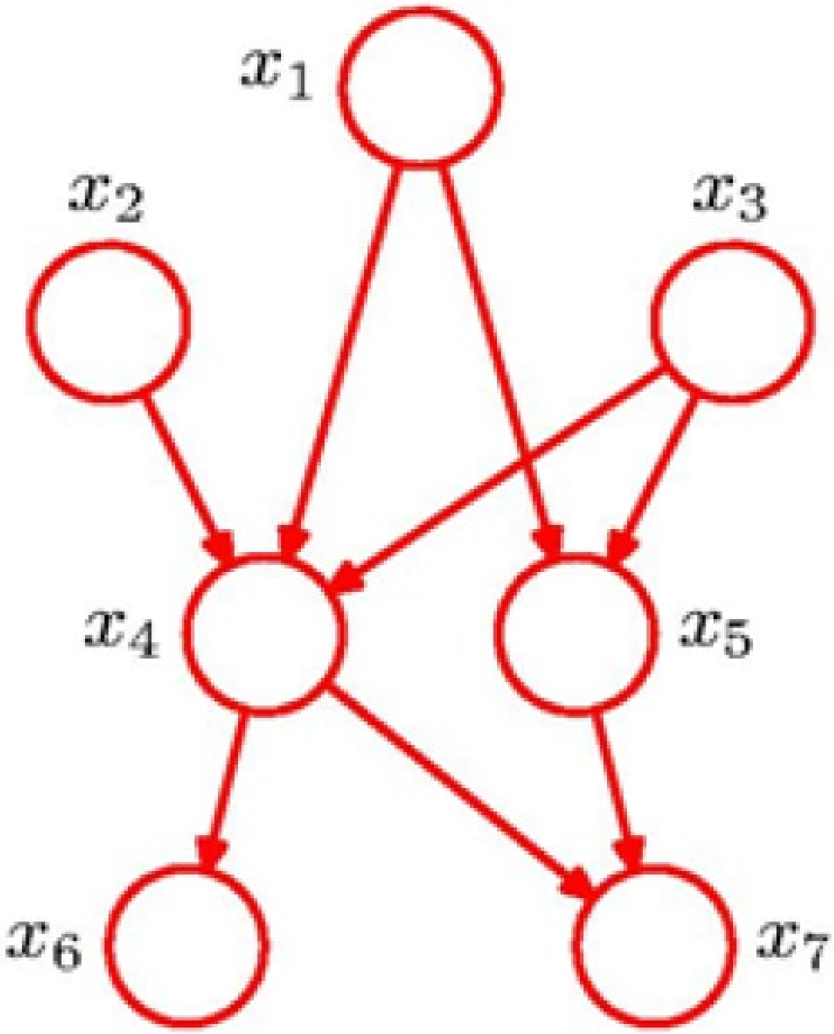


解答:

- **最大似然估计 (MLE):** 假设观测值 t_i 是由真实函数值 $y(x_i, w)$ 加上一个服从高斯分布的噪声 ε_i 得到的，即:
  - $t_i = y(x_i, w) + \varepsilon_i$
  - 其中 $\varepsilon_i \sim N(0, \sigma^2)$
- 因此，给定 x_i 和 w，t_i 服从均值为 $y(x_i, w)$，方差为 σ^2 的高斯分布:
  - $p(t_i|x_i, w, \sigma^2) = (1/(\sqrt{}(2\pi)\sigma))\exp(-(t_i - y(x_i, w))^2 / (2\sigma^2))$
- 假设观测值 $t_1, ..., t_n$ 相互独立，则似然函数为:
  - $p(t_1, ..., t_n|x_1, ..., x_n, w, \sigma^2) = \Pi_i\ p(t_i|x_i, w, \sigma^2)$
- 取对数似然函数:
  - $\ln p(t_1, ..., t_n|x_1, ..., x_n, w, \sigma^2) = \Sigma_i \ln p(t_i|x_i, w, \sigma^2) = -(N/2)\ln(2\pi) - N\ln(\sigma) - (1/(2\sigma^2))\Sigma_i (t_i - y(x_i, w))^2$
- 最大化对数似然函数等价于最小化 $(1/(2\sigma^2))\Sigma_i (t_i - y(x_i, w))^2$，由于 σ^2 是常数，因此等价于最小化:
  - $E(w) = (1/2)\Sigma_i (y(x_i, w) - t_i)^2$
- 因此，通过 MLE 方法，我们得到了与最小化 E(w) 相同的结果。

练习题 4：概率图模型

对如下的 PGM 图:

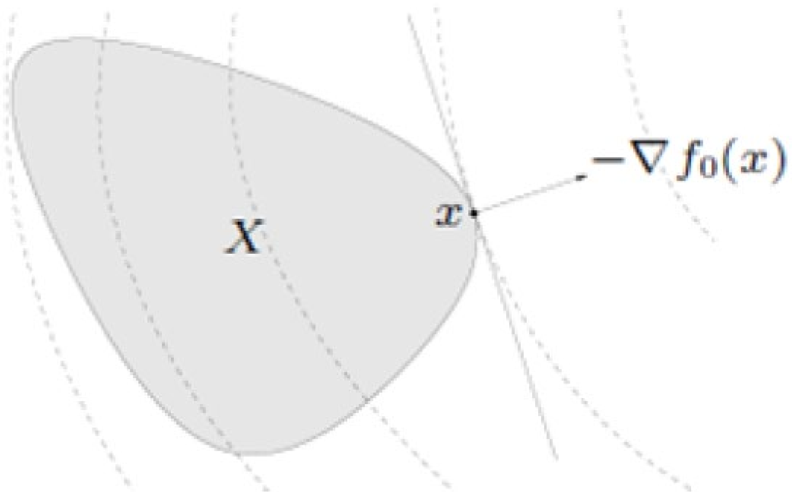


写出 $p(x_1, \ldots, x_7)$ 的公式。

解答:
根据图中的条件独立性假设，可以将联合概率分解为： $p(x_1, \ldots, x_7) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_6)p(x_5|x_2, x_3)p(x_6|x_4, x_5)p(x_7|x_3, x_5, x_6)$

练习题 5：最优性条件

试根据下图，写出 optimality condition (即 x 使 f(x) 达到极值，x 应满足的条件)。



解答:
根据图示，x 位于可行域 X 的边界上，且 $-\nabla f(x)$ 指向可行域外部。在最优点 x 处，目标函数 f(x) 的负梯度 $-\nabla f(x)$ 必须与指向可行域内部的法向量方向一致 (或负梯度与该点处的可行方向均呈钝角)，或者 $-\nabla f(x)$ 为零向量。
因此，最优性条件可以描述为:

- 对于任意可行方向 d (即从 x 出发指向可行域 X 内部的方向)，都有 $\nabla f(x)^T d \geq 0$，或者 $\nabla f(x) = 0$。

练习题 6：拉格朗日乘子法与 KKT 条件

1. 写出 $f: R^n \to R$ 为凸函数的定义。
2. 写出凸优化问题的标准形式。
3. 对以下问题:
4. $\min f(x) = x_1^2 + x_2 + 4$
5. s.t. $-x_1^2 - (x_2 + 4)^2 + 16 \geq 0$
6. $x_1 - x_2 - 6 \geq 0$

写出其相应的 Lagrange 函数，对偶函数和 KKT 条件，并求出相应的解。
解答:

(1) 凸函数的定义：对于函数 $f: R^n \to R$，如果对于任意 $x, y \in R^n$ 和任意 $\alpha \in [0, 1]$，都有： $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$ 则称 f(x) 为凸函数。
(2) 凸优化问题的标准形式： $\min f(x)$ s.t. $g_i(x) \leq 0, i = 1, \ldots, m$ $h_i(x) = 0, i = 1, \ldots, p$ 其中 f(x) 是凸函数，$g_i(x)$ 是凸函数，$h_i(x)$ 是仿射函数。

(3) 针对给定的优化问题:

- **Lagrange 函数:** $L(x, \lambda) = x_1^2 + x_2 + 4 + \lambda_1(-x_1^2 - (x_2 + 4)^2 + 16) + \lambda_2(x_1 - x_2 - 6)$
- **对偶函数:** $g(\lambda) = \inf_x L(x, \lambda)$
- **KKT 条件:**
    1. $\nabla_x L(x, \lambda) = 0$:
        - $2x_1^* - 2\lambda_1 x_1 + \lambda_2^* = 0$
        - $1 - 2\lambda_1(x_2 + 4) - \lambda_2^* = 0$
    2. $\lambda_1^* \geqslant 0$, $\lambda_2^* \geqslant 0$ (对偶可行性)
    3. $-x_1^2 - (x_2 + 4)^2 + 16 \geqslant 0$, $x_1^* - x_2^* - 6 \geqslant 0$ (原始可行性)
    4. $\lambda_1(-x_1^2 - (x_2^* + 4)^2 + 16) = 0$, $\lambda_2(x_1 - x_2^* - 6) = 0$ (互补松弛性)
- **求解:**
    - 分析 λ_1 和 λ_2 的取值情况:
        - **情况 1:** $\lambda_1^* = 0$, $\lambda_2^* = 0$: 此时 $x_1^* = 0$, 但无法满足第一个等式。
        - **情况 2:** $\lambda_1^* = 0$, $\lambda_2^* > 0$: 此时 $x_1^* - x_2^* - 6 = 0, x_1^* = 0$, $\lambda_2^* = -1$, 不满足 $\lambda_2^* > 0$。
        - **情况 3:** $\lambda_1^* > 0$, $\lambda_2^* = 0$: 此时 $-x_1^2 - (x_2 + 4)^2 + 16 = 0$, 且 $x_1^* = 0$, $x_2^* = -3$, $\lambda_1^* = 1/2$。验证得到: $-x_1^2 - (x_2 + 4)^2 + 16 = -1 + 16 > 0, x_1^* - x_2^* - 6 = -3 < 0$, 不满足原始可行性。
        - **情况 4:** $\lambda_1^* > 0$, $\lambda_2^* > 0$: 此时 $-x_1^2 - (x_2 + 4)^2 + 16 = 0$ 且 $x_1^* - x_2^* - 6 = 0$。联立方程求解得到 $x_1^* = (1 + \sqrt{33})/2, x_2^* = (-11 + \sqrt{33})/2$, $\lambda_1^* = (1 + \sqrt{33})/(1 + \sqrt{33} + 8)$, $\lambda_2^* = 16 + 2\sqrt{33} - (1 + \sqrt{33})(1 + \sqrt{33})/(1 + \sqrt{33} + 8)$。
    - 最终解需要验证情况 4 中 $\lambda_1^* > 0$, $\lambda_2^* > 0$ 是否成立。经过计算, 可以验证成立。

因此, 最优解为 $x_1^* = (1 + \sqrt{33})/2, x_2^* = (-11 + \sqrt{33})/2$, $\lambda_1^* = (1 + \sqrt{33})/(1 + \sqrt{33} + 8)$, $\lambda_2^* = 16 + 2\sqrt{33} - (1 + \sqrt{33})(1 + \sqrt{33})/(1 + \sqrt{33} + 8)$。

本章小结

本章通过六道练习题, 复习了向量空间、线性方程组、二次型、概率图模型、最优性条件以及拉格朗日乘子法与 KKT 条件等重要知识点。希望读者能够认真完成这些练习题, 加深对相关概念和方法的理解, 提升解决实际问题的能力。

结束语

本学期我完成了高级工程数学的学习！从线性代数的基石和微积分的精妙，到优化理论的探索和迭代方法的实践，我走过了一段充满挑战与发现的数学旅程。我从向量空间和矩阵的构建开始，逐步深入到线性变换、特征值与特征向量的奥秘之中。领略了微积分的魅力，掌握了序列、极限、导数矩阵以及泰勒级数等重要概念。在优化理论的殿堂里，探索了无约束优化和约束优化的奥秘，学习了梯度下降法、牛顿法、次梯度方法以及临近点算法等强大的工具，更领悟了拉格朗日乘子法和 KKT 条件在解决复杂问题中的精妙运用。

这趟旅程不仅仅是知识的积累，更是思维的锤炼。我学会了如何用严谨的数学语言描述问题，如何运用抽象的数学工具分析问题，更重要的是，学会了如何用数学的思维解决问题。这些知识和能力，将成为未来学习和工作中宝贵的财富。数学的海洋浩瀚无垠，本书仅仅撷取了其中的一小部分。我希望这本教材能够成为我继续探索数学世界的引航灯，激发我对数学的热情，引导我在未来的道路上不断前行。数学不仅仅是一门学科，更是一种思维方式，一种探索未知的工具。在未来的学习和工作中，我将灵活运用所学知识，不断创新，勇攀高峰！