

null

凸分析、非线性规划理论与标准算法

基于GIT的Dr. Arkadi Nemirovski(2018)”Optimization III” 幻灯片整理

2021 年 9 月 7 日

目录

1	引言	5
1.1	基本概念	5
1.2	表述优化问题	7
1.3	主要内容	13
2	凸集	15
2.1	定义与例子	15
2.2	凸组合与凸包	17
2.3	保持凸性的运算	21
2.4	凸集的拓扑性质	23
3	关于凸集的主要定理	27
3.1	Cuthedory定理	27
3.2	Hulley定理	28
3.3	多面集性与Fourier-Motzkin消元	33
3.4	超一定理	35
3.5	线性规划的对偶	38
3.6	凸集分离定理	41
3.7	支撑超平面和极点	50
3.8	多面集的结构及其在线性规划中的应用	53
4	凸函数	57
4.1	定义与例子	57
4.2	保凸运算	60
4.3	如何检测凸性	61
4.4	凸函数的性质	63
4.5	凸函数的次梯度及其运算法则	67
5	拉格朗日对偶和鞍点最优性条件	70
5.1	凸规划	70
5.2	拉格朗日对偶	72
5.3	鞍点与最优性条件	74
5.4	鞍点的存在性	78
6	数学规划的最优性条件	81
6.1	凸规划最优性条件	81

6.2	二阶最优性条件	84
6.3	敏感性分析	91
6.4	应用举例	92
7	最优化算法简介	95
7.1	算法概述	95
7.1.1	草垛中找针有多困难呢?	95
7.1.2	MP算法的分类与收敛速度	97
7.1.3	可解的MP——凸规划	99
7.2	线搜索	99
7.2.1	零阶线搜索与黄金搜索	100
7.2.2	一阶线搜索之二分法	101
7.2.3	非精确线搜索之回溯Armijo线搜索算法	102
8	无约束最小化方法：梯度下降与牛顿法	104
8.1	梯度下降	104
8.1.1	收敛性	105
8.1.2	收敛速度	106
8.1.3	凸的情况	108
8.1.4	小结	110
8.2	牛顿法	112
9	无约束最小化方法：牛顿法的修正	116
9.1	二次正则化牛顿法	116
9.1.1	传统修正：变度量方案(scheme)	118
9.1.2	共轭梯度法	121
9.1.3	非二次扩展	125
9.2	拟牛顿法	127
11	约束最小化方法：惩罚/障碍法	129
11.1	惩罚/障碍策略	129
11.2	探讨惩罚策略	130
11.3	用障碍法求解凸规划	132
11.4	求解LP的原始对偶内点法	135
12	约束最小化方法：增广拉格朗日方法	138
12.1	局部拉格朗日对偶	138
12.2	整合起来：增广拉格朗日方案	141
12.3	纳入不等式约束	142

12.4	特殊情况：增广拉格朗日法	143
13	约束最小化方法：逐步二次规划	147
13.1	牛顿法求解非线性方程组	147
13.2	牛顿位移的结构和解释	149
13.3	一般约束的情况	151
13.3.1	基本SQP法	151
13.3.2	确保全局收敛	151

1 引言

做最优决策是人类最基本的渴望之一. 只要候选决策、设计限制和设计目标能被恰当地量化, 确定最优决策就会产生一个**最优化问题**(optimization problem). 最典型的最优化问题是一个**数学规划**(mathematical programming)问题:

$$\begin{aligned} & \text{minimize } f(\mathbf{x}) \\ & \text{subject to } h_i(\mathbf{x}) = 0, i = 1, \dots, m \\ & \quad g_j(\mathbf{x}) \leq 0, j = 1, \dots, k \\ & \quad \mathbf{x} \in X, \end{aligned} \tag{MP}$$

这里 $f, h_i, g_j : \mathbb{R}^n \rightarrow \mathbb{R}$, f 是要极小化的**目标函数**(objective function), 代表着与决策相关联的损失(负的为利润); $h_i(\mathbf{x}) = 0$ 和 $g_j(\mathbf{x}) \leq 0$ 分别是**等式约束**和**不等式约束**, 代表着对有意义决策的限制, 比如平衡和状态方程, 关于资源的界等; $X \subseteq \mathbb{R}^n$ 是**定义域**(domain). 这样, $\mathbf{x} = (x_1, \dots, x_n)$ 是 n -维向量, 且默认向量的表述是列形式. 称元素 x_1, \dots, x_n 是(MP)的**决策变量**(decision variable). 集合

$$\Omega = X \cap \{\mathbf{x} \in \mathbb{R}^n : h_i(\mathbf{x}) = 0, i = 1, \dots, m, g_j(\mathbf{x}) \leq 0, j = 1, \dots, k\}$$

是**可行域**(feasible region).

1.1 基本概念

如上面表述所表明的, 对(MP)的**全局极小点**(global minimizer)感兴趣, 它定义为: 点 $\mathbf{x}^* \in \Omega$ 使得对所有的 $\mathbf{x} \in \Omega$ 有 $f(\mathbf{x}^*) \leq f(\mathbf{x})$. 需要注意的是, 这样的 \mathbf{x}^* 可能不存在, 比如取 $f(x) = 1/x, \Omega = \mathbb{R}_{++} \equiv \{x \in \mathbb{R} : x > 0\}$. (MP)的最优值定义为集合 $\{f(\mathbf{x}) : \mathbf{x} \in \Omega\}$ 的**最大下界**(greatest lower bound)或者**下确界**(infimum)(参看讲义C), 将其记为

$$v^* = \inf\{f(\mathbf{x}) : \mathbf{x} \in \Omega\}.$$

注意如果 \mathbf{x}^* 是(MP)的全局极小点, 则自然地有 $v^* = f(\mathbf{x}^*)$. 然而, 即使问题(MP)没有全局极小点, v^* 也可以是有限的. 比如, 当 $f(x) = 1/x$ 且 $\Omega = \mathbb{R}_{++}$, 有 $v^* = 0$.

一个相关的概念是(MP)的**局部极小点**(local minimizer), 它定义为: 点 $\mathbf{x}' \in \Omega$ 使得存在某 $\delta > 0$, 使得对所有 $\mathbf{x} \in \Omega \cap B(\mathbf{x}', \delta)$ 有 $f(\mathbf{x}') \leq f(\mathbf{x})$ 成立. 这里,

$$B(\mathbf{x}', \delta) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{x}'\|_2 \leq \delta\}$$

是半径为 $\delta > 0$, 中心在 \mathbf{x}' 的**欧氏球**(Euclidean ball)(回忆对 $\mathbf{x} \in \mathbb{R}^n$, \mathbf{x} 的2-范数定义为 $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{\mathbf{x}^T \mathbf{x}}$). 注意到全局极小点一定是局部极小点, 但是反之不必成立.

首先说明问题(MP)是相当一般的. 比如, 当 $\Omega = \mathbb{R}^n$, 得到无约束优化(unconstrained optimization) 问题; 当定义域 X 是一个离散集, 比如所有的整数或者0/1变量, 得到离散优化(discrete optimization) 问题. 与此相对, 在连续优化(continuous optimization)中将聚焦于 Ω 是一个“连续的”集合, 比如整个 \mathbb{R}^n , 一个盒子 $\{x : a \leq x \leq b\}$, 或者单纯形 $\{x \geq 0 : \sum_{j=1}^n x_j = 1\}$ 等, 目标和约束(至少)在 X 上是连续的. 其它重要的优化问题类包括:

- 线性规划(Linear Programming, LP)问题: 这里, f 是线性函数, 即形如

$$f(\mathbf{x}) = c_1x_1 + c_2x_2 + \cdots + c_nx_n \equiv \mathbf{c}^T \mathbf{x}$$

的函数, 其中 $\mathbf{c} = (c_1, \cdots, c_n) \in \mathbb{R}^n$; 且 X 是由线性不等式定义的集合, 形如

$$X = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_i^T \mathbf{x} \leq b_i, i = 1, \cdots, m\} \quad (1.1)$$

其中 $\mathbf{a}_1, \cdots, \mathbf{a}_m \in \mathbb{R}^n, b_1, \cdots, b_m \in \mathbb{R}$. 用更紧凑的记号, 可以将一个线性规划问题写成

$$\text{minimize } \mathbf{c}^T \mathbf{x}$$

$$\text{subject to } \mathbf{A}\mathbf{x} \leq \mathbf{b}$$

其中 \mathbf{A} 是 $m \times n$ 矩阵, 它的第 i 行是 \mathbf{a}_i^T , 且 $\mathbf{b} = (b_1, \cdots, b_m)$ 是 m -维列向量(对任意两个向量 $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, 不等式 $\mathbf{u} \leq \mathbf{v}$ 意味着对 $i = 1, \cdots, n$, 有 $u_i \leq v_i$ 成立). 诚然如在今后的课程中将看到的那样, 能非常有效地求解线性规划. 与此相对的, 非线性连续优化(Nonlinear Continuous Optimization)指 f, h_i, g_j 中至少有一个是非线性的且 X 是连续的.

- 二次规划(Quadratic Programming, QP) 问题: 这里, X 形如(1.1), 且 f 是一个二次函数, 即形如

$$f(\mathbf{x}) = \sum_{i=1}^n \sum_{j=1}^n Q_{ij}x_i x_j + \sum_{j=1}^n c_j x_j \equiv \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{c}^T \mathbf{x}$$

其中 $\mathbf{Q} = (q_{ij})$ 是 $n \times n$ 矩阵, $\mathbf{c} \in \mathbb{R}^n$. 注意不失一般性, 可以假设 \mathbf{Q} 是对称的. 这源于事实:

$$\mathbf{x}^T \mathbf{Q} \mathbf{x} = \mathbf{x}^T \left(\frac{\mathbf{Q} + \mathbf{Q}^T}{2} \right) \mathbf{x}.$$

- 半定规划(Semidefinite Programming, SDP) 问题: 给定 $n \times n$ 对称矩阵 \mathbf{Q} (记作 $\mathbf{Q} \in \mathcal{S}^n$), 称 \mathbf{Q} 是半正定的(记作 $\mathbf{Q} \succeq \mathbf{0}$ 或者当要显式表示维数时记作 $\mathbf{Q} \in \mathcal{S}_+^n$) 如果 $\mathbf{x}^T \mathbf{Q} \mathbf{x} \geq 0$ 对所有的 $\mathbf{x} \in \mathbb{R}^n$ 成立. 设 $\mathbf{A}_1, \cdots, \mathbf{A}_m, \mathbf{C} \in \mathcal{S}^n$, 且设 $b_1, \cdots, b_m \in \mathbb{R}$. 考虑优化问题

$$\text{minimize } \mathbf{b}^T \mathbf{y}$$

$$\text{subject to } \mathbf{C} - \sum_{i=1}^m y_i \mathbf{A}_i \succeq \mathbf{0} \quad (1.2)$$

$$\mathbf{y} \in \mathbb{R}^m.$$

将问题(1.2)的约束(可以表述成 $-\sum_{i=1}^m y_i \mathbf{A}_i \succeq -\mathbf{C}$)称作**线性矩阵不等式**(linear matrix inequality, LMI), 因为定义为 $M: \mathbb{R}^m \rightarrow S^n$ 的矩阵值函数 $M(\mathbf{y}) = -\sum_{i=1}^m y_i \mathbf{A}_i$ 是线性的(即 M 满足: 对任何 $\mathbf{y}, \mathbf{z} \in \mathbb{R}^m$ 和 $\alpha, \beta \in \mathbb{R}$ 有 $M(\alpha\mathbf{y} + \beta\mathbf{z}) = \alpha M(\mathbf{y}) + \beta M(\mathbf{z})$ 成立). 问题(1.2)是所谓的**半定规划**(Semidefinite Programming, SDP) 问题. 它的可行域是

$$X = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{C} - \sum_{i=1}^m y_i \mathbf{A}_i \succeq \mathbf{0}\}.$$

作为一个常规练习, 可以证明能将优化问题

$$\begin{aligned} & \text{minimize } \mathbf{C} \bullet \mathbf{Z} \equiv \sum_{i=1}^n \sum_{j=1}^n c_{ij} z_{ij} \\ & \text{subject to } \mathbf{A}_i \bullet \mathbf{Z} = b_i, i = 1, \dots, m \\ & \mathbf{Z} \succeq \mathbf{0} \end{aligned} \quad (1.3)$$

表述成(1.2)的形式, 因此(1.3)也是一个半定规划问题. 下面确定(1.3)的可行域. 因为 \mathbf{Z} 是对称的, 所以它由对角线和对角线之上的元素完全确定. 因此, 可将(1.3)的可行域表示为

$$X = \{(z_{11}, z_{12}, \dots, z_{nn}) \in \mathbb{R}^{n(n+1)/2} : \mathbf{A}_i \bullet \mathbf{Z} = b_i, i = 1, \dots, m; \mathbf{Z} \succeq \mathbf{0}\}.$$

与线性规划类似, 也可以有效地求解SDP问题.

- **多项式优化**(Polynomial Optimization, PO) 问题: 这里, f 是一个 d 次多项式. 换句话说, 它可被表示成

$$f(\mathbf{x}) = \sum_{|\alpha| \leq d} f_{\alpha} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$$

其中 $\alpha = (\alpha_1, \dots, \alpha_n), \alpha_i \in \mathbb{N}, i = 1, \dots, n, |\alpha| = \sum_{i=1}^n \alpha_i, f_{\alpha}$ 是 $x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ 的系数. 集合 X 是由**多项式不等式**定义的, 它形如

$$X = \{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) \geq 0, i = 1, \dots, m\}$$

其中 $g_1, \dots, g_m: \mathbb{R}^n \rightarrow \mathbb{R}$ 也是实值多项式. 通常, PO问题非常难于求解. 然而, 在一些温和的假设下, 可以用一系列SDP 问题来近似, 至少在理论上是这样的. 该课程不会深入讨论多项式优化. 推荐感兴趣的读者进一步阅读 [3].

1.2 表述优化问题

上面提到的问题类涵盖了非常广泛的应用. 然而, 为了将一个特定的应用转换成一个形如(MP)的问题, 需要首先识别出数据和决策变量, 然后表述目标函数和约束. 下面用一些例子说明这个过程.

例1.1. 一个航空流量管制问题. 假设有 n 架飞机在首都机场降落. 飞机 i 将在时间区间 $[a_i, b_i]$ 内到达机场, $i = 1, \dots, n$. 为了简单, 假设飞机到达的次序为 $1, 2, \dots, n$. 出于安全考虑, 机场塔台想要最大化所谓的**最短汇合时间**(shortest metering time), 即所有两个相邻飞机之间到港间隔的最小值. 那么, 机场应该如何为每一架飞机指派到达时间?

这里, 决策变量是飞机的到达时间, 记为 t_1, \dots, t_n . 那么, 有下面的最优化问题:

$$\begin{aligned} & \text{maximize } \min_{1 \leq j \leq n-1} (t_{j+1} - t_j) \\ & \text{subject to } a_i \leq t_i \leq b_i, i = 1, \dots, n \\ & \quad t_i \leq t_{i+1}, i = 1, \dots, n-1 \end{aligned} \tag{1.4}$$

不能立即确定(1.4)是否能够被表述成线性规划, 但是可进行如下操作. 设 z 是一个新的决策变量. 则可将(1.4)重写为

$$\begin{aligned} & \text{maximize } z \\ & \text{subject to } t_{i+1} - t_i \geq z, i = 1, \dots, n-1 \\ & \quad a_i \leq t_i \leq b_i, i = 1, \dots, n \\ & \quad t_i \leq t_{i+1}, i = 1, \dots, n-1. \end{aligned}$$

这是一个线性规划. 需要指出的是, 仅当**极大化** $\min_{1 \leq j \leq n-1} (t_{j+1} - t_j)$, 而非极小化时, 上面的表述才是正确的. 具体地, 下面的问题

$$\begin{aligned} & \text{minimize } \min_{1 \leq j \leq n-1} (t_{j+1} - t_j) \\ & \text{subject to } a_i \leq t_i \leq b_i, i = 1, \dots, n \\ & \quad t_i \leq t_{i+1}, i = 1, \dots, n-1 \end{aligned} \tag{1.5}$$

和

$$\begin{aligned} & \text{minimize } z \\ & \text{subject to } t_{i+1} - t_i \geq z, i = 1, \dots, n-1 \\ & \quad a_i \leq t_i \leq b_i, i = 1, \dots, n \\ & \quad t_i \leq t_{i+1}, i = 1, \dots, n-1. \end{aligned} \tag{1.6}$$

不是等价的, 因为(1.5)的最优值是有限的(事实上, 它总是非负的), 而(1.6)的最优值是 $-\infty$.

另一个利用优化技术的流量管控问题的应用参见 [1].

例1.2. 一个数据拟合问题. 前一个例子说明, 有时候能够通过某种变换将一个优化问题转换成一个线性规划. 这里有另一个例子来说明这种可能. 假设给了 m 对

数据 (\mathbf{a}_i, b_i) , 其中 $\mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}, i = 1, \dots, m$, 且 $m \geq n + 1$. 假设这些数据原本是由仿射函数生成的, i.e. 一个函数 $f: \mathbb{R}^n \rightarrow \mathbb{R}$, 形如 $f(\mathbf{y}) = \mathbf{x}^T \mathbf{y} + t$, 其中 $\mathbf{x} \in \mathbb{R}^n, t \in \mathbb{R}$ 是给定的. 然而, 该函数的输出通常会被加性噪声污染. 这样, \mathbf{a}_i 和 b_i 之间的关系用 $b_i = \mathbf{a}_i^T \mathbf{x} + t + \epsilon_i$ 会更好些, 其中 $\mathbf{x} \in \mathbb{R}^n, t \in \mathbb{R}$ 是仿射函数的参数, ϵ_i 是第 i 个测量中的噪声. 目的是确定能最好地拟合数据的仿射函数的参数 $\mathbf{x} \in \mathbb{R}^n$ 和 $t \in \mathbb{R}$. 为了度量拟合的优良性, 可以尝试极小化某种误差度量. 一个特别的度量是残余量误差的1-范数(1-norm of the residual errors), 其定义为

$$\Delta_1 = \sum_{i=1}^m |b_i - \mathbf{a}_i^T \mathbf{x} - t| = \|\mathbf{b} - \mathbf{A}\mathbf{x} - t\mathbf{e}\|_1$$

其中 \mathbf{A} 是 $m \times n$ 矩阵, 它的第 i 行是 $\mathbf{a}_i^T, \mathbf{e} \in \mathbb{R}^m$ 是分量全为1的向量. 换句话说, 优化问题是

$$\min_{\mathbf{x} \in \mathbb{R}^n, t \in \mathbb{R}} \sum_{i=1}^m |b_i - \mathbf{a}_i^T \mathbf{x} - t|. \quad (1.7)$$

这里, 目标函数是非线性的. 然而, 能够将问题(1.7)转换成如下的LP. 首先引入 m 个新的决策变量 $z_1, \dots, z_m \in \mathbb{R}$. 接着, 不难看到(1.7)等价于如下的LP:

$$\begin{aligned} & \text{minimize } \sum_{i=1}^m z_i \\ & \text{subject to } b_i - \mathbf{a}_i^T \mathbf{x} - t \leq z_i, i = 1, \dots, m \\ & \quad -b_i + \mathbf{a}_i^T \mathbf{x} + t \leq z_i, i = 1, \dots, m. \end{aligned}$$

现在, 如果想要极小化残余向量的2-范数, 情况如何? 换句话说, 想要求解如下问题:

$$\min_{\mathbf{x} \in \mathbb{R}^n, t \in \mathbb{R}} \Delta_2 = \|\mathbf{b} - \mathbf{A}\mathbf{x} - t\mathbf{e}\|_2^2 = \sum_{i=1}^m (b_i - \mathbf{a}_i^T \mathbf{x} - t)^2. \quad (1.8)$$

可以说明这是一个特别简单的QP. 事实上, 因为(1.8)是一个目标函数可微(differentiable)的无约束优化问题, 可以用微积分技术求解它. 的确, 如果 $\bar{\mathbf{A}}$ 列满秩(因此 $\bar{\mathbf{A}}^T \bar{\mathbf{A}}$ 可逆), 可给出(1.8)的(唯一)最优解 $(\mathbf{x}^*, t^*) \in \mathbb{R}^n \times \mathbb{R}$ 的显式表示

$$\begin{bmatrix} \mathbf{x}^* \\ t^* \end{bmatrix} = (\bar{\mathbf{A}}^T \bar{\mathbf{A}})^{-1} \bar{\mathbf{A}}^T \mathbf{b}, \text{ 其中 } \bar{\mathbf{A}} = \begin{bmatrix} \mathbf{a}_1^T & 1 \\ \vdots & \vdots \\ \mathbf{a}_m^T & 1 \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}.$$

如果 $\bar{\mathbf{A}}$ 不是列满秩的, 则可以证明对任何 $\mathbf{z} \in \mathbb{R}^{n+1}$, 向量

$$\begin{bmatrix} \mathbf{x}^* \\ t^* \end{bmatrix} = (\bar{\mathbf{A}}^T \bar{\mathbf{A}})^\dagger \bar{\mathbf{A}}^T \mathbf{b} + (\mathbf{I} - (\bar{\mathbf{A}}^T \bar{\mathbf{A}})^\dagger \bar{\mathbf{A}}^T \bar{\mathbf{A}}) \mathbf{z}.$$

是(1.8)的最优解. 这里 $(\cdot)^\dagger$ 表示矩阵的广义逆. 值得注意的是矩阵 $\mathbf{I} - (\bar{\mathbf{A}}^T \bar{\mathbf{A}})^\dagger \bar{\mathbf{A}}^T \bar{\mathbf{A}}$ 就是到 $\bar{\mathbf{A}}^T \bar{\mathbf{A}}$ 的零空间上的正交投影. 特别地, 当 $\bar{\mathbf{A}}$ 不是列满秩的时候, $\bar{\mathbf{A}}^T \bar{\mathbf{A}}$ 的零空间是非平凡的.

在上面的讨论中, 假设观测数据的数量 $m \geq n + 1$. 然而, 在许多现代应用中(比如生物医学图像和基因表达分析), 观测数量要比参数数量少得多. 这样, 人们能找到许多对参数 $(\bar{\mathbf{x}}, \bar{t}) \in \mathbb{R}^n \times \mathbb{R}$ 来完美地拟合数据, 即对 $i = 1, \dots, m$ 有 $b_i = \bar{\mathbf{a}}_i^T \bar{\mathbf{x}} + \bar{t}$. 为了使得数据拟合问题是有意义的, 需要施加额外的假设. 一种直观且流行的假设是真实的适合输入-输出关系的参数数量是小的. 换句话说, 参数向量 $\mathbf{x} \in \mathbb{R}^n$ 的大部分分量应该是零, 尽管事先并不知道是哪些分量. 例如, 人们可以考虑如下的约束优化方法:

$$\begin{aligned} & \text{minimize } \|\mathbf{b} - \mathbf{A}\mathbf{x} - t\mathbf{e}\|_2^2 \\ & \text{subject to } \|\mathbf{x}\|_0 \leq K, \\ & \mathbf{x} \in \mathbb{R}^n, t \in \mathbb{R}. \end{aligned} \tag{1.9}$$

这里, $\|\mathbf{x}\|_0$ 是参数向量 $\mathbf{x} \in \mathbb{R}^n$ 中的非零元素个数, $K \geq 0$ 是一个由用户定义的用于控制 \mathbf{x} 的稀疏度的阈值. 另外, 人们也可以考虑如下的惩罚法:

$$\min_{\mathbf{x} \in \mathbb{R}^n, t \in \mathbb{R}} \{ \|\mathbf{b} - \mathbf{A}\mathbf{x} - t\mathbf{e}\|_2^2 + \mu \|\mathbf{x}\|_0 \}, \tag{1.10}$$

其中 $\mu > 0$ 是一个惩罚参数. 然而, 由于函数 $\mathbf{x} \rightarrow \|\mathbf{x}\|_0$ 的组合本质, 上面的两种表述求解起来都具有组合上的困难. 事实上, 可以证明, 从正式意义上, 不可能存在求解(1.9)和(1.10)的有效算法. 为了得到一种更可处理的表述, 一种应用很广泛的方法是用 $\|\cdot\|_1$ 代替 $\|\cdot\|_0$. 将在后面的课程看到: 从计算的角度讲, 这是一个好想法, 将在后面的课程中看到其原因. 对于目前来讲, 应该注意到这样的一种方法改变了原始问题, 并且一个自然的问题是: 在原始问题的解与修正后问题的解之间是否存在一种对应. 在过去大约十几年里, 高维统计和压缩感知领域已经广泛地研究了这个问题. 建议感兴趣的读者阅读书 [2]以获取细节和进一步的参考文献.

现在反思上面的例子. 直观上, 一个线性问题(比如说LP)应该比非线性问题更容易, 一个可微问题应该比不可微问题更容易. 然而, 上面的例子表明, 也有例外情况. 的确, 尽管2-范数问题(1.8)是一个QP, 但它的最优解有一个优美的刻画, 而对应的1-范数问题(1.7)不具备这种特征. 另一方面, 尽管目标函数(1.7)是不可微的, 这个问题仍然可以通过线性规划来有效求解. 还有, 从问题(1.9)和(1.10)也看到包含简单的约束或者目标有可能使得原本易于求解的问题(即问题(1.8))变成不可处理的.

综合上面的讨论, 一个很自然的问题是: 使得一个优化问题变得困难的原因是什么? 尽管要不以偏概全地回答该问题很困难, 但是至少可以识别一种可

能的困难来源. 看起来很相似的问题(1.8)和(1.9)之间的区别是: (1.8)是所谓的凸优化问题(convex optimization problem), 当然问题(1.9)不是. 定义凸性的概念, 并在后面详细地研究它.

例1.3. 特征值优化. 假设给了 k 个 $n \times n$ 的对称矩阵 $\mathbf{A}_1, \dots, \mathbf{A}_k$. 考虑函数 $\mathbf{A} : \mathbb{R}^k \rightarrow \mathbb{R}^{n \times n}$.

$$\mathbf{A}(\mathbf{x}) = \sum_{i=1}^k x_i \mathbf{A}_i.$$

注意到由定义, 对任何 $\mathbf{x} \in \mathbb{R}^k$, 矩阵 $\mathbf{A}(\mathbf{x})$ 是对称矩阵. 现在, 在实际中经常遇到的一个问题是: 选择 $\mathbf{x} \in \mathbb{R}^k$ 使得 $\mathbf{A}(\mathbf{x})$ 的最大特征值被最小化(比如, 详见 [4]). 可将该问题表述成一个SDP. 为此, 需要下面的结论.

命题1 设 \mathbf{A} 是一个 $n \times n$ 对称矩阵, 且令 $\lambda_{\max}(\mathbf{A})$ 表示 \mathbf{A} 的最大特征值. 则, 有 $t\mathbf{I} \succeq \mathbf{A}$ 当且仅当 $t \geq \lambda_{\max}(\mathbf{A})$.

证明 假设 $t\mathbf{I} \succeq \mathbf{A}$, 或者等价地, $t\mathbf{I} - \mathbf{A} \succeq \mathbf{0}$. 则对任何 $\mathbf{u} \in \mathbb{R}^n \setminus \{0\}$, 有 $\mathbf{u}^T(t\mathbf{I} - \mathbf{A})\mathbf{u} = t\mathbf{u}^T\mathbf{u} - \mathbf{u}^T\mathbf{A}\mathbf{u} \geq 0$, 或者等价地

$$t \geq \frac{\mathbf{u}^T\mathbf{A}\mathbf{u}}{\mathbf{u}^T\mathbf{u}}.$$

因为该事实对任意的 $\mathbf{u} \in \mathbb{R}^n \setminus \{0\}$ 都成立, 所以有

$$t \geq \max_{\mathbf{u} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{u}^T\mathbf{A}\mathbf{u}}{\mathbf{u}^T\mathbf{u}}. \quad (1.11)$$

由Courant-Fischer定理, (1.11)的右边恰好是 $\lambda_{\max}(\mathbf{A})$.

逆向进行上面的讨论, 也可以到逆命题成立. 定理得证. \square

由命题1, 可将上面的特征值优化问题表述为

$$\begin{aligned} & \text{minimize } t \\ & \text{subject to } t\mathbf{I} - \mathbf{A}(\mathbf{x}) \succeq \mathbf{0}. \end{aligned} \quad (1.12)$$

因为函数 $\mathbb{R}^n \times \mathbb{R} \ni (\mathbf{x}, t) \rightarrow t\mathbf{I} - \mathbf{A}(\mathbf{x})$ 关于 (\mathbf{x}, t) 是线性的, 约束是一个LMI. 因此问题(1.12)是一个SDP.

鉴于半定规划比较抽象, 下面给出一个具体例子来帮助理解. 在(1.2)中, 考虑 $n = 3, m = 2$. 令

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 3 & 7 \\ 1 & 7 & 5 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 0 & 2 & 8 \\ 2 & 6 & 0 \\ 8 & 0 & 4 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 9 & 0 \\ 3 & 0 & 7 \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} -11 \\ -19 \end{pmatrix}.$$

这时, 对应的问题(1.2)为

$$\begin{aligned} & \text{minimize} \quad -11y_1 - 19y_2 \\ & \text{subject to} \quad \begin{bmatrix} 1 - y_1 & 2 - 2y_2 & 3 - y_1 - 8y_2 \\ 2 - 2y_2 & 9 - 3y_1 - 6y_2 & -7y_1 - 0y_2 \\ 3 - y_1 - 8y_2 & -7y_1 - 0y_2 & 7 - 5y_1 - 4y_2 \end{bmatrix} \succeq \mathbf{0}. \end{aligned}$$

此外, 利用矩阵半正定当且仅当它的所有主子式皆大于等于零的事实, 可将上述问题等价表述为

$$\begin{aligned} & \text{minimize} \quad -11y_1 - 19y_2 \\ & \text{subject to} \quad 1 - y_1 \geq 0, 9 - 3y_1 - 6y_2 \geq 0, 7 - 5y_1 - 4y_2 \geq 0 \\ & \quad (1 - y_1)(9 - 3y_1 - 6y_2) - (2 - 2y_2)^2 \geq 0 \\ & \quad (1 - y_1)(7 - 5y_1 - 4y_2) - (3 - 1y_1 - 8y_2)^2 \geq 0 \\ & \quad (9 - 3y_1 - 6y_2)(7 - 5y_1 - 4y_2) - (7y_1)^2 \geq 0 \\ & \quad \left| \begin{bmatrix} 1 - 1y_1 - 0y_2 & 2 - 0y_1 - 2y_2 & 3 - 1y_1 - 8y_2 \\ 2 - 0y_1 - 2y_2 & 9 - 3y_1 - 6y_2 & 0 - 7y_1 - 0y_2 \\ 3 - 1y_1 - 8y_2 & 0 - 7y_1 - 0y_2 & 7 - 5y_1 - 4y_2 \end{bmatrix} \right| \geq 0 \end{aligned}$$

这个例子表明, SDP可以等价处理一类凸的非线性规划问题. 需要注意的是, 所有顺序主子式大于或者等于零是不能保证一个对称矩阵是半正定的. 比如

$$\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$$

就是一个反例.

下面用一个例子说明形如(1.3)的问题可以表述成(1.2)的问题. 这里的做法对于将一般的问题(1.3)表述成(1.2)也是适用的. 考虑 $n = 2, m = 1$. 令

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad b = 1.$$

将上述数据代入问题(1.3), 得

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix}.$$

对应的问题是

$$\begin{aligned} & \text{minimize } z_{11} + z_{22} \\ & \text{subject to } z_{11} + 4z_{12} + 3z_{22} = 1 \\ & \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix} \succeq \mathbf{0}. \end{aligned}$$

该问题可等价表述为

$$\begin{aligned} & \text{minimize } z_{11} + z_{22} \\ & \text{subject to } \begin{bmatrix} z_{11} & z_{12} & 0 & 0 \\ z_{21} & z_{22} & 0 & 0 \\ 0 & 0 & z_{11} + 4z_{12} + 3z_{22} - 1 & 0 \\ 0 & 0 & 0 & -z_{11} - 4z_{12} - 3z_{22} + 1 \end{bmatrix} \succeq \mathbf{0}. \end{aligned} \quad (1.13)$$

令

$$\mathbf{A}_1 = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \mathbf{A}_2 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}, \mathbf{A}_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \mathbf{y} = (z_{11}, z_{12}, z_{22})$$

将这些参数代入(1.2), 得到的问题即为(1.13).

1.3 主要内容

最优化理论的**数学基础**是凸分析. 凸分析是实分析和几何的特定组合, 强调与凸性相关的概念. 该讲义的目标是

- (a) 给出与最优化理论相关的凸分析概念和结论及其在优化中的应用;
- (b) 给出连续优化的基本理论, 强调最优解的存在性及唯一性, 特别是最优解的刻画(比如必要和/或者充分的最优性条件).
- (c) 构建(近似)连续优化问题最优解的传统算法;

- (d) 研究 f 和 Ω 的结构如何影响求解(MP) 的能力，给出一些适用于大规模问题的算法.

2 凸集

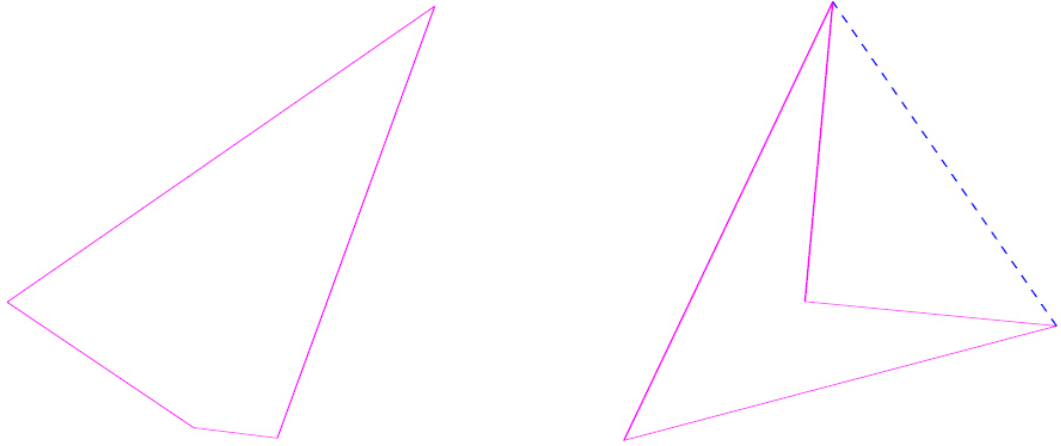
如讲义1中简要提到的, 凸性的概念对于最优化的理论和方法都非常重要. 在深入地讨论最优化中与凸性相关的内容之前, 首先引入凸集和凸函数的概念, 并学习它们的重要性质.

2.1 定义与例子

定义2.1. 称集合 $X \subseteq \mathbb{R}^n$ 是凸的(convex), 如果它包含以自己中任一对点 \mathbf{x}, \mathbf{y} 为顶点的线段, 即

$$\mathbf{x}, \mathbf{y} \in X \Rightarrow (1 - \theta)\mathbf{x} + \theta\mathbf{y} \in X \quad \forall \theta \in [0, 1].$$

按语2.1. 当 θ 取遍 $[0, 1]$, 点 $(1 - \theta)\mathbf{x} + \theta\mathbf{y} \equiv \mathbf{x} + \theta(\mathbf{y} - \mathbf{x})$ 跑遍线段 $[\mathbf{x}, \mathbf{y}]$.



(a) 以红线为边界的 \mathbb{R}^2 中的集合是凸的 (b) 以红线为边界的 \mathbb{R}^2 中的集合是非凸的

图 2.1: \mathbb{R}^2 中的凸集与非凸集的例子.

由凸集的定义知图2.1中左边的集合是凸的, 但右边的集合是非凸的. 此外, 单点集 $\{\mathbf{x}\}$ 和 \emptyset 均是 \mathbb{R}^n 中的凸集. 下面是更多凸集的例子.

例2.1. 已知 $\mathbf{0} \neq \mathbf{s} \in \mathbb{R}^n, c \in \mathbb{R}$.

超平面(Hyperplane): $H(\mathbf{s}, c) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{s}^T \mathbf{x} = c\}$. 称 \mathbf{s} 为超平面的法向量. 超平面是凸集.

半空间(Halfspaces): $H^+(\mathbf{s}, c) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{s}^T \mathbf{x} \leq c\}, H^-(\mathbf{s}, c) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{s}^T \mathbf{x} \geq c\}$ 是凸集.

多面集(polyhedral set)有限个半空间的交集 $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$. 易见凸集的交还是凸集, 所以多面集也是凸集.

例2.2. \mathbb{R}^n 中的仿射集 M 是能表述成将线性子空间 $L \subset \mathbb{R}^n$ 平移一个向量 $\mathbf{a} \in \mathbb{R}^n$ 的集合:

$$M = \mathbf{a} + L = \{\mathbf{a} + \mathbf{y} \in \mathbb{R}^n : \mathbf{y} \in L\}. \quad (2.1)$$

其中的线性子空间是由仿射子空间 M 唯一定义的, 并且是 M 中向量之差形成的集合:

$$(2.1) \Rightarrow L = M - M = \{\mathbf{y} = \mathbf{x}' - \mathbf{x}'' : \mathbf{x}', \mathbf{x}'' \in M\}.$$

平移向量 \mathbf{a} 并不能由仿射子空间唯一确定; 在(2.1)中, 能取 M 中的任意向量作为 \mathbf{a} (并且只有 M 中的向量):

$$(2.1) \Rightarrow M = \mathbf{a}' + L \quad \forall \mathbf{a}' \in M.$$

仿射子空间的常用例子: 可解线性方程组的解集

$$M \text{ 是 } \mathbb{R}^n \text{ 中的仿射子空间} \Leftrightarrow \emptyset \neq M \equiv \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{b}\} \equiv \underbrace{\mathbf{a}}_{\mathbf{A}\mathbf{a}=\mathbf{b}} + \underbrace{\{\mathbf{x} : \mathbf{A}\mathbf{x} = 0\}}_{\text{Ker } \mathbf{A}}.$$

由该事实, 再结合例2.1中多面集是凸集的事实, 得到仿射子空间也是凸集.

例2.3. 范数单位球(Unit balls of norms): 设 $\|\cdot\|$ 是 \mathbb{R}^n 上的范数, 该范数的单位球 $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq 1\}$ 是凸的, 任何其它球 $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq r\}$ 也是凸的, 其中 $r \geq 0$ 是已知的.

证明. 设 $\|\mathbf{x} - \mathbf{a}\| \leq r, \|\mathbf{y} - \mathbf{a}\| \leq r, \theta \in [0, 1]$. 则

$$\begin{aligned} \|[(1-\theta)\mathbf{x} + \theta\mathbf{y}] - \mathbf{a}\| &= \|(1-\theta)(\mathbf{x} - \mathbf{a}) + \theta(\mathbf{y} - \mathbf{a})\| \\ &\leq \|(1-\theta)(\mathbf{x} - \mathbf{a})\| + \|\theta(\mathbf{y} - \mathbf{a})\| \\ &= (1-\theta)\|(\mathbf{x} - \mathbf{a})\| + \theta\|(\mathbf{y} - \mathbf{a})\| \\ &\leq (1-\theta)r + \theta r \\ &= r, \end{aligned}$$

其中第一个不等式和第二个不等式分别源于范数的三角不等式和齐次性. □

\mathbb{R}^n 上范数的标准例子: ℓ_p 范数

$$\|\mathbf{x}\|_p = \begin{cases} (\sum_{i=1}^n |x_i|^p)^{1/p}, & 1 \leq p < \infty \\ \max_i |x_i|, & p = \infty \end{cases}$$

请注意: $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$ 是标准的欧氏范数;

$$\|\mathbf{x}\|_1 = \sum_i |x_i|, \quad \|\mathbf{x}\|_\infty = \max_i |x_i| \text{ (uniform 范数)}.$$

例2.4. \mathbb{R}^n 中的椭球是由 $n \times n$ 对称正定矩阵 \mathbf{Q} , 中心 $\mathbf{a} \in \mathbb{R}^n$ 及半径 $r > 0$ 定义的集合

$$E(\mathbf{a}, \mathbf{Q}/r) := \{\mathbf{x} \in \mathbb{R}^n : (\mathbf{x} - \mathbf{a})^T \mathbf{Q} (\mathbf{x} - \mathbf{a}) \leq r^2\}.$$

椭球是凸集.

证明. 由于 Q 是对称正定的, 由谱分解定理知存在对称正定矩阵 $Q^{1/2}$ 使得 $Q = Q^{1/2}Q^{1/2}$. 置

$$\|x\|_Q = \|Q^{1/2}x\|_2,$$

因为 $\|\cdot\|_2$ 是范数且 $Q^{1/2}$ 是非奇异的, 所以 $\|\cdot\|_Q$ 是 \mathbb{R}^n 上的范数. 有

$$(x-a)^T Q(x-a) = [(x-a)^T Q^{1/2}][Q^{1/2}(x-a)] = \|Q^{1/2}(x-a)\|_2^2 = \|x-a\|_Q^2.$$

因此, $E(a, Q/r)$ 是一个 $\|\cdot\|_Q$ 球, 因此是凸集. \square

例2.5 (凸集的 ϵ -邻域). 设 M 是 \mathbb{R}^n 中的非空凸集, 且 $\epsilon \geq 0$. 那么集合

$$X = \{x : \text{dist}_{\|\cdot\|}(x, M) := \inf_{y \in M} \|x - y\| \leq \epsilon\}$$

是凸的.

证明. $x \in X$ 当且仅当对每个 $\epsilon' > \epsilon$ 存在 $y \in M$ 使得 $\|x - y\| \leq \epsilon'$. 现在有

$$x, y \in X, \theta \in [0, 1]$$

$$\Rightarrow \forall \epsilon' > \epsilon \exists u, v \in M : \|x - u\| \leq \epsilon', \|y - v\| \leq \epsilon'$$

$$\Rightarrow \forall \epsilon' > \epsilon \exists u, v \in M : \underbrace{\theta\|x - u\| + (1-\theta)\|y - v\|}_{\geq \|\theta x + (1-\theta)y - [\theta u + (1-\theta)v]\|} \leq \epsilon', \forall \theta \in [0, 1]$$

$$\Rightarrow \forall \epsilon' > \epsilon, \forall \theta \in [0, 1], \exists w = \theta u + (1-\theta)v \in M : \|\theta x + (1-\theta)y - w\| \leq \epsilon'$$

$$\Rightarrow \theta x + (1-\theta)y \in X, \forall \theta \in [0, 1].$$

\square

2.2 凸组合与凸包

设 k 是正整数. 点 $x_1, \dots, x_k \in \mathbb{R}^n$ 的**凸组合**(convex combination)是系数非负且系数之和为1的线性组合:

$$\sum_{i=1}^k \theta_i x_i, \theta_1, \dots, \theta_k \geq 0, \sum_{i=1}^k \theta_i = 1. \quad (2.2)$$

下面的命题给出了凸集的一种刻画.

命题2.1 (凸集的内表示). 集合 $X \subset \mathbb{R}^n$ 是凸的当且仅当它关于取凸组合运算是封闭的, 即 X 中点的凸组合属于 X .

证明. 假设 X 是凸的. 下面对凸组合中属于 X 的点数 $k \geq 1$ 用归纳法证明 X 关于取凸组合运算是封闭的. $k = 1$ 的情况是平凡的. 现在假设对 k 成立. 设

$$\mathbf{x}_1, \dots, \mathbf{x}_{k+1} \in X, \theta_i \geq 0, \sum_{i=1}^{k+1} \theta_i = 1.$$

不失一般性, 假设 $0 \leq \theta_{k+1} < 1$. 那么

$$\sum_{i=1}^{k+1} \theta_i \mathbf{x}_i = (1 - \theta_{k+1}) \underbrace{\left(\sum_{i=1}^k \frac{\theta_i}{1 - \theta_{k+1}} \mathbf{x}_i \right)}_{\in X} + \theta_{k+1} \mathbf{x}_{k+1} \in X.$$

反之, 由 X 关于凸组合运算封闭, 取 $m = 2$ 即得 X 是凸集. \square

由定义易于得到如下结论.

命题2.2. \mathbb{R}^n 中任意凸子集族 $\{X_\alpha\}_{\alpha \in \mathcal{A}}$ 的交集 $X = \bigcap_{\alpha \in \mathcal{A}} X_\alpha$ 是凸的.

设 $X \subset \mathbb{R}^n$ 是任意集合. 那么在包含 X 的凸集(其存在, 比如 \mathbb{R}^n)中存在一个最小的, 即所有包含 X 的凸集之交. 由此可得如下定义.

定义2.2. 称包含 X 的最小凸集是 X 的**凸包**(convex hull), 记作 $\text{conv} X$.

可将以上定义看作从**外部**刻画一个集合 X 的凸包. 然而, 给定 X 的凸包中的点, 由上述定义并不清楚这些给定的点与 X 中点的关系. 这激发了下一个命题, 其在某种意义上从**内部**刻画 X 的凸包.

命题2.3 (由凸组合得到凸包). 对 \mathbb{R}^n 的每个子集 X , 它的凸包 $\text{conv} X$ 恰好是由 X 中点的所有凸组合组成的集合 \hat{X} .

证明. 每个包含 X 的凸集也包含由 X 中的点得到的凸组合, 所以有 $\hat{X} \subset \text{conv} X$.

下面证明反包含 $\text{conv} X \subset \hat{X}$. 因为 $\text{conv} X$ 是包含 X 的最小凸集, 由此证明 \hat{X} 包含 X (显然)并且是凸集即可. 为了证明 \hat{X} 是凸集, 设 $\mathbf{x}, \mathbf{y} \in \hat{X}, \theta \in [0, 1]$. 由 \hat{X} 的定义, 存在 $\mathbf{y}_1, \dots, \mathbf{y}_p, \mathbf{y}_{p+1}, \dots, \mathbf{y}_q \in X$ 和

$$\theta_1, \dots, \theta_p, \theta_{p+1}, \dots, \theta_q \geq 0, \sum_{i=1}^p \theta_i = 1, \sum_{i=p+1}^q \theta_i = 1.$$

使得

$$\mathbf{z} = \sum_{i=1}^p \theta_i \mathbf{y}_i, \quad \mathbf{y} = \sum_{i=p+1}^q \theta_i \mathbf{y}_i.$$

那么

$$\mathbf{z}(\theta) = \theta \mathbf{x} + (1 - \theta) \mathbf{y} = \sum_{i=1}^p \theta \theta_i \mathbf{y}_i + \sum_{i=p+1}^q (1 - \theta) \theta_i \mathbf{y}_i$$

及

$$\sum_{i=1}^p \theta \theta_i + \sum_{i=p+1}^q (1-\theta) \theta_i = 1.$$

换句话说, $z(\theta)$ 是 X 中点的凸组合. 这样, 有 $z(\theta) \in \hat{X}$. 这表明 \hat{X} 是凸集. 证毕. \square

推论 设集合 $X \subseteq \mathbb{R}^n$ 是任意的. 则 X 是凸的当且仅当 $\text{conv} X = X$.

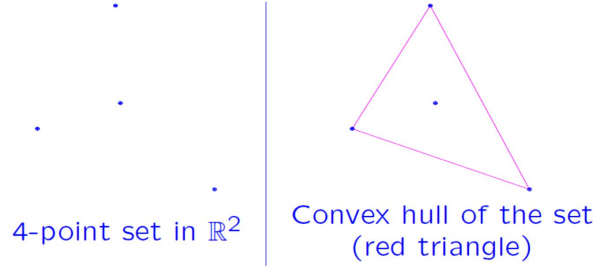


图 2.2: \mathbb{R}^2 中的集合及其凸包的例子.

定义 2.3. 已知 \mathbb{R}^n 中的 $m+1$ 个点 $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_m$. 如果向量 $\mathbf{x}_1 - \mathbf{x}_0, \mathbf{x}_2 - \mathbf{x}_0, \dots, \mathbf{x}_m - \mathbf{x}_0$ 是线性无关的, 则称向量 $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_m$ 仿射无关(affine independent).

动机: 设 $X \subset \mathbb{R}^n$ 非空.

包含 X 的仿射子空间的交集也是一个仿射子空间. 它显然是包含 X 的最小仿射子空间; 它被称作 X 的仿射包(affine hull), 记作 $\text{aff} X$. 易于看到 $\text{aff} X$ 就是 X 中点的所有仿射组合, 即组合系数之和为 1 的线性组合:

$$\text{aff} X = \left\{ \mathbf{x} = \sum_i \theta_i \mathbf{x}_i : \mathbf{x}_i \in X, \sum_i \theta_i = 1 \right\}.$$

$m+1$ 个点 $\mathbf{x}_0, \dots, \mathbf{x}_m$ 是仿射无关的当且仅当每个 $\mathbf{x} \in \text{aff} X$ 能被唯一地表示成 $\mathbf{x}_0, \dots, \mathbf{x}_m$ 的仿射组合:

$$\sum_i \theta_i \mathbf{x}_i = \sum_i \mu_i \mathbf{x}_i \text{ \& } \sum_i \theta_i = \sum_i \mu_i = 1 \Rightarrow \theta_i \equiv \mu_i \quad \forall i.$$

此时, 点 $\mathbf{x} \in M = \text{aff}(\{\mathbf{x}_0, \dots, \mathbf{x}_m\})$ 作为 $\mathbf{x}_0, \dots, \mathbf{x}_m$ 的仿射组合, 具有表示

$$\mathbf{x} = \sum_{i=0}^m \theta_i \mathbf{x}_i, \quad \sum_i \theta_i = 1.$$

称该表示中的系数 θ_i 是 $\mathbf{x} \in M$ 关于 M 的仿射基 $\mathbf{x}_0, \dots, \mathbf{x}_m$ 的重心坐标(barycentric coordinates).

定义 2.4. 顶点是 $\mathbf{x}_0, \dots, \mathbf{x}_m$ 的 m -维单纯形(simplex) Δ 是 $m+1$ 个仿射无关点 $\mathbf{x}_0, \dots, \mathbf{x}_m$ 的凸包:

$$\Delta = \Delta(\mathbf{x}_0, \dots, \mathbf{x}_m) = \text{conv}\{\mathbf{x}_0, \dots, \mathbf{x}_m\} = \left\{ \sum_{i=0}^m \theta_i \mathbf{x}_i : \sum_{i=0}^m \theta_i = 1, \theta_i \geq 0, i = 0, \dots, m \right\}.$$

顶点为 $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_m$ 的 m -维单纯形是顶点集的凸包, 因此是凸集. 单纯形中每个点是顶点的凸组合, 并且系数由该点唯一确定.

例2.6 (单纯形). (A) 2-维单纯形是由3个不共线的点确定的, 是以这三个点为顶点的三角形;

(B) 设 e_1, \dots, e_n 是 \mathbb{R}^n 中的标准正交基. 这 n 个点是仿射无关的, 由它们确定的 $(n-1)$ -维单纯形是**标准单纯形**(standard simplex)

$$\Delta_n = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \mathbf{0}, \sum_{i=1}^n x_i = 1\}.$$

(C) 将 e_1, \dots, e_n 加入 $e_0 = \mathbf{0}$, 得到 $n+1$ 个仿射无关的点. 对应的 n 维单纯形是

$$\Delta_n^+ = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \mathbf{0}, \sum_{i=1}^n x_i \leq 1\}.$$

定义2.5. 称非空集合 $K \subseteq \mathbb{R}^n$ 是**锥**(Cone), 若对任意的 $\mathbf{x} \in K$ 有

$$\{t\mathbf{x} : t > 0\} \subseteq K.$$

若 K 是凸的, 则称 K 是**凸锥**(convex cone).

例2.7 (凸锥). (A) 非负卦限(Non-Negative Orthant): $\mathbb{R}_+^n = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \mathbf{0}\}$.

(B) Lorentz锥(Lorentz Cone): $L^n = \{\mathbf{x} \in \mathbb{R}^n : x_n \geq \sqrt{x_1^2 + \dots + x_{n-1}^2}\}$.

(C) 半定锥(Semidefinite Cone): $\mathcal{S}_+^n = \{\mathbf{X} \in \mathcal{S}^n : \mathbf{x}^T \mathbf{Q} \mathbf{x} \geq 0, \forall \mathbf{x} \in \mathbb{R}^n\}$.

(D) 任意(有限或无限)齐次非严格线性不等式组的解集

$$P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_\alpha^T \mathbf{x} \geq 0, \forall \alpha \in \mathcal{A}\}$$

是闭锥, 其中 \mathcal{A} 是有限或者无限指标集. 特别地, 已知 $\mathbf{A} \in \mathbb{R}^{m \times n}$, 集合 $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A} \mathbf{x} \leq \mathbf{0}\}$ 是**多面锥**(Polyhedral Cone).

请注意: \mathbb{R}^n 中的每个闭锥是可数个非严格齐次线性不等式组的解集.

命题2.4. \mathbb{R}^n 的非空子集 K 是凸锥当且仅当

(i) K 是锥, 并且

(ii) K 关于加法封闭, 即 $\mathbf{x}, \mathbf{y} \in K \Rightarrow \mathbf{x} + \mathbf{y} \in K$.

证明 \Rightarrow : 设 K 是凸的, 且 $\mathbf{x}, \mathbf{y} \in K$. 则由凸性 $\frac{1}{2}(\mathbf{x} + \mathbf{y}) \in K$. 再由 K 是锥有 $\mathbf{x} + \mathbf{y} \in K$. 这样, 凸锥关于加法封闭.

\Leftarrow : 设 K 是锥且关于加法封闭. 这时, K 中的向量 \mathbf{x}, \mathbf{y} 的凸组合 $(1-\theta)\mathbf{x} + \theta\mathbf{y}$ 分别是属于 K 的向量 $(1-\theta)\mathbf{x}$ 和 $\theta\mathbf{y}$ 之和, 由于 K 关于加法封闭, 所以这个凸组合属于 K . 这样, 关于加法封闭的锥是凸的.

锥是极为重要的一类凸集, 它的性质与一般凸集的性质是“平行的”. 比如

- (i) 任意一族凸锥的交也是凸锥. 因此, 对于每个非空集合 X , 在包含 X 的锥中存在最小的凸锥, 称其为 X 的**锥包**(cone hull), 记作 $\text{cone } X$.
- (ii) 非空集合 K 是凸锥当且仅当 K 关于取**锥组合**(conic combination)运算是封闭的, 即非负系数的线性组合. 从而 K 是锥当且仅当 $K = \text{cone}(K)$.
- (iii) 非空集合 X 的锥包 $\text{cone } X$ 恰好是 X 中元素的所有锥组合组成的.

2.3 保持凸性的运算

尽管从理论上讲总是能够直接由定义建立一个集合的凸性, 但该方法通常不易于实施. 许多情况下, 需要对一族凸集执行某种运算, 然后希望知道运算后得到的集合是否是凸的. 这样, 将遇到如下问题: 哪些集合运算是**保持凸性**的?

命题2.5. 如下命题成立:

- (a) 已知 $\alpha \in \mathcal{A} \subseteq \mathbb{R}$. 如果 $\forall \alpha \in \mathcal{A}$, $X_\alpha \subseteq \mathbb{R}^n$ 是凸集, 那么 $\bigcap_{\alpha \in \mathcal{A}} X_\alpha$ 是凸的.
- (b) 如果 $X_\ell \subseteq \mathbb{R}^{n_\ell}, 1 \leq \ell \leq L$ 是凸集, 那么集合

$$X = X_1 \times \cdots \times X_L := \{(\mathbf{x}^1, \dots, \mathbf{x}^L) : \mathbf{x}^\ell \in X_\ell, 1 \leq \ell \leq L\} \subseteq \mathbb{R}^{n_1 + \cdots + n_L}$$

是凸的.

- (c) 如果 X_1, \dots, X_L 是 \mathbb{R}^n 中的非空凸集, 且 β_1, \dots, β_L 是实数, 那么集合

$$X = \beta_1 X_1 + \cdots + \beta_L X_L := \{\beta_1 \mathbf{x}_1 + \cdots + \beta_L \mathbf{x}_L : \mathbf{x}_\ell \in X_\ell, 1 \leq \ell \leq L\}$$

是凸的.

例2.8. 一个点 $\mathbf{x} \in \mathbb{R}^n$ 是

- (i) “良好的”, 如果 \mathbf{x} 满足已知线性约束组 $\mathbf{A}\mathbf{x} \leq \mathbf{b}$,
- (ii) “优秀的”, 如果它控制(dominate)一个良好点: $\exists \mathbf{y} : \mathbf{y}$ 是 “良好的” 且 $\mathbf{x} \geq \mathbf{y}$.
- (iii) “半优秀的”, 如果它以逐坐标在精度为0.1的方式下可由一个优秀点来近似:

$$\forall (i, \epsilon' > 0.1) \exists \mathbf{y} \text{ s.t. } |y_i - x_i| \leq \epsilon' \text{ 且 } \mathbf{y} \text{ 是优秀的.}$$

问半优秀点形成的集合 Y 是凸的吗?

答案: 是的. 的确是凸的.

良好点集 X_g 是多面集, 因而是凸的.

优秀点集 X_{exc} 是良好点集 X_g 与非负卦限 \mathbb{R}_+^n 之和, 因此是凸的.

对每个指标 i , 良好点集的第 i 个坐标集合 X_{exc}^i 是 X_{exc} 在第 i 个坐标轴的投影; 因为该投影是仿射变换, 所以 X_{exc}^i 是凸的.

对每个指标 i , 作为坐标轴上凸集 X_{exc}^i 的0.1-邻域的集合 Y^i 是凸的.

半优秀点集 Y 是凸集族 Y^1, \dots, Y^n 的直积, 所以 Y 也是凸的.

易于看到求集合的并集不能保持凸性. 下面再介绍一些保持凸性的运算.

命题2.6. 设 $\mathbf{x} \mapsto \mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ 是从 $\mathbb{R}^n \rightarrow \mathbb{R}^m$ 的仿射映射(affine mapping). 如果 $X \subseteq \mathbb{R}^n$ 是凸集, 那么 X 在映射下的像

$$\mathcal{A}(X) = \{\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} : \mathbf{x} \in X\}$$

是凸的. 反之, 如果 $Y \subseteq \mathbb{R}^m$ 是凸集, 则 Y 在映射下的逆像——集合

$$\mathcal{A}^{-1}(Y) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} + \mathbf{b} \in Y\}$$

是凸的.

定义透视函数(perspective function) $P : \mathbb{R}^n \times \mathbb{R}_{++} \rightarrow \mathbb{R}^n$ 为

$$P(\mathbf{x}, t) = \mathbf{x}/t. \quad (2.3)$$

下面的命题表明透视函数能保持凸性.

命题2.7. 设 $P : \mathbb{R}^n \times \mathbb{R}_{++} \rightarrow \mathbb{R}^n$ 是透视函数, $X \subseteq \mathbb{R}^n \times \mathbb{R}_{++}$ 是凸集. 则像

$$P(X) = \{\mathbf{x}/t \in \mathbb{R}^n : (\mathbf{x}, t) \in X\}$$

是凸的. 反之, 如果 $T \subseteq \mathbb{R}^n$ 是凸集, 则原像

$$P^{-1}(T) = \{(\mathbf{x}, t) \in \mathbb{R}^n \times \mathbb{R}_{++} : \mathbf{x}/t \in T\}$$

是凸的.

证明 对任一 $z_i = (\mathbf{x}_i, t_i) \in \mathbb{R}^n \times \mathbb{R}_{++}, i = 1, 2$ 和 $\theta \in [0, 1]$, 有

$$P(\theta z_1 + (1 - \theta)z_2) = \frac{\theta \mathbf{x}_1 + (1 - \theta)\mathbf{x}_2}{\theta t_1 + (1 - \theta)t_2} = \beta P(z_1) + (1 - \beta)P(z_2),$$

其中

$$\beta = \frac{\theta t_1}{\theta t_1 + (1 - \theta)t_2} \in [0, 1].$$

此外, 随着 θ 从0增加到1, β 也从0增加到1. 从而

$$P([z_1, z_2]) = [P(z_1), P(z_2)] \subseteq \mathbb{R}^n.$$

证毕. □

下面的结论是命题2.6和命题2.7的直接应用，它表明线性分式映射(射影映射(projective mappings))也可以保持凸性.

推论 设 $A \in \mathbb{R}^n \rightarrow \mathbb{R}^{m+1}$ 是仿射映射

$$A(\mathbf{x}) = \begin{bmatrix} \mathbf{Q} \\ \mathbf{c}^T \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{u} \\ d \end{bmatrix},$$

其中 $\mathbf{Q} \in \mathbb{R}^{m \times n}$, $\mathbf{c} \in \mathbb{R}^n$, $\mathbf{u} \in \mathbb{R}^m$, $d \in \mathbb{R}$. 进一步, 设 $D = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{c}^T \mathbf{x} + d > 0\}$. 定义线性-分式映射(linear-fractional map) $f : D \rightarrow \mathbb{R}^m$ 为 $f = P \circ A$, 其中 P 是透视函数(2.3). 如果 $X \subseteq D$ 是凸的, 则像 $f(X)$ 是凸的. 反之, 如果 $T \subseteq \mathbb{R}^m$ 是凸的, 则逆向 $f^{-1}(T)$ 是凸的.

2.4 凸集的拓扑性质

称集合 $X \subseteq \mathbb{R}^n$ 是**闭的**(closed), 如果 X 中收敛点列的极限仍包含在 X 中:

$$\mathbf{x}_i \in X \ \& \ \mathbf{x}_i \rightarrow \mathbf{x}, i \rightarrow \infty \Rightarrow \mathbf{x} \in X.$$

称 X 是**开的**(open), 如果对于它中的每个点 \mathbf{x} , 存在 $r > 0$ 使得以 \mathbf{x} 为中心 r 为半径的球包含在 X 中:

$$\mathbf{x} \in X \Rightarrow \exists r > 0 : \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\|_2 \leq r\} \subseteq X.$$

比如任意个非严格线性不等式组的解集

$$\{\mathbf{x} : \mathbf{a}_\alpha^T \mathbf{x} \leq \mathbf{b}_\alpha, \alpha \in \mathcal{A}\}$$

是闭集; **有限个**严格线性不等式组的解集

$$\{\mathbf{x} : \mathbf{A}\mathbf{x} < \mathbf{b}\}$$

是开集, 其中 $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$ 是已知的.

关于开集和闭集, 有以下事实:

- A. X 是闭的当且仅当它的补集 $\mathbb{R}^n \setminus X$ 是开的.
- B. 任意闭集族的交是闭集, 且有限个闭集的并是闭集.
- B'. 任意开集族的并是开集, 且有限个开集的交是开集.

由B知所有包含已知集合 X 的闭集之交仍是闭集; 称这个交集为 X 的**闭包**, 记为 $\text{cl}X$, 它是包含 X 的最小闭集. $\text{cl}X$ 恰好是 X 中收敛序列的极限点所组成的集合:

$$\text{cl}X = \{\mathbf{x} : \exists x_i \in X \text{ s.t. } \mathbf{x} = \lim_{i \rightarrow \infty} \mathbf{x}_i\}.$$

由 B' 知所有包含在已知集合 X 中的开集之并仍是开集；称这个并集为 X 的内部，记为 $\text{int}X$ ，是包含在 X 中的最大开集. 它恰好是 X 的所有内点，即以 X 中的点为中心且半径为正的球包含在 X 中，组成的集合：

$$\text{int}X = \{\mathbf{x} : \exists r > 0 \text{ s.t. } \{\mathbf{y} : |\mathbf{y} - \mathbf{x}|_2 \leq r\} \subseteq X\}.$$

设 $X \subseteq \mathbb{R}^n$. 则

$$\text{int}X \subseteq X \subset \text{cl}X. \quad (2.4)$$

称“差”

$$\partial X = \text{cl}X \setminus \text{int}X$$

是 X 的边界(boundary)；因为闭集 $\text{cl}X$ 和开集 $\text{int}X$ 的补集之交是闭集，所以边界总是闭的.

一般地， $\text{int}X$ 与 $\text{cl}X$ 之间的差异有可能很大. 比如设 $X \subset \mathbb{R}$ 是区间 $[0, 1]$ 上的无理数. 那么 $\text{int}X = \emptyset$, $\text{cl}X = [0, 1]$. 因此 $\text{int}X$ 与 $\text{cl}X$ 很有显著的差别. 幸运地是，凸集可由它的闭包(和内部，如果其非空)完美地近似.

命题2.8. 设 $X \subseteq \mathbb{R}^n$ 是非空凸集. 那么

- (a) $\text{int}X$ 和 $\text{cl}X$ 都是凸的,
- (b) 如果 $\text{int}X$ 非空, 那么 $\text{int}X$ 是 $\text{cl}X$ 的稠密子集. 此外,

$$\mathbf{x} \in \text{int}X, \mathbf{y} \in \text{cl}X \Rightarrow (1 - \theta)\mathbf{x} + \theta\mathbf{y} \in \text{int}X \quad \forall \theta \in [0, 1]. \quad (2.5)$$

证明. (a) 设 X 是凸的, 那么 $\text{int}X$ 和 $\text{cl}X$ 都是凸的几乎是显然的. 的确, 为了证明 $\text{int}X$ 是凸的, 请注意对每两个点 $\mathbf{x}, \mathbf{y} \in \text{int}X$, 存在一个共同的 $r > 0$ 使得分别以 \mathbf{x} 和 \mathbf{y} 为中心, 半径均为 r 的球 $B_{\mathbf{x}}$ 和 $B_{\mathbf{y}}$ 包含在 X 中. 因为 X 是凸的, 对每个 $\theta \in [0, 1]$, X 包含集合 $(1 - \theta)B_{\mathbf{x}} + \theta B_{\mathbf{y}}$, 这显然是中心为 $(1 - \theta)\mathbf{x} + \theta\mathbf{y}$ 半径为 r 的球. 这样, 对所有 $\theta \in [0, 1]$, $(1 - \theta)\mathbf{x} + \theta\mathbf{y} \in \text{int}X$. 所以 $\text{int}X$ 是凸集.

类似地, 为了证明 $\text{cl}X$ 是凸的, 假设 $\mathbf{x}, \mathbf{y} \in \text{cl}X$, 且对选取的 $\mathbf{x}_i, \mathbf{y}_i \in X$ 使得 $\mathbf{x} = \lim_{i \rightarrow \infty} \mathbf{x}_i, \mathbf{y} = \lim_{i \rightarrow \infty} \mathbf{y}_i$. 则对 $\theta \in [0, 1]$, 有 $(1 - \theta)\mathbf{x}_i + \theta\mathbf{y}_i \in X$ 且

$$(1 - \theta)\mathbf{x} + \theta\mathbf{y} = \lim_{i \rightarrow \infty} (1 - \theta)\mathbf{x}_i + \theta\mathbf{y}_i,$$

因此对所有 $\theta \in [0, 1]$ 有 $(1 - \theta)\mathbf{x} + \theta\mathbf{y} \in X$. 所以 $\text{cl}X$ 是凸集.

(b) 设 X 是凸集, 且 $\text{int}X$ 非空. 易见(2.5)成立是 $\text{int}X$ 在 $\text{cl}X$ 中稠密的充分条件. 如果(2.5)成立, 设 $\bar{\mathbf{x}} \in \text{int}X$ (因为后者非空). 每个点 $\mathbf{x} \in \text{cl}X$ 是序列

$$\mathbf{x}_i = \frac{1}{i}\bar{\mathbf{x}} + (1 - \frac{1}{i})\mathbf{x}$$

的极限. 由(2.5), 所有的 \mathbf{x}_i 属于 $\text{int}X$, 这样 $\text{int}X$ 在 $\text{cl}X$ 中是稠密的.

下面证明(2.5)成立. 设 $\mathbf{x} \in \text{int}X$, $\mathbf{y} \in \text{cl}X$, $\theta \in [0, 1)$. 下面证明 $(1 - \theta)\mathbf{x} + \theta\mathbf{y} \in \text{int}X$. 因为 $\mathbf{x} \in \text{int}X$, 存在 $r > 0$ 使得以 \mathbf{x} 为中心 r 为半径的球 B 包含于 X . 因为 $\mathbf{y} \in \text{cl}X$, 从而存在序列 $\mathbf{x}_i \in X$ 使得 $\mathbf{y} = \lim_{i \rightarrow \infty} \mathbf{y}_i$. 现在设

$$\begin{aligned} B^i &= (1 - \theta)B + \theta\mathbf{y}_i \\ &= \{z = \underbrace{[(1 - \theta)\mathbf{x} + \theta\mathbf{y}_i]}_{z_i} + (1 - \theta)\mathbf{h} : \|\mathbf{h}\|_2 \leq r\} \\ &= \{z = z_i + \boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq r'\}, \end{aligned}$$

其中 $r' = (1 - \theta)r$. 因为 $B \subseteq X$, $\mathbf{y}_i \in X$, 且 X 是凸的, 集合 B^i (是中心为 z_i 半径为 $r' > 0$ 的球)包含在 X 中. 因为当 $i \rightarrow \infty$ 时,

$$z_i \rightarrow z = (1 - \theta)\mathbf{x} + \theta\mathbf{y},$$

所以从某个指标开始, 所有这些球都包含中心为 z 半径为 $r'/2$ 的球 B' . 这样, $B' \subseteq X$, 即 $z \in \text{int}X$. \square

设 X 是凸集. 有可能恰好 $\text{int}X = \emptyset$ (比如 X 是 \mathbb{R}^3 中的线段). 在这种情况下, 内部肯定不能逼近 X 和 $\text{cl}X$. 这时怎么做?

克服该困难的一种自然的方式是定义**相对内部**(relative interior), 这里仅关于 X 的仿射包, 而不是关于 \mathbb{R}^n 取内部. 从几何上讲, 这个仿射包仅仅是某个 \mathbb{R}^m , $m \leq n$; 如有必要, 用这个 \mathbb{R}^m 代替 \mathbb{R}^n , 就能抵达 $\text{int}X$ 非空这种情况. 下面给出上述思想的实现.

定义2.6 (相对内部和相对边界). 设 X 是 \mathbb{R}^n 的非空子集, 且 M 是 X 的仿射包. X 的**相对内部**(relative interior)是所有 $\mathbf{x} \in X$ 中那些在 M 中存在以 \mathbf{x} 为中心半径为正的球包含于 X 的点组成的集合:

$$\text{rint}X = \{\mathbf{x} \in X : \exists r > 0 \text{ s.t. } \{\mathbf{y} \in \text{aff}X : \|\mathbf{y} - \mathbf{x}\|_2 \leq r\} \subseteq X\}.$$

由定义, X 的相对边界是 $\text{cl}X \setminus \text{rint}X$.

仿射子空间 M 是线性方程组的解集, 从而是闭的; 因此, 每个子集 $Y \subseteq M$ 的闭包都包含于 M ; Y 的闭包并不是别的什么, 仅仅是当把原始的全集 \mathbb{R}^n 用仿射子空间 M (从几何上看, M 仅仅是某个 \mathbb{R}^m , $m \leq n$) 替换后得到的 Y 的闭包. 本质见如下命题.

命题2.9. 设 $X \subseteq \mathbb{R}^n$ 是非空凸集. 那么 $\text{rint}X \neq \emptyset$.

证明. 若 X 是单点集, 易见 $X = \text{rint}X \neq \emptyset$. 否则, 由线性代数, 能找到 X 的仿射包 $\text{aff}X$ 的仿射基: $\exists \mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_m, m \geq 1$ 使得每个 $\mathbf{x} \in \text{aff}X$ 拥有表示

$$\mathbf{x} = \sum_{i=0}^m \theta_i \mathbf{x}_i, \sum_{i=0}^m \theta_i = 1$$

且这个表示中的系数由 \mathbf{x} 唯一确定.

从而, 只要 $\mathbf{x} \in \text{aff } X$, 关于变量 $\theta_0, \theta_1, \dots, \theta_m$ 的线性方程组

$$\begin{aligned}\sum_{i=0}^m \theta_i \mathbf{x}_i &= \mathbf{x} \\ \sum_{i=0}^m \theta_i &= 1\end{aligned}$$

有唯一解. 由于这个解是唯一的, 再次由线性代数, 该解对 $\mathbf{x} \in \text{aff } X$ 是连续依赖的. 特别地, 当

$$\mathbf{x} = \bar{\mathbf{x}} = \frac{1}{m+1} \sum_{i=0}^m \mathbf{x}_i$$

时, 解 $\theta_i = \frac{1}{m+1}$ 是正的; 由连续性, 当 $\mathbf{x} \in \text{aff } X$ 充分接近 $\bar{\mathbf{x}}$ 时, 这个解仍能保持是正的:

$$\exists r > 0 \text{ s.t. } \mathbf{x} \in \text{aff } X, \|\mathbf{x} - \bar{\mathbf{x}}\|_2 \leq r \Rightarrow \mathbf{x} = \sum_{i=0}^m \theta_i(\mathbf{x}) \mathbf{x}_i, \sum_{i=0}^m \theta_i(\mathbf{x}) = 1, \theta_i(\mathbf{x}) > 0.$$

从而当 X 是凸集时, $\bar{\mathbf{x}} \in \text{rint } X$. □

这样, 如有必要, 用一个类似的全集替换原始“全集” \mathbb{R}^n , 把研究任意非空凸集 X 转化成该集合内部非空这种情况(这仅仅是 X 的相对内部). 特别地, 关于“满维”情况的结论蕴含着: 对非空凸集 X , $\text{rint } X$ 和 $\text{cl } X$ 都是凸集, 满足

$$\emptyset \neq \text{rint } X \subseteq X \subseteq \text{cl } X \subseteq \text{aff } X,$$

并且 $\text{rint } X$ 在 $\text{cl } X$ 中是稠密的. 从而, 只要 $\mathbf{x} \in \text{rint } X, \mathbf{y} \in \text{cl } X$ 且 $\theta \in [0, 1)$, 有

$$(1 - \theta)\mathbf{x} + \theta\mathbf{y} \in \text{rint } X.$$

3 关于凸集的主要定理

3.1 Carathéodory定理

定义3.1. 设 M 是 \mathbb{R}^n 的仿射子空间, 因此存在线性子空间 L 和 $\mathbf{a} \in M$ 使得 $M = \mathbf{a} + L$. 称 L 的线性维数是 M 的仿射维数, 记为 $\dim M$.

例3.1. 单点集的仿射维数是0, \mathbb{R}^n 的仿射维数是 n . 仿射子空间 $M = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}\}$ 的仿射维数是 $n - \text{rank}(\mathbf{A})$.

这样, 结合仿射包的概念, 可以为 \mathbb{R}^n 中任意一个集合定义维数的概念. 特别地, 有如下定义.

定义3.2. 已知非空集合 $X \subseteq \mathbb{R}^n$. 称仿射包 $\text{aff } X$ 的仿射维数是 X 的仿射维数(affine dimension), 记作 $\dim X$.

易见 $0 \leq \dim X \leq n$. 粗略地讲, $\dim X$ 是 X 的固有维数. 像将要看到的, 这个量在优化中扮演着基石的角色. 为了更好地理解集合的仿射维数这一概念, 考虑下面的例子.

例3.2 (集合的仿射维数). 考虑两点集合 $X = \{(1, 1), (3, 2)\} \subseteq \mathbb{R}^2$. 由命题2.3 (i), 有

$$\text{aff } X = \{(1 - \theta)(1, 1) + \theta(3, 2) : \theta \in \mathbb{R}\}.$$

易于看到

$$\text{aff } X = \{(0, 1/2)\} + L,$$

其中 $L = \{\beta(1, 1/2) : \beta \in \mathbb{R}\}$ 是由向量 $(1, 1/2)$ 生成的子空间. 因此, 得

$$\dim X = \dim L = 1.$$

定理3.1 (Carathéodory). 设 $\emptyset \neq X \subset \mathbb{R}^n$. 则每个点 $\mathbf{x} \in \text{conv } X$ 是 X 中最多 $\dim X + 1$ 个点的凸组合.

证明. 将证明如果 \mathbf{x} 是 X 中有限个点 $\mathbf{x}_1, \dots, \mathbf{x}_k$ 的凸组合, 则 \mathbf{x} 至多是这些点中 $m + 1$ 个点的凸组合, 其中 $m = \dim X$. 如有必要的话, 用 $\text{aff } X$ 替换 \mathbb{R}^n , 所以证明 $m = n$ 的情况是充分的.

考虑将 \mathbf{x} 写成 $\mathbf{x}_1, \dots, \mathbf{x}_k$ 的非零系数最少的凸组合, 然后证明这个数小于等于 $n + 1$ 即可. 考虑凸组合

$$\mathbf{x} = \sum_{i=1}^k \theta_i \mathbf{x}_i, \text{ 其中 } \mathbf{x}_1, \dots, \mathbf{x}_k \in X, \theta_1, \dots, \theta_k > 0, \sum_{i=1}^k \theta_i = 1.$$

若 $k \leq n + 1$, 结论成立. 现设 $k > n + 1$. 欲证明: 在不改变 \mathbf{x} 的情况下, 可将系数 θ_i 中的一个置为0. 首先, 由于 $k \geq n + 2$, 因此向量 $\mathbf{x}_2 - \mathbf{x}_1, \mathbf{x}_3 - \mathbf{x}_1, \dots, \mathbf{x}_k - \mathbf{x}_1$ 在 \mathbb{R}^n

中一定线性相关. 确切地, 存在不全为零的 $\beta_1, \dots, \beta_k \in \mathbb{R}$, 使得 $\sum_{i=1}^k \beta_i x_i = 0$ 且 $\sum_{i=1}^k \beta_i = 0$. 不妨设 $\beta_1 > 0$. 对于 $i = 1, \dots, k$, 定义

$$\theta_i(t^*) = \theta_i - t^* \beta_i,$$

其中

$$t^* = \min_{j: \beta_j > 0} \frac{\theta_j}{\beta_j} = \max\{t \geq 0 : \theta_i - t\beta_i \geq 0, i = 1, \dots, k\}.$$

由于至少有 $\beta_1 > 0$, 因此 $t^* < \infty$. 接下来可以直接验证

$$\mathbf{x} = \sum_{i=1}^k \theta_i(t^*) \mathbf{x}_i, \sum_{i=1}^k \theta_i(t^*) = 1, \theta_1(t^*), \dots, \theta_k(t^*) \geq 0$$

以及 $|\{i : \theta_i(t^*) > 0\}| \leq k - 1$. 重复该处理过程, 直到最多只有 $n + 1$ 个正系数为止. \square

定理3.2 (Caratheodory, 锥版本). 设 $\emptyset \neq X \subseteq \mathbb{R}^n$. 则每个向量 $\mathbf{x} \in \text{cone} X$ 是 X 中至多 n 个向量的锥组合.

注记: 由Caratheodory定理给出的界(普通版本和锥版本)是紧的, 比如(i) 对于有 $m + 1$ 个顶点 v_0, \dots, v_m 的单纯形 Δ , 有 $\dim \Delta = m$, 并且需要用所有的顶点将质心 $\frac{1}{m+1} \sum_{i=0}^m v_i$ 表示为顶点的凸组合. (ii) \mathbb{R}^n 中 n 个标准正交基的锥包恰好是非负卦限 \mathbb{R}_+^n , 并且需要取全部 n 个向量的锥组合, 才可得到 \mathbb{R}_+^n 中的所有 n -维向量.

尽管Carathéodory定理指出: 任意 $\mathbf{x} \in \text{conv} X$ 可表述为 X 中最多 $n + 1$ 个点的凸组合, 这并不意味着存在 X 中 $n + 1$ 个点的“基”. 换句话说, X 中可能不存在 $n + 1$ 个固定不变的点 $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_n$, 使得

$$\text{conv} X = \text{conv}(\{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_n\}).$$

比如单位圆盘 $B(\mathbf{0}, 1)$. 与此形成对比的是, 在线性组合情况下, 一个 n 维子空间可由 n 个固定的向量(基)生成.

然而, 上述所说也不是绝对的. 在特定情况下, 存在由 X 中固定点组成的(可能是无限的)集合 X' 使得 $X = \text{conv}(X')$, 详见3.4节的Krein-Milman定理.

3.2 Helley定理

在3.4节证明齐次Farkas引理时要用Helley定理; 在5.4节证明Sion-Kakutani鞍点存在定理时, 需要用提高版 Helley 定理. 为了证明Helley定理, 需要用Radon定理. 出了用Radon定理来证明Helley定理, 也可以利用它证明 \mathbb{R}^n 中超平面族的VC维是 $n + 1$.

定理3.3 (Radon). 设 $\mathbf{x}_1, \dots, \mathbf{x}_m$ 是 \mathbb{R}^n 中的 m 个向量, 并且 $m \geq n + 2$. 可以将这些向量拆成两个非空且不重叠的子集 A, B 使得

$$\text{conv} A \cap \text{conv} B \neq \emptyset.$$

证明. 考虑关于 m 个变量 $\delta_1, \dots, \delta_m$ 的齐次线性方程组

$$\sum_{i=1}^m \delta_i \mathbf{x}_i = 0$$

$$\sum_{i=1}^m \delta_i = 0$$

因为 $m \geq n + 2$, 所以方程组有非平凡解 δ . 因为 $\delta \neq 0$ 并且 $\sum_{i=1}^m \delta_i = 0$, 所以可将指标集 $\{1, \dots, m\}$ 剖分成两个非空组

$$\mathcal{I} = \{i : \delta_i > 0\}, \quad \mathcal{J} = \{i : \delta_i \leq 0\}.$$

易见这两个指标集满足

$$\sum_{i \in \mathcal{I}} \delta_i \mathbf{x}_i = \sum_{j \in \mathcal{J}} [-\delta_j] \mathbf{x}_j$$

$$\gamma = \sum_{i \in \mathcal{I}} \delta_i = \sum_{j \in \mathcal{J}} (-\delta_j) > 0.$$

因此

$$\sum_{i \in \mathcal{I}} \frac{\delta_i}{\gamma} \mathbf{x}_i = \sum_{j \in \mathcal{J}} \left(\frac{-\delta_j}{\gamma} \right) \mathbf{x}_j.$$

令 $A = \{\mathbf{x}_i : i \in \mathcal{I}\}, B = \{\mathbf{x}_j : j \in \mathcal{J}\}$. 则上述等式左侧向量属于 $\text{conv} A$, 等式右侧向量属于 $\text{conv} B$. 从而定理成立. \square

定理3.4 (Helley). 设 A_1, \dots, A_M 是 \mathbb{R}^n 中的凸集. 假设这族集合中每 $n + 1$ 个集合有一个公共点. 则所有 M 个集合有一个公共点.

证明. 关于 M 进行归纳. 起点 $M = n + 1$ 是平凡成立的. 假设对某个 $M \geq n + 1$, 命题对凸集族中的每 M 个成员是成立的. 现在来证明命题对凸集族 A_1, \dots, A_M, A_{M+1} 中的 $M + 1$ 个成员也是成立的. 由归纳假设, $M + 1$ 个集合

$$A_{-\ell} = A_1 \cap A_2 \cap \dots \cap A_{\ell-1} \cap A_{\ell+1} \cap \dots \cap A_{M+1}$$

中的每个都是非空的. 由此可选择 $\mathbf{x}_\ell \in A_{-\ell}, \ell = 1, \dots, M + 1$. 由Radon定理, 可将 $\mathbf{x}_1, \dots, \mathbf{x}_{M+1}$ 拆分成凸包相交的两个子集. 不失一般性, 设拆分成 $\{\mathbf{x}_1, \dots, \mathbf{x}_{J-1}\}$ 和 $\{\mathbf{x}_J, \dots, \mathbf{x}_{M+1}\}$, 并设

$$\mathbf{z} \in (\text{conv}\{\mathbf{x}_1, \dots, \mathbf{x}_{J-1}\}) \cap (\text{conv}\{\mathbf{x}_J, \dots, \mathbf{x}_{M+1}\}).$$

由 A_ℓ 的构造和 \mathbf{x}_j 的选取, 知 $\mathbf{x}_j \in A_\ell, j \neq \ell$. 由此断言对所有 $\ell \leq M+1$, $\mathbf{z} \in A_\ell$. 的确, 对于 $\ell \leq J-1$, 点 $\mathbf{x}_J, \mathbf{x}_{J+1}, \dots, \mathbf{x}_{M+1}$ 属于凸集 A_ℓ , 因此

$$\mathbf{z} \in \text{conv}(\{\mathbf{x}_J, \dots, \mathbf{x}_{M+1}\}) \subset A_\ell.$$

对于 $\ell \geq J$, 点 $\mathbf{x}_1, \dots, \mathbf{x}_{J-1}$ 属于凸集 A_ℓ , 因此

$$\mathbf{z} \in \text{conv}(\{\mathbf{x}_1, \dots, \mathbf{x}_{J-1}\}) \subset A_\ell.$$

□

提高: 假设 A_1, \dots, A_M 是 \mathbb{R}^n 中的凸集. 如果集合之并 $A_1 \cup A_2 \cup \dots \cup A_M$ 属于仿射维数是 m 的仿射子空间 P , 并且这族集合中每 $n+1$ 个集合有一个公共点, 那么所有集合有一个属于 P 的公共点.

证明. 可将 A_j 作为 P 中的集合或者将 A_j 作为 \mathbb{R}^m 中的集合, 然后应用Helley定理可知结论成立. □

定理3.5 (Helley II). 设 $A_\alpha, \alpha \in \mathcal{A}$, 是 \mathbb{R}^n 中的一族凸集, 其满足每 $n+1$ 个集合有一个公共点. 此外假设集合 A_α 是闭的, 并且可以找到有限个集合 $A_{\alpha_1}, \dots, A_{\alpha_M}$ 的交集有界. 那么所有集合族 $A_\alpha, \alpha \in \mathcal{A}$, 有一个公共点.

证明. 首先由Helley定理, 集合 A_α 的每个有限组有一个公共元素. 下面利用分析中的标准事实: 设 B_α 是 \mathbb{R}^n 中的闭集族, 满足每个有限组有非空交集, 并且在集合族中存在有限组的交集是有界的. 那么该族中的所有集合有一个公共点. 基于 \mathbb{R}^n 中的每个有界闭子集是紧集这一基本性质可以证明这个标准事实.

首先回忆紧集的两个等价定义:

(a) 称度量空间 M 中的子集 X 是紧的, 如果 X 中的每个点列都能找到子列收敛到 X 中的某点.

(b) 称度量空间 M 中的子集 X 是紧的, 如果 X 的任何开覆盖(开集族使得 X 的每个点至少属于其中一个集合)都有有限子覆盖.

现在, 设 B_α 是 \mathbb{R}^n 中的闭集族, 满足任一有限子族的交集非空, 并且这些交集中至少有一个, 设为 B , 是有界的. 下面证明所有集合 B_α 有一个公共点. 假设不是这种情况. 那么对每个点 $\mathbf{x} \in B$, 都存在集合 B_α 不包含 \mathbf{x} . 因为 B_α 是闭的, 它会与一个中心在 \mathbf{x} 的恰当开球 V_x 不相交. 请注意 $\{V_x : \mathbf{x} \in B\}$ 是 B 的开覆盖.

由 B 的构造知, 它是闭集的交, 从而 B 是闭的, 并且 B 还是有界的, 从而是紧集. 因此可以找到有限个 $V_{\mathbf{x}_1}, \dots, V_{\mathbf{x}_M}$ 覆盖 B . 对每个 $i \leq M$, 在集族中存在 B_{α_i} 与 $V_{\mathbf{x}_i}$ 没有交集; 因此 $\bigcap_{i=1}^M B_{\alpha_i}$ 与 B 不相交. 由于 B 本身是有限个集合 B_α 的交集, 从而证明了有限个集合 B_α (那些参与描述 B 的和集合 $B_{\alpha_1}, \dots, B_{\alpha_M}$)的交集是空的, 这与已知矛盾. □

例3.3. 已知定义在7,000,000个点组成的集合 $X \subset \mathbb{R}$ 上的函数 $f(x)$. 在 X 的每7-点子集上, 可以用恰当的5次多项式逼近该函数, 使得在每个点的精度为0.001. 想在整个 X 上使用5次样条(每段是多项式函数且每段的次数是5)逼近函数. 请问需要多少段以保证在每个点的精度是0.001.

解. 仅需1个. 的确, 设 $A_x, x \in X$, 是所有以精度0.001来再生 $f(x)$ 的所有5次多项式系数的集合, 即

$$A_x = \left\{ \mathbf{p} = (p_0, p_1, \dots, p_5) \in \mathbb{R}^6 : |f(x) - \sum_{i=0}^5 p_i x^i| \leq 0.001 \right\}.$$

集合 A_x 是多面集, 因此是凸的, 且已知集族 $\{A_x\}_{x \in X}$ 中的每6 + 1 = 7个集合有一个公共点. 由Helly定理, 所有集合 $A_x, x \in X$ 有一个公共点, 这表明存在单个5次多项式在 X 的每个点近似 f 的精度是0.001.

例3.4. 将设计一个制造厂, 数学上可用如下线性规划来描述它:

$$\begin{aligned} \mathbf{A}\mathbf{x} &\geq \mathbf{d} \quad [d_1, \dots, d_{1000}: \text{需求}] \\ \mathbf{B}\mathbf{x} &\leq \mathbf{f} \quad [f_1 \geq 0, \dots, f_{10} \geq 0: \text{设备容量}] \\ \mathbf{C}\mathbf{x} &\leq \mathbf{c} \quad [\text{其它约束}] \end{aligned} \tag{F}$$

数据 $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{c}$ 是提前给定的. 应该提前确定设备容量 $f_i \geq 0, i = 1, \dots, 10$, 使得工厂能够满足从已知的有限集合 D 中取出的所有需求场景 \mathbf{d} , 即(F)对每个 $\mathbf{d} \in D$ 应该是可行的. 设创建第 i 个容量为 f_i 设备所需成本是 $a_i f_i$.

已知为了满足每个源于 D 的单一需求, 投资1美元创建设备是充分的. 现在的问题是: 当 D 包含

- (a) 仅有一个场景,
- (b) 3个场景,
- (c) 10个场景,
- (d) 2004个场景

时, 用于设备的投资应该是多少?

解. 考虑 D 的各种情况:

$$\begin{aligned} D = \{\mathbf{d}^1\} &\Rightarrow 1 \text{ 美元就够了} \\ D = \{\mathbf{d}^1, \mathbf{d}^2, \mathbf{d}^3\} &\Rightarrow 3 \text{ 美元就够了} \\ D = \{\mathbf{d}^1, \dots, \mathbf{d}^{10}\} &\Rightarrow 10 \text{ 美元就够了} \\ D = \{\mathbf{d}^1, \dots, \mathbf{d}^{2004}\} &\Rightarrow 11 \text{ 美元就够了} \end{aligned}$$

的确, 对于 $\mathbf{d} \in D$, 设 $F[\mathbf{d}]$ 是所有 $\mathbf{f} \in \mathbb{R}^{10}, \mathbf{f} \geq \mathbf{0}$ 中满足成本最多为11美元, 并对已知 \mathbf{d} 产生的关于变量 \mathbf{x} 的系统

$$\begin{aligned} A\mathbf{x} &\geq \mathbf{d} \\ B\mathbf{x} &\leq \mathbf{f} \\ C\mathbf{x} &\leq \mathbf{c} \end{aligned} \quad (F[\mathbf{d}])$$

是可解的 \mathbf{f} 组成的集合. 那么集合 $F[\mathbf{d}]$ 是凸的¹, 且每11个这种类型的集合有一个公共点. 的确, 已知 D 中的11个场景 $\mathbf{d}^1, \dots, \mathbf{d}^{11}$, 能够以成本1美元和恰当的 \mathbf{f}^i 来“物化” \mathbf{d}^i , 因此能用11美元的成本及容量为 $\mathbf{f}^1 + \dots + \mathbf{f}^{11}$ 的单个向量来“物化”11个场景 $\mathbf{d}^1, \dots, \mathbf{d}^{11}$ 中的每一个, 因此该向量属于 $F[\mathbf{d}^1], \dots, F[\mathbf{d}^{11}]$.

因为2004个凸集 $F[\mathbf{d}] \subset \mathbb{R}^{10}, \mathbf{d} \in D$ 中的每11个有一个公共点, 所以这些集合有公共点 \mathbf{f} ; 对于这个 \mathbf{f} , 系统 $F[\mathbf{d}], \mathbf{d} \in D$ 中的每个都是可解的.

例3.5. 考虑具有11个变量 x_1, \dots, x_{11} 的优化问题

$$c_* = \min\{\mathbf{c}^T \mathbf{x} : g_i(\mathbf{x}) \leq 0, i = 1, \dots, 2004\}.$$

假设约束是凸的, 即集合

$$X_i = \{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) \leq 0\}, i = 1, \dots, 2004$$

中的每个都是凸的. 也假设问题的最优值是0, 且是可解的.

显然, 当去掉一个或者更多的约束时, 最优值仅能减小或者保持不变. 有可能找到一个约束使得去掉它后, 最优值保持不变? 可否同时去掉两个约束并且不影响最优值? 类似地, 可否去掉三个?

解. 可以去掉 $2004 - 11 = 1993$ 个精心选取的约束同时不改变最优值!

假设, 相反地原始问题每11个约束得到的松弛问题的最优值是负的. 因为这种松弛是有限的, 所以存在 $\epsilon > 0$ 使得每个形如

$$\min_{\mathbf{x}} \{\mathbf{c}^T \mathbf{x} : g_{i_1}(\mathbf{x}) \leq 0, \dots, g_{i_{11}}(\mathbf{x}) \leq 0\}.$$

的问题有一个可行解, 其对应的目标值小于 $-\epsilon$. 由于该问题存在一个目标值为0的可行解(即原始问题的最优解), 且它的可行集是凸的, 所以问题有可行解 \mathbf{x} 使得 $\mathbf{c}^T \mathbf{x} = -\epsilon$. 换句话说, 2004个集合

$$Y_i = \{\mathbf{x} \in \mathbb{R}^{11} : \mathbf{c}^T \mathbf{x} = -\epsilon, g_i(\mathbf{x}) \leq 0\}, i = 1, \dots, 2004$$

中的每11个集合有一个公共点.

因为集合 Y_i 是凸集 X_i 和仿射子空间的交, 所以是凸的. 如果 $c \neq 0$, 则这些集合属于一个仿射维数为10的仿射子空间, 且由于它们中的每11个相交, 从而所有这2004个相交; 取交集中的一点 \mathbf{x} , 则 \mathbf{x} 是原始问题的可行解且 $\mathbf{c}^T \mathbf{x} < 0$, 这是不可能的. 当 $c = 0$ 时, 可以去掉所有的2004个约束而不改变最优值.

¹可用3.3节的结论说明.

3.3 多面集性与Fourier-Motzkin消元

定义3.3. 多面集 $X \subset \mathbb{R}^n$ 是能表示成有限个非严格线性不等式组的解集的集合:

$$X = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}.$$

定义3.4. 集合 $X \subset \mathbb{R}^n$ 的多面集表示(polyhedral representation) 是 X 的形如

$$X = \{\mathbf{x} \in \mathbb{R}^n : \exists \mathbf{w} \text{ s.t. } \mathbf{P}\mathbf{x} + \mathbf{Q}\mathbf{w} \leq \mathbf{r}\} \quad (3.1)$$

的表示, 即 X 是 (\mathbf{x}, \mathbf{w}) -变量空间的多面集

$$Y = \{(\mathbf{x}, \mathbf{w}) : \mathbf{P}\mathbf{x} + \mathbf{Q}\mathbf{w} \leq \mathbf{r}\} \quad (3.2)$$

在 \mathbf{x} -变量空间的投影.

例3.6. 下面是一些多面集表示的例子:

(i) 集合 $X = \{\mathbf{x} \in \mathbb{R}^n : \sum_i |x_i| \leq 1\}$ 具有多面集表示

$$X = \left\{ \mathbf{x} \in \mathbb{R}^n : \exists \mathbf{w} \in \mathbb{R}^n \text{ s.t. } -w_i \leq x_i \leq w_i, 1 \leq i \leq n, \sum_i w_i \leq 1 \right\}.$$

(ii) 集合

$$X = \{\mathbf{x} \in \mathbb{R}^6 : \max\{x_1, x_2, x_3\} + 2 \max\{x_4, x_5, x_6\} \leq x_1 - x_6 + 5\}$$

具有多面集表示

$$X = \left\{ \mathbf{x} \in \mathbb{R}^6 : \exists \mathbf{w} \in \mathbb{R}^2 \text{ s.t. } \begin{array}{l} x_1 \leq w_1, x_2 \leq w_1, x_3 \leq w_1 \\ x_4 \leq w_2, x_5 \leq w_2, x_6 \leq w_2 \\ w_1 + 2w_2 \leq x_1 - x_6 + 5 \end{array} \right\}.$$

设 X 是由多面集表示(3.1)给出的, 即 X 是关于 (\mathbf{x}, \mathbf{w}) 变量的有限个线性不等式组的解集(3.2)在 \mathbf{x} -变量空间的投影. 那么 X 是多面集吗? 就是 X 是一个仅关于 \mathbf{x} 的有限个不等式组的解集吗?

定理3.6. 每个多面集可表示集合(3.1)是多面集.

证明. 由以下的Fourier-Motzkin消元机制可证明(3.2) 在 \mathbf{x} -变量空间的投影是多面集.

消元步: 消去单个松弛变量. 给定(3.2), 假设 $\mathbf{w} = (w_1, \dots, w_m)$ 是非空的, $m \geq 1$, 并设

$$Y = \{(\mathbf{x}, w_1, \dots, w_{m-1}) : \exists w_m \text{ s.t. } \mathbf{P}\mathbf{x} + \mathbf{Q}\mathbf{w} \leq \mathbf{r}\} \quad (3.3)$$

是(3.2)在 $\mathbf{x}, w_1, \dots, w_{m-1}$ 变量空间的投影. 的确, 根据 w_m 的系数 q_{im} 的符号可将定义 Y 的线性不等式

$$\mathbf{p}_i^T \mathbf{x} + \mathbf{q}_i^T \mathbf{w} \leq r_i, 1 \leq i \leq I$$

拆分成三组:

$$\mathcal{I}_0 := \{i \in \{1, \dots, I\} : q_{im} = 0\},$$

$$\mathcal{I}_+ := \{i \in \{1, \dots, I\} : q_{im} > 0\},$$

$$\mathcal{I}_- := \{i \in \{1, \dots, I\} : q_{im} < 0\}.$$

记 $\mathbf{y} = (w_1, \dots, w_{m-1})$. 当 $i \in \mathcal{I}_0$ 时, 记

$$c_i = r_i, \mathbf{a}_i = \mathbf{p}_i, \mathbf{b}_i = (q_{i1}, \dots, q_{i(m-1)});$$

当 $i \in \mathcal{I}_+ \cup \mathcal{I}_-$ 时, 记

$$c_i = q_{im}^{-1} r_i, \mathbf{a}_i = -q_{im}^{-1} \mathbf{p}_i, \mathbf{b}_i = -q_{im}^{-1} (q_{i1}, \dots, q_{i(m-1)}).$$

那么 $\mathbf{p}_i^T \mathbf{x} + \mathbf{q}_i^T \mathbf{w} \leq r_i$ 分别等价于

$$\mathbf{a}_i^T \mathbf{x} + \mathbf{b}_i^T \mathbf{y} \leq c_i, i \in \mathcal{I}_0,$$

$$w_m \leq \mathbf{a}_i^T \mathbf{x} + \mathbf{b}_i^T \mathbf{y} + c_i, i \in \mathcal{I}_+,$$

$$w_m \geq \mathbf{a}_i^T \mathbf{x} + \mathbf{b}_i^T \mathbf{y} + c_i, i \in \mathcal{I}_-.$$

从而

$$Y^+ = \left\{ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^{m-1} : \begin{array}{ll} \mathbf{a}_i^T \mathbf{x} + \mathbf{b}_i^T \mathbf{y} \leq c_i, & i \in \mathcal{I}_0; \\ \mathbf{a}_j^T \mathbf{x} + \mathbf{b}_j^T \mathbf{y} + c_j \geq \mathbf{a}_k^T \mathbf{x} + \mathbf{b}_k^T \mathbf{y} + c_k, & j \in \mathcal{I}_+, k \in \mathcal{I}_- \end{array} \right\}.$$

易见 Y^+ 是多面集. 这样, 看到了多面集(3.3)的投影 Y^+ 是多面集. 迭代该过程, 得到集合(3.1)是多面集. \square

已知线性规划问题

$$\text{Opt} = \max_{\mathbf{x}} \{\mathbf{c}^T \mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}, \quad (3.4)$$

观察到可行解的目标值形成的集合可被表示成

$$\begin{aligned} T &= \{\tau \in \mathbb{R} : \exists \mathbf{x} \text{ s.t. } \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{c}^T \mathbf{x} - \tau = 0\} \\ &= \{\tau \in \mathbb{R} : \exists \mathbf{x} \text{ s.t. } \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{c}^T \mathbf{x} \leq \tau, \mathbf{c}^T \mathbf{x} \geq \tau\}, \end{aligned} \quad (3.5)$$

即 T 是多面集可表示的. 由定理知 T 可表示为有限个仅关于变量 τ 的线性不等式组的解集. 由此立即得到: 如果 T 非空有上界, 那么 T 有最大元素. 这样, 证明了以下推论:

推论3.7. 可行且有界的线性规划问题(3.4)有最优解, 因此是可解的.

Fourier-Motzkin消元机制表明了一个求解线性规划问题(3.4)的步骤有限的算法, 其中首先应用这种机制将 T 表示成关于变量 τ 的有限个不等式组的解集 S ; 其次, 分析 S 以确实它是否非空且有上界. 若 S 是非空且有上界的, 找到规划(3.4)的最优值 $\text{Opt} \in T$. 然后, 以反方向回退方式使用Fourier-Motzkin消元机制, 找到 \mathbf{x} 使得 $\mathbf{Ax} \leq \mathbf{b}$ 且 $\mathbf{c}^T \mathbf{x} = \text{Opt}$, 这样就恢复了感兴趣问题的最优解. 坏消息是所得算法是根本不实用, 因为每一步需要处理的不等式的个数通常会快速增长, 当消去几十个变量时, 就需要处理天文数字般大的不等式约束. 综上, 上述线性规划的求解方法是构造性的. 但是具体的构造方法并不是实用算法.

3.4 择一定理

考虑齐次线性不等式

$$\mathbf{a}^T \mathbf{x} \geq 0 \quad (3.6)$$

和类似的有限个不等式组

$$\mathbf{a}_i^T \mathbf{x} \geq 0, \quad 1 \leq i \leq m. \quad (3.7)$$

关心的问题是: 何时(3.6)是(3.7)的后果, 即每个满足(3.7)的 \mathbf{x} 也满足(3.6)? 仔细观察, 发现如果 \mathbf{a} 是 $\mathbf{a}_1, \dots, \mathbf{a}_m$ 的锥组合, 即

$$\exists \lambda_i \geq 0 \text{ s.t. } \mathbf{a} = \sum_i \lambda_i \mathbf{a}_i, \quad (3.8)$$

则(3.6)是(3.7)的后果. 的确, (3.8)蕴含着

$$\mathbf{a}^T \mathbf{x} = \sum_i \lambda_i \mathbf{a}_i^T \mathbf{x}, \quad \forall \mathbf{x},$$

从而对满足 $\mathbf{a}_i^T \mathbf{x} \geq 0$ 的 \mathbf{x} , 有 $\mathbf{a}^T \mathbf{x} \geq 0$.

定理3.8 (齐次Farkas引理). 系统(3.6)是(3.7)的后果当且仅当 \mathbf{a} 是 $\mathbf{a}_1, \dots, \mathbf{a}_m$ 的锥组合.²

已知向量 $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{R}^n$, 设

$$K = \text{cone}\{\mathbf{a}_1, \dots, \mathbf{a}_m\} = \left\{ \sum_{i=1}^m \lambda_i \mathbf{a}_i : \lambda \in \mathbb{R}_+^m \right\} \quad (3.9)$$

²系统(3.6)是(3.7)的结果等价于(3.7)与 $\mathbf{a}^T \mathbf{x} < 0$ 组成的系统无解. 所以齐次Farkas引理即该系统无解当且仅当系统(3.8)有界. 从而一个系统无解当且仅当关联的系统有解, 从而是一种形式的择一定理.

是已知向量的锥包. 已知向量 \mathbf{a} , 易于验证 $\mathbf{a} \in \text{cone}\{\mathbf{a}_1, \dots, \mathbf{a}_m\}$: 找到非负权重向量 $\lambda \in \mathbb{R}_+^m$ 使得 $\sum_{i=1}^m \lambda_i \mathbf{a}_i$ 即可认证. 易于验证 $\mathbf{a} \notin \text{cone}\{\mathbf{a}_1, \dots, \mathbf{a}_m\}$: 找到向量 $\mathbf{d} \in \mathbb{R}^n$ 使得 $\mathbf{a}_i^T \mathbf{d} \geq 0, \forall i$ 且 $\mathbf{a}^T \mathbf{d} < 0$ 是一种认证.

证明. 仅需证明如果 \mathbf{a} 不是 $\mathbf{a}_1, \dots, \mathbf{a}_m$ 的锥组合, 则存在 \mathbf{d} 使得

$$\mathbf{d}^T \mathbf{a} < 0, \quad \mathbf{d}^T \mathbf{a}_i \geq 0, i = 1, \dots, m, \quad (3.10)$$

即存在超平面 $\mathbf{d}^T \mathbf{x} = 0$ 严格分离点 \mathbf{a} 和锥包 K . 因为

$$K = \text{cone}\{\mathbf{a}_1, \dots, \mathbf{a}_m\} = \{\mathbf{x} \in \mathbb{R}^n : \exists \lambda \in \mathbb{R}^m \text{ s.t. } \mathbf{x} = \sum_{i=1}^m \lambda_i \mathbf{a}_i, \lambda \geq 0\},$$

所以集合(3.9)是多面集可表示的. 由Fourier-Motzkin定理, K 是多面集, 即可将 K 表示为

$$K = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{d}_\ell^T \mathbf{x} \geq c_\ell, 1 \leq \ell \leq L\}. \quad (3.11)$$

首先观察到 $\mathbf{0} \in K$, 由此知 $c_\ell \leq 0, \forall \ell$. 其次, 观察到

$$\forall \lambda > 0, \lambda \mathbf{a}_i \in \text{cone}\{\mathbf{a}_1, \dots, \mathbf{a}_m\},$$

所以 $\forall \lambda > 0, \lambda \mathbf{d}_\ell^T \mathbf{a}_i \geq c_\ell$. 由此得 $\mathbf{d}_\ell^T \mathbf{a}_i \geq 0, \forall \ell$. 现在, $\mathbf{a} \notin \text{cone}\{\mathbf{a}_1, \dots, \mathbf{a}_m\}$, 从而由 K 的多面集表示(3.11) 知存在 ℓ_* 使得 $\mathbf{a}^T \mathbf{d}_{\ell_*} < c_{\ell_*} \leq 0$. 所以 $\mathbf{a}^T \mathbf{d}_{\ell_*} < 0$. 这样,

$$\mathbf{d} = \mathbf{d}_{\ell_*}$$

满足(3.10). □

利用齐次Farkas引理可以证明下面的广义择一定理和5.4节的Sion-Kakutani鞍点存在定理. 可将一般的以 $\mathbf{x} \in \mathbb{R}^n$ 为未知数的有限线性不等式组写为

$$\begin{aligned} \mathbf{a}_i^T \mathbf{x} &> b_i, i = 1, \dots, m_s \\ \mathbf{a}_i^T \mathbf{x} &\geq b_i, i = m_s + 1, \dots, m. \end{aligned} \quad (\text{S})$$

为了验证(S)是可解的, 找到一个解即可! 现在的问题是如何验证(S) 是不可解的.

考虑用反证法. 即假设 \mathbf{x} 是(S)的解, 则通过线性聚合可得到矛盾: 设(S) 的第 i 个不等式与非负权重 λ_i 对应, 给不等式的两边同时乘以对应权重并求和, 由 λ_i 非负, 对(S) 的每个解 \mathbf{x} 得到不等式

$$\left(\sum_{i=1}^m \lambda_i \mathbf{a}_i \right)^T \mathbf{x} \begin{cases} > \sum_i \lambda_i b_i, \sum_{i=1}^{m_s} \lambda_i > 0 \\ \geq \sum_i \lambda_i b_i, \sum_{i=1}^{m_s} \lambda_i = 0. \end{cases} \quad (\text{C})$$

因此, 如果存在 $\lambda \geq 0$ 使得(C)压根没有解, 那么(S)无解. 现在问题转化成何时线性不等式

$$\mathbf{d}^T \mathbf{x} \begin{cases} > \\ \geq \end{cases} e \quad (3.12)$$

压根没有解? 线性不等式(3.12)无解当且仅当 $d = 0$ 且(i) 或者符号是“ $>$ ”且 $e \geq 0$, (ii) 或者符号是“ \geq ”且 $e > 0$. 从而, 当考虑关于变量 \mathbf{x} 的线性不等式组(S)时, 将其与两个关于变量 λ 的线性不等式组

$$\mathcal{T}_I : \begin{cases} \lambda \geq 0, \sum_{i=1}^{m_s} \lambda_i > 0 \\ \sum_{i=1}^m \lambda_i \mathbf{a}_i = 0 \\ \sum_{i=1}^m \lambda_i b_i \geq 0 \end{cases} \quad \mathcal{T}_{II} : \begin{cases} \lambda \geq 0, \sum_{i=1}^{m_s} \lambda_i = 0 \\ \sum_{i=1}^m \lambda_i \mathbf{a}_i = 0 \\ \sum_{i=1}^m \lambda_i b_i > 0 \end{cases} \quad (3.13)$$

关联. 如果系统 $\mathcal{T}_I, \mathcal{T}_{II}$ 中的之一可解, 那么(S)不可解. 请注意, 如果 \mathcal{T}_{II} 是可解的, 则系统

$$\mathbf{a}_i^T \mathbf{x} \geq b_i, i = m_s + 1, \dots, m \quad (3.14)$$

是不可解的! 进一步, (S)的子系统(3.14)不可解当且仅当 \mathcal{T}_{II} 是可解的.

定理3.9 (广义择一定理). 线性不等式组(S)不可解当且仅当(3.13)中的系统之一是可解的.

证明. 已经知道系统 $\mathcal{T}_I, \mathcal{T}_{II}$ 之一可解是(S)不可解的充分条件. 仅需证明的是如果(S)不可解, 那么系统 $\mathcal{T}_I, \mathcal{T}_{II}$ 之一可解. 假设关于变量 \mathbf{x} 的系统(S)无解. 那么齐次不等式组

$$\begin{aligned} \tau - \epsilon &\geq 0, \\ \mathbf{a}_i^T \mathbf{x} - b_i \tau - \epsilon &\geq 0, i = 1, \dots, m_s \\ \mathbf{a}_i^T \mathbf{x} - b_i \tau &\geq 0, i = m_s + 1, \dots, m \end{aligned} \quad (U)$$

的每个解 $\mathbf{x}, \tau, \epsilon$ 有 $\epsilon \leq 0$. 的确, 对于一个其中 $\epsilon > 0$ 的解, 也能得到 $\tau > 0$, 并且向量 $\tau^{-1} \mathbf{x}$ 是(S)的解. 假设齐次不等式组(U)的每个解有 $\epsilon \leq 0$, 即齐次不等式

$$-\epsilon \geq 0 \quad (I)$$

是齐次不等式组(U)的结果, 那么由齐次Farkas引理, 不等式(I)左侧的系数向量是(U)的左侧系数向量的锥组合, 即 $\exists \lambda \geq 0, \nu \geq 0$ 满足

$$\begin{aligned} \sum_{i=1}^m \lambda_i \mathbf{a}_i &= 0 \\ -\sum_{i=1}^m \lambda_i b_i + \nu &= 0 \\ -\sum_{i=1}^{m_s} \lambda_i - \nu &= -1 \end{aligned}$$

假设 $\lambda_1 = \cdots = \lambda_s = 0$, 由第三个等式得到 $\nu = 1$. 因此由前两个等式知 λ 求解 \mathcal{T}_{Π} . 在 $\sum_{i=1}^{m_s} \lambda_i > 0$ 的情况下, λ 显然求解 \mathcal{T}_I . \square

推论3.10. 有限个线性不等式组无解当且仅当对它进行线性聚合得到矛盾, 即用“合法”权重得到不等式的一个恰当组合或者与不等式

$$\mathbf{0}^T \mathbf{x} > a \quad [a \geq 0]$$

矛盾或者与不等式

$$\mathbf{0}^T \mathbf{x} \geq a \quad [a > 0]$$

矛盾.

推论3.11 (非齐次Farkas引理). 线性不等式

$$\mathbf{a}^T \mathbf{x} \leq b \quad (3.15)$$

是可解线性不等式组

$$\mathbf{a}_i^T \mathbf{x} \leq b_i, i = 1, \cdots, m$$

的结果当且仅当通过线性聚合能得到目标不等式(3.15)和恒成立不等式

$$\mathbf{0}^T \mathbf{x} \leq 1,$$

即当且仅当存在非负 $\lambda_0, \lambda_1, \cdots, \lambda_m$ 使得

$$\begin{aligned} \mathbf{a} &= \sum_{i=1}^m \lambda_i \mathbf{a}_i \\ b &= \lambda_0 + \sum_{i=1}^m \lambda_i b_i \end{aligned} \quad \Leftrightarrow \quad \begin{cases} \mathbf{a} = \sum_{i=1}^m \lambda_i \mathbf{a}_i \\ b \geq \sum_{i=1}^m \lambda_i b_i \end{cases}$$

后一小节证明线性规划的对偶定理时, 将用到非齐次Farkas引理.

3.5 线性规划的对偶

线性规划问题

$$\text{Opt}(P) = \min_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A}\mathbf{x} \geq \mathbf{b} \}, \quad (\text{P})$$

的对偶规划源于寻找一种为(P)的最优值确定下界的系统方式. 概念上最简单的定界方式是将约束不等式进行线性聚合. 经观察知对每个非负权重向量 λ , 约束

$$[\mathbf{A}^T \lambda]^T \mathbf{x} = \lambda^T \mathbf{A}\mathbf{x} \geq \lambda^T \mathbf{b}$$

是(P)的约束的结果, 因此该约束在(P)的每个可行解处都满足.

命题3.12. 对每个满足 $\mathbf{A}^T \lambda = \mathbf{c}$ 的向量 $\lambda \geq \mathbf{0}$, $\lambda^T \mathbf{b}$ 是 $\text{Opt}(P)$ 的下界.

原始问题(P)的对偶问题是

$$\text{Opt}(D) = \max_{\lambda} \{ \mathbf{b}^T \boldsymbol{\lambda} : \lambda \geq 0, \mathbf{A}^T \boldsymbol{\lambda} = \mathbf{c} \}, \quad (\text{D})$$

即最大化由命题3.12给出的 $\text{Opt}(P)$ 的下界. (D) 的来源蕴含着如下事实.

定理3.13 (弱对偶性). 原始问题(P)在每个可行解处的原始目标值大于等于对偶问题(D) 在每一个可行解处的目标值, 即 \mathbf{x} 和 $\boldsymbol{\lambda}$ 分别是(P) 和(D) 的可行解, 那么

$$\mathbf{c}^T \mathbf{x} \geq \mathbf{b}^T \boldsymbol{\lambda}.$$

特别地,

$$\text{Opt}(P) \geq \text{Opt}(D).$$

定理3.14 (线性规划对偶性). 考虑线性规划问题(P)及其对偶(D), 那么

- (i) 对偶是对称的: 对偶的对偶问题是(等价于)原始问题.
- (ii) 每个对偶可行解处的对偶目标值小于等于每个原始可行解处的原始目标值.
- (iii) 下面的五个性质是相互等价的:
 - (a) 原始问题(P)是可行有(下)界的,
 - (b) 对偶问题(D)是可行有(上)界的,
 - (c) 原始问题(P)是可解的,
 - (d) 对偶问题(D)是可解的,
 - (e) 原始问题(P)和对偶问题(D)都是可行的,

并且只要它们发生, 就有 $\text{Opt}(P) = \text{Opt}(D)$.

证明. (i) 将(D)重新写成(P)的形式, 得到

$$\min_{\lambda} \left\{ -\mathbf{b}^T \boldsymbol{\lambda} : \begin{bmatrix} \mathbf{A}^T \\ -\mathbf{A}^T \\ I \end{bmatrix} \boldsymbol{\lambda} \geq \begin{bmatrix} \mathbf{c} \\ -\mathbf{c} \\ \mathbf{0} \end{bmatrix} \right\},$$

由此得到该问题的对偶是

$$\max_{\mathbf{u}, \mathbf{v}, \mathbf{w}} \{ \mathbf{c}^T \mathbf{u} - \mathbf{c}^T \mathbf{v} + \mathbf{0}^T \mathbf{w} : \mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0}, \mathbf{w} \geq \mathbf{0}, \mathbf{A}\mathbf{u} - \mathbf{A}\mathbf{v} + \mathbf{w} = -\mathbf{b} \}$$

这等价于

$$\max_{\mathbf{x}=\mathbf{v}-\mathbf{u}, \mathbf{w}} \{ -\mathbf{c}^T \mathbf{x} : \mathbf{w} \geq \mathbf{0}, \mathbf{A}\mathbf{x} = \mathbf{b} + \mathbf{w} \}$$

该问题等价于

$$\max_{\mathbf{x}} \{-\mathbf{c}^T \mathbf{x} : \mathbf{A}\mathbf{x} \geq \mathbf{b}\},$$

该问题等价于(P).

(ii) 即为弱对偶定理.

(iii) (a) \Rightarrow (d). 由Opt(P)的定义, 不等式

$$\mathbf{c}^T \mathbf{x} \geq \text{Opt}(P)$$

是不等式组

$$\mathbf{A}\mathbf{x} \geq \mathbf{b}$$

的结果. 由非齐次Farkas引理(推论3.11)知

$$\exists \boldsymbol{\lambda} \geq \mathbf{0} : \mathbf{A}^T \boldsymbol{\lambda} = \mathbf{c} \ \& \ \mathbf{b}^T \boldsymbol{\lambda} \geq \text{Opt}(P).$$

因此对偶问题有可行解并且其对偶目标值大于等于Opt(P). 由弱对偶性, 这个解也是最优的, 并且Opt(D) = Opt(P).

(d) \Rightarrow (b)是显然的.

(b) \Rightarrow (c). 由原始一对偶的对称性和(a) \Rightarrow (d)成立知(b) \Rightarrow (c)成立.

(c) \Rightarrow (a)是显然的.

综上, 证明了(a) \Leftrightarrow (b) \Leftrightarrow (c) \Leftrightarrow (d), 并且当这四个等价性质发生时Opt(D) = Opt(P).

剩下的就是证明性质(a)-(d)等价于(e). 当(e)成立时, 由弱对偶性知(P)是可行且有下界的, 因此(e) \Rightarrow (a). 当(a)成立时, 因为(a)与(b)等价, 所以(b)也成立, 从而(P)和(D)都是可行的. 因此(a) \Rightarrow (e). \square

定理3.15 (线性规划的最优性条件). 考虑均可行的线性规划原始一对偶对(P)和(D), 并假设 \mathbf{x} 和 $\boldsymbol{\lambda}$ 是各自的可行解. 这些解是各自的最优解当且仅当

$$\mathbf{c}^T \mathbf{x} - \mathbf{b}^T \boldsymbol{\lambda} = 0 \text{ [零对偶间隙]}$$

和当且仅当

$$(\mathbf{A}\mathbf{x} - \mathbf{b})_i \cdot \lambda_i = 0, i = 1, \dots, m. \text{ [互补松弛性]}$$

证明. 在定理的已知条件下, 由定理(3.14)的(iii)知Opt(P) = Opt(D). 因此

$$\mathbf{c}^T \mathbf{x} - \mathbf{b}^T \boldsymbol{\lambda} = \underbrace{\mathbf{c}^T \mathbf{x} - \text{Opt}(P)}_{\geq 0} + \underbrace{\text{Opt}(D) - \mathbf{b}^T \boldsymbol{\lambda}}_{\geq 0}.$$

这样, 对偶间隙总是非负的, 其等于零当且仅当 \mathbf{x} 和 $\boldsymbol{\lambda}$ 是各自问题的最优解. 因为 $\mathbf{A}\mathbf{x} - \mathbf{b}$ 和 $\boldsymbol{\lambda}$ 都是非负的, 由恒等式

$$\mathbf{c}^T \mathbf{x} - \mathbf{b}^T \boldsymbol{\lambda} = (\mathbf{A}^T \boldsymbol{\lambda})^T \mathbf{x} - \mathbf{b}^T \boldsymbol{\lambda} = (\mathbf{A}\mathbf{x} - \mathbf{b})^T \boldsymbol{\lambda}$$

可得到互补松弛条件, 并且对偶间隙为零当且仅当互补松弛性成立. \square

3.6 凸集分离定理

\mathbb{R}^n 上的每个线性形 $f(\mathbf{x})$ 可用内积表示为

$$f(\mathbf{x}) = \mathbf{f}^T \mathbf{x},$$

这里的 f 是由 $f(\mathbf{x})$ 唯一确定的恰当向量. 非平凡(不恒等于零)线性形与 \mathbb{R}^n 上的非零向量 f 对应. \mathbb{R}^n 上非平凡线性形的等高线

$$M = \{\mathbf{x} : \mathbf{f}^T \mathbf{x} = a\} \quad (\text{HL})$$

是仿射维数为 $n-1$ 的仿射子空间; 反之亦然; 具体地, 通过选取合适的 $f \neq 0$ 和 a , \mathbb{R}^n 上每个仿射维数为 $n-1$ 的仿射子空间 M 可表述为(HL); f 和 a 由 M 定义, 最多乘以非零公共因子. 称 \mathbb{R}^n 上的 $(n-1)$ -维仿射子空间是超平面(hyperplane). 非平凡线性形的等高线(HL)把 \mathbb{R}^n 分成两部分:

$$M_+ = \{\mathbf{x} : \mathbf{f}^T \mathbf{x} \geq a\}, M_- = \{\mathbf{x} : \mathbf{f}^T \mathbf{x} \leq a\}$$

称它们是由 (f, a) 确定的闭半空间. 超平面 M 是这些闭半空间的公共边界. M_+ 的内部 M_{++} 和 M_- 的内部 M_{--} 分别表示为

$$M_{++} = \{\mathbf{x} : \mathbf{f}^T \mathbf{x} > a\}, M_{--} = \{\mathbf{x} : \mathbf{f}^T \mathbf{x} < a\}.$$

称它们是由 (f, a) 给定的开半空间. 有

$$\mathbb{R}^n = M_- \cup M_+, [M_- \cap M_+ = M], \mathbb{R}^n = M_{--} \cup M \cup M_{++}.$$

定义3.5. 设 S, T 是 \mathbb{R}^n 上的两个非空集合. 称超平面(HL)分离 S 与 T , 如果

$$S \subset M_-, T \subset M_+ (S \text{不在} M \text{之上}, T \text{不在} M \text{之下}), \text{并且 } S \cup T \not\subset M.$$

进一步, 称非平凡线性形 $\mathbf{f}^T \mathbf{x}$ 分离 S 和 T , 如果对于恰当选取的 a , 超平面(HL)分离 S 和 T .

例3.7. 考虑 \mathbb{R}^2 上的线性形 x_1 , 易见它

(i) 分离集合 $S = \{\mathbf{x} \in \mathbb{R}^2 : x_1 \leq 0, x_2 \leq 0\}$ 和 $T = \{\mathbf{x} \in \mathbb{R}^2 : x_1 \geq 0, x_2 \geq 0\}$.

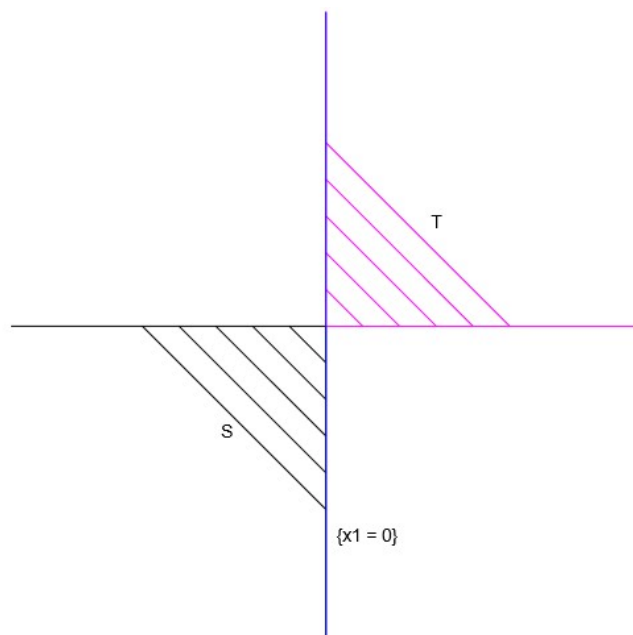


图 3.1: 超平面 $x_1 = 0$ 能分离 S 和 T .

(ii) 分离集合 $S = \{\mathbf{x} \in \mathbb{R}^2 : x_1 \leq 0, x_2 \leq 0\}$ 和 $T = \{\mathbf{x} \in \mathbb{R}^2 : x_1 + x_2 \geq 0, x_2 \leq 0\}$.

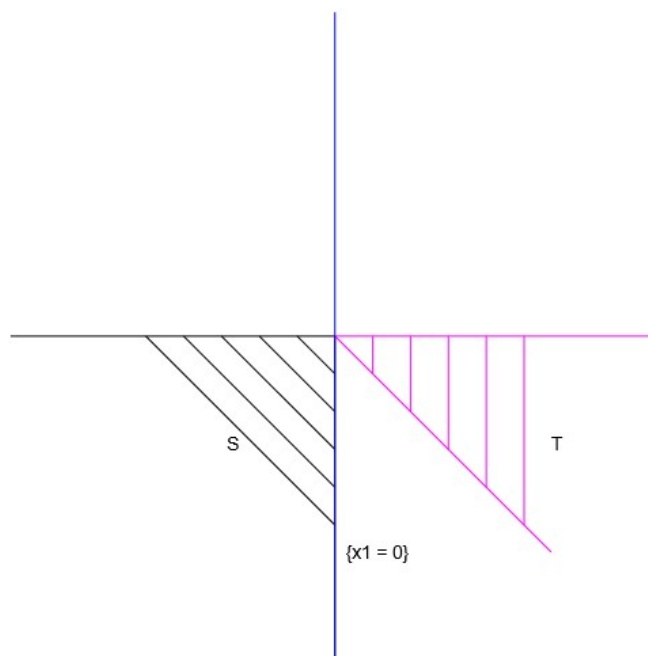


图 3.2: 超平面 $x_1 = 0$ 不能分离 S 和 T .

(iii) 不能分离集合 $S = \{\mathbf{x} \in \mathbb{R}^2 : x_1 = 0, 1 \leq x_2 \leq 2\}$ 和 $T = \{\mathbf{x} \in \mathbb{R}^2 : x_1 = 0, -2 \leq x_2 \leq -1\}$.



图 3.3: 任何形如 $x_1 = a$ 的超平面都不能分离 S 和 T .

经观察, 发现线性形 $f^T x$ 分离非空集合 S 和 T 当且仅当

$$\sup_{x \in S} f^T x \leq \inf_{y \in T} f^T y, \quad \inf_{x \in S} f^T x < \sup_{y \in T} f^T y. \quad (3.16)$$

成立. 在(3.16)的场景下, 分离 S 与 T 的(与 f 相关联的)超平面恰好是

$$\{x : f^T x = a\},$$

这里 $\sup_{x \in S} f^T x \leq a \leq \inf_{y \in T} f^T y$.

下面的凸集分离定理表明: 两个非空凸集可分离的充要条件是它们的相对内部没有公共点. 为了证明这个至关重要的定理, 需要如下两个引理, 其中证明必要性时需要第一个引理(推广版本见命题4.14 (a)), 证明充分性时需要第二个引理.

引理3.16. 设 X 是凸集, $f(x) = f^T x$ 是线性形, 且 $a \in \text{rint} X$. 那么

$$f^T a = \max_{x \in X} f^T x \quad \text{当且仅当} \quad f(\cdot)|_X = \text{const.}$$

证明. 不失一般性, 假设 $a = 0$ (如有必要, 平移 X 即可). 下面用反证法证明结论. 假设 $f^T x$ 在 X 上不是常数, 因此存在 $y \in X$ 使得

$$f^T y \neq f^T a = 0.$$

由于 $f^T x$ 在 X 上的最大值 $f^T a = 0$, 所以 $f^T y > 0$ 的情况是不可能的. 这样, $f^T y < 0$. 因为通过 0 和 y 的直线 $\{ty : t \in \mathbb{R}\}$ 包含于 $\text{aff} X$; 因为 $0 \in \text{rint} X$, 假如 $\epsilon > 0$ 充分小, 这条线上的所有点 $z = -\epsilon y$ 属于 X . 在每个这种类型的点处有 $f^T z > 0$, 这与事实

$$\max_{x \in X} f^T x = f^T a = 0$$

矛盾. □

引理3.17. \mathbb{R}^n 上每个非空子集 S 都是可分的: 能找到由 S 中的点组成的序列 $\{x_i\}$, 并且该序列在 S 里是稠密的, 即任何点 $x \in S$ 均是该序列中某个子序列的极限.

证明. 设 r_1, r_2, \dots 是 \mathbb{R}^n 里所有有理向量组成的可数集合. 对于任何正整数 t , 设 $X_t \subset S$ 是按如下方式构造得到的可数集:

在点 r_1, r_2, \dots 处逐点依次检查: 对于点 r_s , 检查是否存在 S 中的点 z , 它与 r_s 的距离最多为 $1/t$. 若存在这样的点 z , 取其中的一个点并将它增加到 X_t 中, 然后检查点 r_{s+1} ; 否则, 直接检查点 r_{s+1} .

由 X_t 的构造, 可断言点 $x \in S$ 与 X_t 中的某确定点之间的距离最多为 $2/t$. 显然成立. 的确, 由于有理向量在 \mathbb{R}^n 中是稠密的, 从而存在 s 使得 r_s 与 x 之间的距离 $\leq \frac{1}{t}$. 因此, 当处理 r_s 时, 肯定会向 X_t 中增加点 z , 且该点与 r_s 之间的距离 $\leq 1/t$, 这样, z 与 x 之间的距离 $\leq 2/t$. 根据构造方法, 可数集 $X_t \subset S$ 的可数个之并 $\bigcup_{t=1}^{\infty} X_t$ 是 S 中的可数集, 并且由上述断言知该集合在 S 中是稠密的. □

请注意在如下的分离定理中, S 与 T 的凸性很关键!

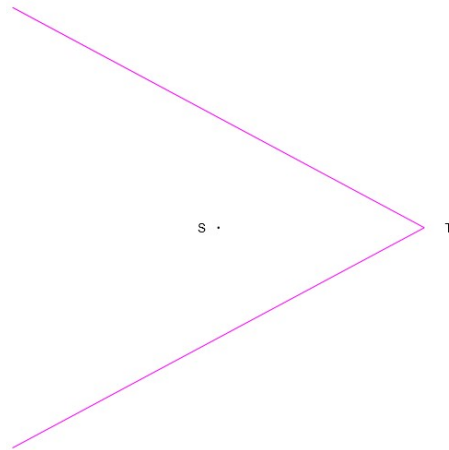


图 3.4: 这里 S 是单点集, T 是折线形成的集合. 这里 S 的相对内部是自己, T 的相对内部是 T 去掉尖点, 所以 S 与 T 的相对内部没有交集. 易见 S 与 T 是不可分离的.

定理3.18 (分离定理). 非空凸集 S 和 T 能被分离当且仅当它们的相对内部不相交.

证明. 用反证法证明必要性: 假设 $f^T x$ 分离 S 和 T . 因此

$$\sup_{x \in S} f^T x \leq \inf_{y \in T} f^T y.$$

假设与要证明的结论相反, 即 $\exists a \in \text{rint}(S) \cap \text{rint}(T)$. 由于 $a \in T$, 得到 $f^T a \geq \sup_{x \in S} f^T x$, 即

$$f^T a = \max_{x \in S} f^T x.$$

根据引理3.16可得, 对于所有 $x \in S$, $f^T x = f^T a$. 由于 $a \in S$, 得到

$$f^T a \leq \inf_{y \in T} f^T y,$$

即, $f^T a = \min_{y \in T} f^T y$. 根据引理3.16可得, 对于所有 $y \in T$, $f^T y = f^T a$. 这样, 对 $z \in S \cup T$, 有 $f^T z \equiv f^T a$. 因此 f 不能分离 S 与 T . 这与假设矛盾.

充分性. 假设 S, T 为非空凸集且满足 $\text{rint}(S) \cap \text{rint}(T) = \emptyset$, 分四步来证明 S, T 是可分离的.

Step 1: 分离点与有限点集的凸包. 设

$$S = \text{conv}(\{b_1, \dots, b_m\}), T = \{b\}, b \notin S.$$

证明 S 与 T 是可分离的. 为此, 设

$$\beta_i = \begin{bmatrix} b_i \\ 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} b \\ 1 \end{bmatrix}.$$

观察到 β 不是 β_1, \dots, β_m 的锥组合, 否则存在 $\lambda \geq 0$ 满足

$$\begin{bmatrix} b \\ 1 \end{bmatrix} = \sum_{i=1}^m \lambda_i \begin{bmatrix} b_i \\ 1 \end{bmatrix}, \quad (3.17)$$

即 $b = \sum_i \lambda_i b_i$, $\sum_i \lambda_i = 1$, $\lambda_i \geq 0$. 这与 $b \notin S$ 矛盾! 由于 β 不是 β_i 的锥组合, 根据齐

次Farkas 引理(定理3.8)知, 存在 $h = \begin{bmatrix} f \\ -a \end{bmatrix}$ 使得

$$f^T b - a \equiv h^T \beta > 0 \geq h^T \beta_i \equiv f^T b_i - a, \quad i = 1, \dots, m$$

即,

$$f^T b > \max_{i=1, \dots, m} f^T b_i = \max_{x \in S = \text{conv}\{b_1, \dots, b_m\}} f^T x.$$

请注意，这里运用了如下的明显事实：

$$\begin{aligned}
\max_{\mathbf{x} \in \text{conv}(\{\mathbf{b}_1, \dots, \mathbf{b}_m\})} \mathbf{f}^T \mathbf{x} &\equiv \max_{\lambda \geq 0, \sum_i \lambda_i = 1} \mathbf{f}^T \left[\sum_i \lambda_i \mathbf{b}_i \right] \\
&= \max_{\lambda \geq 0, \sum_i \lambda_i = 1} \sum_i \lambda_i \left[\mathbf{f}^T \mathbf{b}_i \right] \\
&= \max_i \mathbf{f}^T \mathbf{b}_i.
\end{aligned}$$

Step 2: 分离点与(不包含该点的)凸集. 设 S 是非空凸集, $\mathbf{b} \notin S$, $T = \{\mathbf{b}\}$. 证明 S 与 T 是可分离的.

将 S 与 T 均平移 $-\mathbf{b}$ (显然不影响集合的分离性), 可假设

$$T = \{\mathbf{0}\} \not\subset S.$$

如有必要, 把 \mathbb{R}^n 替换为 $\text{lin}(S)$, 从而可进一步假设 $\mathbb{R}^n = \text{lin}(S)$.

由引理3.17, 设 $\{\mathbf{x}_i \in S\}$ 是 S 中的稠密序列. 由于 S 是凸集且不包含 $\mathbf{0}$, 有

$$\mathbf{0} \notin \text{conv}(\{\mathbf{x}_1, \dots, \mathbf{x}_i\}) \quad \forall i,$$

因此由Step 1知

$$\exists \mathbf{f}_i : \mathbf{0} = \mathbf{f}_i^T \mathbf{0} > \max_{1 \leq j \leq i} \mathbf{f}_i^T \mathbf{x}_j. \quad (3.18)$$

经过伸缩, 可假设 $\|\mathbf{f}_i\|_2 = 1$.

单位向量序列 $\{\mathbf{f}_i\}$ 拥有收敛子序列 $\{\mathbf{f}_{i_s}\}_{s=1}^\infty$, 且该子序列的极限 \mathbf{f} 自然是单位向量. 根据(3.18)可知, 对于每个已知的 j 和所有足够大的 s , 有 $\mathbf{f}_{i_s}^T \mathbf{x}_j < 0$, 因此

$$\mathbf{f}^T \mathbf{x}_j \leq 0 \quad \forall j. \quad (3.19)$$

由于 $\{\mathbf{x}_j\}$ 在 S 中是稠密的, (3.19)意味着对于所有 $\mathbf{x} \in S$ 有 $\mathbf{f}^T \mathbf{x} \leq 0$, 因此

$$\sup_{\mathbf{x} \in S} \mathbf{f}^T \mathbf{x} \leq 0 = \mathbf{f}^T \mathbf{0}.$$

针对 $\text{lin}(S) = \mathbb{R}^n$ 和 $T = \{\mathbf{0}\}$, 找到单位向量 \mathbf{f} 使得

$$\sup_{\mathbf{x} \in S} \mathbf{f}^T \mathbf{x} \leq 0 = \mathbf{f}^T \mathbf{0}. \quad (3.20)$$

这样, 由(3.20)知, 证明 \mathbf{f} 分离 S 和 $T = \{\mathbf{0}\}$ 转化成验证

$$\inf_{\mathbf{x} \in S} \mathbf{f}^T \mathbf{x} < \mathbf{f}^T \mathbf{0} = 0.$$

假设该事实不成立, 那么结合(3.20)表明对于所有 $\mathbf{x} \in S$, 都有 $\mathbf{f}^T \mathbf{x} = 0$. 这与 $\text{lin}(S) = \mathbb{R}^n$ 和 \mathbf{f} 非零矛盾. 所以, 所以这是不可能的.

Step 3: 分离两个不相交的非空凸集. 设 S, T 是非空凸集, 并且它们不相交; 下面证明 S 和 T 是可分离的. 为此设

$$\hat{S} = S - T, \hat{T} = \{0\}.$$

显然集合 \hat{S} 是凸集且不包含 0 (因为 $S \cap T = \emptyset$). 根据**Step 2**可知, \hat{S} 与 $\{0\} = \hat{T}$ 是可分离的, 即存在 f 使得

$$\overbrace{\sup_{x \in S} f^T x - \inf_{y \in T} f^T y}^{\sup_{x \in S, y \in T} [f^T x - f^T y]} \leq 0 = \inf_{z \in \{0\}} f^T z, \quad \overbrace{\inf_{x \in S} f^T x - \sup_{y \in T} f^T y}^{\inf_{x \in S, y \in T} [f^T x - f^T y]} < 0 = \sup_{z \in \{0\}} f^T z.$$

因此

$$\sup_{x \in S} f^T x \leq \inf_{y \in T} f^T y, \quad \inf_{x \in S} f^T x < \sup_{y \in T} f^T y.$$

Step 4: 结束分离定理的证明. 最后, 设 S 和 T 是非空凸集, 且它们的相对内部不相交. 下面证明 S 和 T 是可分离的.

由命题2.9和2.4节最后一段的讨论, 知道集合

$$S' = \text{rint}(S), T' = \text{rint}(T)$$

是非空凸集; 这里考虑的是这些集合不相交的情况. 由**Step 3**知道, S' 与 T' 是可分离的, 即存在 f 使得

$$\sup_{x \in S'} f^T x \leq \inf_{y \in T'} f^T y, \quad \inf_{x \in S'} f^T x < \sup_{y \in T'} f^T y. \quad (3.21)$$

再次由2.4节最后一段的讨论, 知道 S' 在 S 中是稠密的, T' 在 T 中是稠密的, 在(3.21)式中将 S' 替换为 S , T' 替换为 T 时, \inf 和 \sup 运算保持不变. 因此, f 分离 S 和 T . \square

定义3.6. 称线性形 $f^T x$ 严格分离集合 S, T , 如果

$$\sup_{x \in S} f^T x < \inf_{y \in T} f^T y.$$

分离定理的另一种证明方法可从点 $T = \{a\}$ 与闭凸集 S 并且 $a \notin S$ 的分离开始着手, 且基于如下命题.

命题3.19. 设 S 是非空闭凸集, 并设 $a \notin S$. 那么在 S 中存在唯一距离 a 最近的点

$$\text{Proj}_S(a) = \operatorname{argmin}_{x \in S} \|a - x\|_2,$$

并且向量 $\mathbf{e} = \mathbf{a} - \text{Proj}_S(\mathbf{a})$ 分离 \mathbf{a} 与 S , 即

$$\max_{\mathbf{x} \in S} \mathbf{e}^T \mathbf{x} = \mathbf{e}^T \text{Proj}_S(\mathbf{a}) = \mathbf{e}^T \mathbf{a} - \|\mathbf{e}\|_2^2 < \mathbf{e}^T \mathbf{a}.$$

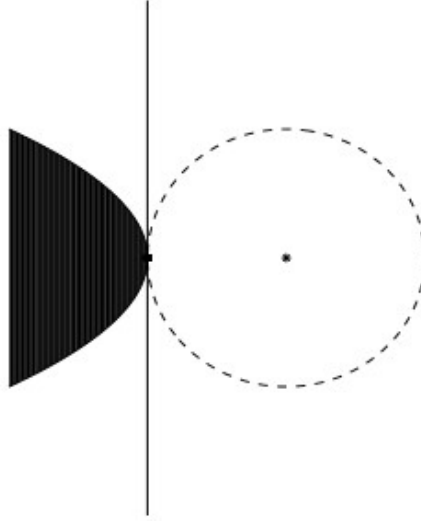


图 3.5: 投影的几何直观

证明. 先证明 S 中距 \mathbf{a} 最近的点确实存在. 事实上, 设 $\mathbf{x}_i \in S$ 是满足

$$\|\mathbf{a} - \mathbf{x}_i\|_2 \rightarrow \inf_{\mathbf{x} \in S} \|\mathbf{a} - \mathbf{x}\|_2, \quad i \rightarrow \infty$$

的序列. 显然序列 $\{\mathbf{x}_i\}$ 是有界的; 从而有收敛子列. 不妨设当 $i \rightarrow \infty$ 时, $\mathbf{x}_i \rightarrow \bar{\mathbf{x}}$. 由于 S 是闭集, 从而 $\bar{\mathbf{x}} \in S$, 并且

$$\|\mathbf{a} - \bar{\mathbf{x}}\|_2 = \lim_{i \rightarrow \infty} \|\mathbf{a} - \mathbf{x}_i\|_2 = \inf_{\mathbf{x} \in S} \|\mathbf{a} - \mathbf{x}\|_2.$$

接下来证明 S 中距离 \mathbf{a} 最近的点是唯一的. 事实上, 令 \mathbf{x}, \mathbf{y} 为 S 中距离 \mathbf{a} 最近的两个点, 因此

$$\|\mathbf{a} - \mathbf{x}\|_2 = \|\mathbf{a} - \mathbf{y}\|_2 = d.$$

由于 S 为凸集, 点 $\mathbf{z} = \frac{1}{2}(\mathbf{x} + \mathbf{y})$ 属于 S ; 再由 d 的最优性, 有 $\|\mathbf{a} - \mathbf{z}\|_2 \geq d$. 现在有

$$\overbrace{\|[\mathbf{a} - \mathbf{x}] + [\mathbf{a} - \mathbf{y}]\|_2^2}^{= \|2(\mathbf{a} - \mathbf{z})\|_2^2 \geq 4d^2} + \overbrace{\|[\mathbf{a} - \mathbf{x}] - [\mathbf{a} - \mathbf{y}]\|_2^2}^{= \|(\mathbf{x} - \mathbf{y})\|_2^2} = \overbrace{2\|\mathbf{a} - \mathbf{x}\|_2^2 + 2\|\mathbf{a} - \mathbf{y}\|_2^2}^{4d^2}.$$

因此 $\|\mathbf{x} - \mathbf{y}\|_2 = 0$.

综上, S 距 \mathbf{a} 最近的点存在且唯一. 由 $\mathbf{e} = \mathbf{a} - \text{Proj}_S(\mathbf{a})$, 对任何 $\mathbf{x} \in S$,

设 $\mathbf{x} \in S$, $\mathbf{f} = \mathbf{x} - \text{Proj}_S(\mathbf{a})$, 有

$$\begin{aligned}\phi(t) &\equiv \|\mathbf{e} - t\mathbf{f}\|_2^2 \\ &= \|\mathbf{a} - [\text{Proj}_S(\mathbf{a}) + t(\mathbf{x} - \text{Proj}_S(\mathbf{a}))]\|_2^2 \\ &\geq \|\mathbf{a} - \text{Proj}_S(\mathbf{a})\|_2^2 \\ &= \phi(0), \quad 0 \leq t \leq 1.\end{aligned}$$

所以 $0 \leq \phi'(0) = -2\mathbf{e}^T(\mathbf{x} - \text{Proj}_S(\mathbf{a}))$. 由此得

$$\forall \mathbf{x} \in S : \mathbf{e}^T \mathbf{x} \leq \mathbf{e}^T \text{Proj}_S(\mathbf{a}) = \mathbf{e}^T \mathbf{a} - \|\mathbf{e}\|_2^2.$$

□

定理3.20. 设 S 和 T 是非空凸集. 那么 S 和 T 是严格可分的当且仅当它们之间的距离大于 0, 即

$$\text{dist}(S, T) = \inf_{\mathbf{x} \in S, \mathbf{y} \in T} \|\mathbf{x} - \mathbf{y}\|_2 > 0.$$

证明. 必要性. 设 f 严格分离 S 和 T ; 用反证法证明 S 和 T 之间的距离大于 0. 假设结论不成立, 那么可找到序列 $\mathbf{x}_i \in S$ 和 $\mathbf{y}_i \in T$ 满足: 当 $i \rightarrow \infty$ 时 $\|\mathbf{x}_i - \mathbf{y}_i\|_2 \rightarrow 0$, 从而当 $i \rightarrow \infty$ 时 $f^T(\mathbf{y}_i - \mathbf{x}_i) \rightarrow 0$. 由此可构造数轴上距离为 0 的集合

$$\hat{S} = \{a = \mathbf{f}^T \mathbf{x} : \mathbf{x} \in S\}, \quad \hat{T} = \{b = \mathbf{f}^T \mathbf{y} : \mathbf{y} \in T\}.$$

这与 $\sup_{a \in \hat{S}} a < \inf_{b \in \hat{T}} b$ 矛盾.

充分性. 设 S 和 T 是非空凸集, 它们之间的距离是 $2\delta > 0$, 即

$$2\delta = \inf_{\mathbf{x} \in S, \mathbf{y} \in T} \|\mathbf{x} - \mathbf{y}\|_2 > 0.$$

设 $S^+ = S + \{\mathbf{z} : \|\mathbf{z}\|_2 \leq \delta\}$, 那么集合 S^+ 和 T 都是凸集且不相交, 因此二者是可分离的, 即存在 $f \neq 0$ 满足

$$\sup_{\mathbf{x}_+ \in S^+} \mathbf{f}^T \mathbf{x}_+ \leq \inf_{\mathbf{y} \in T} \mathbf{f}^T \mathbf{y}.$$

由于

$$\sup_{\mathbf{x}_+ \in S^+} \mathbf{f}^T \mathbf{x}_+ = \sup_{\mathbf{x} \in S, \|\mathbf{z}\|_2 \leq \delta} [\mathbf{f}^T \mathbf{x} + \mathbf{f}^T \mathbf{z}],$$

从而得到

$$\sup_{\mathbf{x} \in S} \mathbf{f}^T \mathbf{x} < \inf_{\mathbf{y} \in T} \mathbf{f}^T \mathbf{y}.$$

□

例3.8. 下面的问题中, S 是非空凸集, $T = \{a\}$.

命题	对/错
若 T 与 S 可被分离, 那么 $a \notin S$	
若 $a \notin S$, 那么 T 与 S 可被分离	
若 T 与 S 可被严格分离, 那么 $a \notin S$	
若 $a \notin S$, 那么 T 与 S 可被严格分离	
若 S 是闭集且 $a \notin S$, 那么 T 与 S 可被严格分离	

3.7 支撑超平面和极点

设 Q 是 \mathbb{R}^n 中的闭凸集, \bar{x} 是 Q 相对边界上的一点. 称超平面

$$H = \{x : f^T x = a\} \quad [a \neq 0]$$

是 Q 在点 \bar{x} 处的支撑超平面, 如果该超平面分离 Q 与 $\{\bar{x}\}$, 即

$$\sup_{x \in Q} f^T x \leq f^T \bar{x}, \quad \inf_{x \in Q} f^T x < f^T \bar{x}.$$

等价地, 超平面 $H = \{x : f^T x = a\}$ 在 \bar{x} 处支撑 Q 当且仅当线性形 $f^T x$ 在 \bar{x} 处取到其在 Q 中的最大值, 该最大值等于 a , 且该线性形在 Q 上不是常数.

命题3.21. 设 Q 是 \mathbb{R}^n 中的闭凸集, \bar{x} 是 Q 相对边界上的点. 那么存在至少一个超平面 H , 其在 \bar{x} 处支撑 Q . 进一步, 对于每个这样的超平面 H , 集合 $Q \cap H$ 的维数小于 Q 的维数.

证明. 因为由 $\bar{x} \notin \text{rint } Q$ 知

$$\{\bar{x}\} \cap \text{rint } Q = \emptyset.$$

从而分离定理(定理3.18)保证了支撑超平面的存在性. 进一步, 由 $Q \not\subseteq H$ 知 $\text{aff } Q \not\subseteq H$, 从而 $\text{aff}(H \cap Q) \subsetneq \text{aff } Q$. 如果两个互不相同的仿射子空间中的一个可以嵌入到另一个中, 那么已经嵌入子空间的维数严格小于即将嵌入子空间的维数. \square

定义3.7. 设 Q 是 \mathbb{R}^n 中的凸集, \bar{x} 是 Q 中的一点. 若该点不是 Q 中异于 \bar{x} 的两点的正加权凸组合, 那么称该点是 Q 的极点, 即 $\bar{x} \in \text{ext } Q$ 当且仅当 $\bar{x} \in Q$ 且如果

$$u, v \in Q, \lambda \in (0, 1) \text{ s.t. } \bar{x} = \lambda u + (1 - \lambda)v,$$

那么 $u = v = \bar{x}$.

等价地, 点 $\bar{x} \in Q$ 是极点当且仅当它不是 Q 中非平凡线段的中点, 即

$$\bar{x} \pm \mathbf{h} \in Q \Rightarrow \mathbf{h} = 0.$$

等价地, 点 $\bar{x} \in Q$ 是极点当且仅当集合 $Q \setminus \{\bar{x}\}$ 是凸集.

例3.9. (i) $[x, y]$ 的极点是 x 和 y .

(ii) $\triangle ABC$ 的极点是顶点 A, B 和 C .

(iii) 球 $\{x : \|x\|_2 \leq 1\}$ 的极点是 x , 其中 $\|x\|_2 = 1$.

下面证明著名的 Krein-Milman 定理. 为此需要如下两个引理.

引理3.22. 设 S 是闭凸集, 超平面 $H = \{x : f^T x = a\}$ 在某已知点支撑 S . 那么

$$\text{ext}(H \cap S) \subset \text{ext}S.$$

证明. 设 $\bar{x} \in \text{ext}(H \cap S)$. 下面用反证法证明 $\bar{x} \in \text{ext}S$. 假设 \bar{x} 是非平凡线段 $[u, v] \subset S$ 的中点. 那么

$$f^T \bar{x} = a = \max_{x \in S} f^T x,$$

因此

$$f^T \bar{x} = a = \max_{x \in [u, v]} f^T x.$$

由引理3.16, 线性形在区间段中点取得最大值当且仅当线性形在该区间上是常数; 这样,

$$a = f^T \bar{x} = f^T u = f^T v,$$

即, $[u, v] \subset (H \cap S)$. 这与 \bar{x} 是 $H \cap S$ 的极点相矛盾! □

引理3.23. 设 S 是闭凸集, 其满足对于某已知 \bar{x} , 有 $\{\bar{x} + t\mathbf{h} : t \geq 0\} \subset S$. 那么

$$\{x + t\mathbf{h} : t \geq 0\} \subset S, \forall x \in S.$$

证明. 对于每个 $s > 0$ 和 $x \in S$, 有

$$x + s\mathbf{h} = \lim_{i \rightarrow \infty} \underbrace{[(1 - s/i)x + (s/i)[\bar{x} + i\mathbf{h}]]}_{\in S}.$$

□

请注意, 称由对于某(并且之后, 对所有) $x \in S$ 有 $\{x + t\mathbf{h} : t \geq 0\} \subset S$ 成立的所有方向 $\mathbf{h} \in \mathbb{R}^n$ 组成的集合是闭凸集 S 的回收锥(recessive cone), 并记为 $\text{rec}S$. 回收锥 $\text{rec}S$ 是锥, 并且

$$S + \text{rec}S = S.$$

□

推论3.24. 如果闭凸集 Q 包含直线 ℓ , 那么过 Q 中各点并且与 ℓ 平行的直线也属于 Q . 特别地, Q 没有极点.

定理3.25 (Krein-Milman). 设 Q 是 \mathbb{R}^n 中的非空闭凸集, 那么

- (i) Q 有极点当且仅当它不包含直线;
- (ii) 如果 Q 是有界的, 那么它是其极点的凸包:

$$Q = \text{conv}(\text{ext}Q).$$

因此 Q 的每个点都是 Q 极点的凸组合.

请注意, 如果 $Q = \text{conv}A$, 那么 $\text{ext}Q \subset A$. 因此, 有界闭凸集 Q 的极点给出了 Q 的最小表示.

证明. 首先证明: 如果闭凸集 Q 不包含直线, 那么 $\text{ext}Q \neq \emptyset$.

设 Q 是不包含直线的非空闭凸集. 下面应用纯化(Purification)算法 找到 Q 的极点:

初始化: 置 $S_0 = Q$ 并选择 $\mathbf{x}_0 \in Q$.

步骤 t : 给定不包含直线的非空闭凸集 S_t , 其满足

$$\text{ext}S_t \subset \text{ext}Q, \mathbf{x}_t \in S_t,$$

- (1) 检查 S_t 是否是单个元素集合 $\{\mathbf{x}_t\}$. 若是, 终止, 且

$$\mathbf{x}_t \in \text{ext}S_t \subset \text{ext}Q.$$

(2) 如果 S_t 不是单个元素集合, 在其相对边界上找到点 \mathbf{x}_{t+1} , 由命题3.21知, 可确定在该点支撑 S_t 的超平面 H_t . 为此, 取平行于 $\text{aff}S_t$ 的方向 $\mathbf{h} \neq 0$. 由于 S_t 不包含直线, 当从 \mathbf{x}_t 沿方向 \mathbf{h} 或者方向 $-\mathbf{h}$ 移动时, 最终会离开 S_t , 从而会穿过 S_t 的相对边界, 得到的交点就是希望找的 \mathbf{x}_{t+1} .

- (3) 置 $S_{t+1} = S_t \cap H_t$, 用 $t+1$ 替换 t , 并循环到(1).

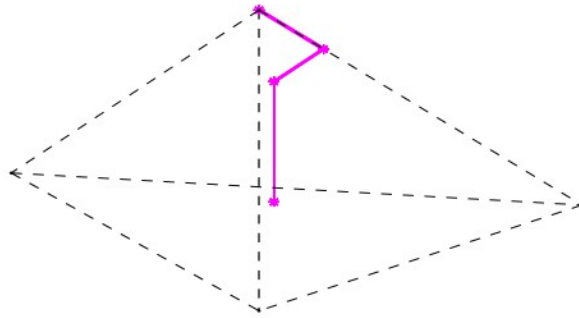


图 3.6: 非空闭凸集 Q 是 \mathbb{R}^3 中的四面体, \mathbf{x}_0 位于底面的中心, \mathbf{x}_1 在侧面上, \mathbf{x}_2 位于棱上, \mathbf{x}_3 是找到的极点.

解释：根据引理3.22知

$$\text{ext}S_{t+1} \subset \text{ext}S_t$$

因此

$$\text{ext}S_t \subset \text{ext}Q \quad \forall t.$$

此外 $\dim S_{t+1} < \dim S_t$, 因此纯化算法会有限步终止.

请注意, 给定线性形 $g^T x$, 它在 Q 上有界, 那么在纯化算法中很容易保证 $g^T x_{t+1} \geq g^T x_t$, 这样, 如果 Q 是 \mathbb{R}^n 上不包含直线的非空闭集, $f^T x$ 是在 Q 上有界的线性形, 那么对于每个点 $x_0 \in Q$, 存在(可通过纯化算法找出)点 $\bar{x} \in \text{ext}Q$ 使得 $g^T \bar{x} \geq g^T x_0$. 特别地, 若 $g^T x$ 在 Q 上能取到最大值, 那么在 Q 的极点中可以找到最大值点.

其次, 由重要引理3.22的推论知: 如果闭凸集 Q 包含直线, 那么它没有极点.

最后, 如果非空闭凸集 Q 有界, 那么 $Q = \text{conv}(\text{ext}Q)$. 这里的包含关系 $\text{conv}(\text{ext}Q) \subset Q$ 是显然的. 下面证明反包含关系, 即证明 Q 的每个点都是 Q 极点的凸组合. 这里对 $k = \dim Q$ 进行归纳. 当 $k = 0$ (Q 由单个元素组成) 时, 结论显然成立.

Step $k \rightarrow k+1$: 给定 $(k+1)$ -维有界闭凸集 Q 以及 $x \in Q$, 与纯化算法一样, 可将 x 表示成 Q 的相对边界上两点 x_+ , x_- 的凸组合. 设 H_+ 是在 x_+ 支撑 Q 的超平面, 并设 $Q_+ = H_+ \cap Q$. 易见 Q_+ 是闭凸集, 且满足

$$\dim Q_+ < \dim Q, \quad \text{ext}Q_+ \subset \text{ext}Q, \quad x_+ \in Q_+.$$

由归纳假设, 得

$$x_+ \in \text{conv}(\text{ext}Q_+) \subset \text{conv}(\text{ext}Q).$$

同理, $x_- \in \text{conv}(\text{ext}Q)$. 由于 $x \in [x_-, x_+]$, 得到 $x \in \text{conv}(\text{ext}Q)$. □

3.8 多面集的结构及其在线性规划中的应用

定义3.8. \mathbb{R}^n 中的多面集 Q 是 \mathbb{R}^n 的子集, 是有限个非严格线性不等式组的解集:

$$Q \text{ 是多面集当且仅当 } Q = \{x : Ax \geq b\},$$

其中 $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

任何多面集都是闭凸集. 后面所讨论的多面集都假设是非空的. 现在的问题: 多面集 $Q = \{x : Ax \geq b\}$ 什么情况下包含直线? 如果包含直线, 如何刻画这些直线?

答案: Q 包含直线当且仅当 A 有非平凡零空间:

$$\text{null}(A) \equiv \{h : Ah = 0\} \neq \{0\}.$$

事实上, 对于 $\mathbf{h} \neq 0$, 直线 $\ell = \{\mathbf{x} = \bar{\mathbf{x}} + t\mathbf{h} : t \in \mathbb{R}\}$ 属于 Q 当且仅当

$$\forall t : \mathbf{A}(\bar{\mathbf{x}} + t\mathbf{h}) \geq \mathbf{b} \Leftrightarrow \forall t : t\mathbf{A}\mathbf{h} \geq \mathbf{b} - \mathbf{A}\bar{\mathbf{x}} \Leftrightarrow \mathbf{A}\mathbf{h} = \mathbf{0} \ \& \ \bar{\mathbf{x}} \in Q.$$

重要事实: 多面集 $Q = \{\mathbf{x} : \mathbf{A}\mathbf{x} \geq \mathbf{b}\}$ 总是可以表示成

$$Q = Q_* + L,$$

其中 Q_* 是不包含直线的多面集, L 是线性子空间. 在该表述中, L 由 Q 唯一定义, 且与 $\text{null}A$ 是一致的; 可将 Q_* 选为(例如)

$$Q_* = Q \cap L^\perp.$$

定理3.26 (不包含直线的多面集的结构). 设 $Q = \{\mathbf{x} : \mathbf{A}\mathbf{x} \geq \mathbf{b}\} \neq \emptyset$ 是不包含直线的多面集(或者 $\text{null}A = \{0\}$ 也是同样的意思). 那么 Q 的极点集合 $\text{ext}Q$ 非空且有限, 对于正确选取的向量 $\mathbf{r}_1, \dots, \mathbf{r}_S$,

$$Q = \text{conv}(\text{ext}Q) + \text{cone}\{\mathbf{r}_1, \dots, \mathbf{r}_S\} \quad (3.22)$$

请注意, $\text{cone}\{\mathbf{r}_1, \dots, \mathbf{r}_S\}$ 恰好是 Q 的回收锥:

$$\text{cone}\{\mathbf{r}_1, \dots, \mathbf{r}_S\} = \{\mathbf{r} : \mathbf{x} + t\mathbf{r} \in Q \ \forall (\mathbf{x} \in Q, t \geq 0)\} = \{\mathbf{r} : \mathbf{A}\mathbf{r} \geq \mathbf{0}\}.$$

该锥是平凡锥 $\{0\}$ 当且仅当 Q 是有界多面集(称为多胞形(polytope)). 综合上述各定理, 得到如下结论: (非空)多面集 Q 总可以表述为

$$Q = \left\{ \mathbf{x} = \sum_{i=1}^I \lambda_i \mathbf{v}_i + \sum_{j=1}^J \mu_j \mathbf{w}_j : \lambda \geq 0, \mu \geq 0, \sum_i \lambda_i = 1 \right\} \quad (3.23)$$

其中 I, J 是正整数, $\mathbf{v}_1, \dots, \mathbf{v}_I, \mathbf{w}_1, \dots, \mathbf{w}_J$ 是恰当选取的点和方向. 反之亦然, 形如(3.23)的集合 Q 是多面集. 请注意, 当(3.23)的集合具有“平凡 \mathbf{w} -部分”的属性时, 即 $\mathbf{w}_1 = \dots = \mathbf{w}_J = \mathbf{0}$, 那么 Q 正好是多胞形(有界多面集).

练习 1 若两个多面集的交集非空, 它是否是多面集?

练习 2 多面集 Q 的仿射映射下的像 $\{\mathbf{y} = \mathbf{P}\mathbf{x} + \mathbf{p} : \mathbf{x} \in Q\}$ 是否是多面集?

下面将这些理论应用到线性规划中. 为此, 考虑可行线性规划

$$\min_{\mathbf{x}} \mathbf{c}^T \mathbf{x} \text{ subject to } \mathbf{x} \in Q = \{\mathbf{x} : \mathbf{A}\mathbf{x} \geq \mathbf{b}\}. \quad (\text{LP})$$

首先观察到假设 $\text{null}A = \{0\}$ 不影响要解决的问题. 事实上, 有

$$Q = Q_* + \text{null}A,$$

其中 Q_* 是不包含直线的多面集. 如果 c 与 $\text{null}A$ 不正交, 那么(LP)显然无界. 如果 c 与 $\text{null}A$ 正交, 那么(LP)等价于

$$\min_x c^T x \text{ subject to } x \in Q_*,$$

其中 $Q_* = \{x : \tilde{A}x \geq \tilde{b}\}$, 并且矩阵 \tilde{A} 的零空间是平凡的. 这表明已经证明了命题: 假设 $\text{null}A = \{0\}$, 设(LP)有界(因此可解). 由于 Q 是闭凸集且不包含直线, 因此目标函数在 Q 上的最小值点集合(非空!)里有 Q 的极点.

命题3.27. 假设(LP)可行有界(因此可解), 且 $\text{null}A = \{0\}$. 那么至少存在 Q 的一个极点是(LP)的最优解.

假设 A 是 $m \times n$ 矩阵, 且 $\text{null}A = \{0\}$. 现在的问题是**如何刻画集合**

$$Q = \{x : Ax \geq b\} \neq \emptyset$$

的极点.

命题3.28. \bar{x} 是 Q 的极点当且仅当 $A\bar{x} \geq b$, 且对于约束 $Ax \geq b$, 那些 \bar{x} 处的积极约束(也就是等式约束)中有 n 个是线性无关的.

证明. 必要性: 如果 \bar{x} 是 Q 的极点, 那么对于约束 $Ax \geq b$, \bar{x} 处的那些积极约束中有 n 个是线性无关的.

不失一般性, 假设 \bar{x} 处的积极约束是下述前 k 个约束

$$a_i^T x \geq b_i \quad i = 1, \dots, k.$$

欲证明在 n -维向量 a_1, \dots, a_k 中, 有 n 个是线性无关的. 反证法: 假设存在非零向量 h , 使得 $a_i^T h = 0, i = 1, \dots, k$, 也就是说, 对于所有 $\epsilon > 0$,

$$a_i^T [\bar{x} \pm \epsilon h] = a_i^T \bar{x} = b_i, \quad i = 1, \dots, k.$$

由于剩下的约束 $a_i^T x \geq b_i, i > k$, 在 \bar{x} 处得到严格满足, 因此断定: 对于所有足够小的 $\epsilon > 0$,

$$a_i^T [\bar{x} \pm \epsilon h] \geq b_i, \quad i = k+1, \dots, m.$$

综上可以断定: 对于所有足够小的 $\epsilon > 0$, $\bar{x} \pm \epsilon h \in Q = \{x : Ax \geq b\}$. 由于 $h \neq 0$ 以及 \bar{x} 是 Q 的极点, 因此得到矛盾结果.

充分性: 如果 $\bar{x} \in Q$ 使得约束 $a_i^T x \geq b_i$ 中有 n 个系数向量线性无关的等式成立, 那么 $\bar{x} \in \text{ext}Q$.

不失一般性, 假设如下前 n 个约束

$$a_i^T x \geq b_i, \quad i = 1, \dots, n$$

的系数向量线性无关, 且满足 $\mathbf{a}_i^T \bar{\mathbf{x}} = b_i, \quad i = 1, \dots, n$. 欲证明: 如果 \mathbf{h} 使得 $\bar{\mathbf{x}} \pm \mathbf{h} \in Q$, 那么 $\mathbf{h} = \mathbf{0}$. 事实上, 由 $\bar{\mathbf{x}} \pm \mathbf{h} \in Q$ 可以推出

$$\mathbf{a}_i^T [\bar{\mathbf{x}} \pm \mathbf{h}] \geq b_i, \quad i = 1, \dots, n.$$

对于 $i \leq n$, 由于 $\mathbf{a}_i^T \bar{\mathbf{x}} = b_i$, 得到

$$\mathbf{a}_i^T \bar{\mathbf{x}} \pm \mathbf{a}_i^T \mathbf{h} = \mathbf{a}_i^T [\bar{\mathbf{x}} \pm \mathbf{h}] \geq \mathbf{a}_i^T \bar{\mathbf{x}}, \quad i = 1, \dots, n,$$

因此

$$\mathbf{a}_i^T \mathbf{h} = 0, \quad i = 1, \dots, n. \tag{3.24}$$

由于 n -维向量 $\mathbf{a}_1, \dots, \mathbf{a}_n$ 线性无关, 因此(3.24)意味着 $\mathbf{h} = \mathbf{0}$. □

4 凸函数

4.1 定义与例子

定义4.1. 设 f 是定义在非空子集 $\text{dom } f \subseteq \mathbb{R}^n$ 上的实值函数. 称 f 是凸的(convex), 如果 $\text{dom } f$ 是凸集, 并且对于所有 $\mathbf{x}, \mathbf{y} \in \text{dom } f, \theta \in [0, 1]$, 有

$$f((1-\theta)\mathbf{x} + \theta\mathbf{y}) \leq (1-\theta)f(\mathbf{x}) + \theta f(\mathbf{y}). \quad (4.1)$$

凸函数的等价定义: 设 f 是定义在非空子集 $\text{dom } f \subseteq \mathbb{R}^n$ 上的实值函数. 如果它的上镜图(epigraph)

$$\text{epi } f = \{(\mathbf{x}, t) \in \mathbb{R}^{n+1} : f(\mathbf{x}) \leq t\}$$

是 \mathbb{R}^{n+1} 中的凸集, 那么 f 是凸函数.

下面来看凸性定义实际表示的是什么? 考虑不等式(4.1), 其中 $\mathbf{x}, \mathbf{y} \in \text{dom } f, \theta \in [0, 1]$. 当 $\mathbf{x} = \mathbf{y}$ 或者 $\theta = 0/1$ 时, 不等式自动满足. 从而仅需讨论 \mathbf{x}, \mathbf{y} 互不相同和 $\theta \in (0, 1)$ 的情形, 这时点 $\mathbf{z} = (1-\theta)\mathbf{x} + \theta\mathbf{y}$ 是(相比较而言)区间 $[\mathbf{x}, \mathbf{y}]$ 的内点. 具体讨论如下:

我们观察到 $\mathbf{z} = (1-\theta)\mathbf{x} + \theta\mathbf{y} = \mathbf{x} + \theta(\mathbf{y} - \mathbf{x})$, 因此

$$\|\mathbf{y} - \mathbf{x}\| : \|\mathbf{y} - \mathbf{z}\| : \|\mathbf{z} - \mathbf{x}\| = 1 : (1-\theta) : \theta. \quad (4.2)$$

给(4.1)两边同时减去 $f(\mathbf{x})$, 并将由(4.2)得到的 $\theta = \frac{\|\mathbf{z}-\mathbf{x}\|}{\|\mathbf{y}-\mathbf{x}\|}$ 代入, 得

$$\frac{f(\mathbf{z}) - f(\mathbf{x})}{\|\mathbf{z} - \mathbf{x}\|} \leq \frac{f(\mathbf{y}) - f(\mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|}.$$

同理, 整理(4.1)得

$$(1-\theta)[f(\mathbf{y}) - f(\mathbf{x})] \leq f(\mathbf{y}) - f(\mathbf{z}),$$

并将由(4.2)得到的 $1-\theta = \frac{\|\mathbf{y}-\mathbf{z}\|}{\|\mathbf{y}-\mathbf{x}\|}$ 代入, 得

$$\frac{f(\mathbf{y}) - f(\mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} \leq \frac{f(\mathbf{y}) - f(\mathbf{z})}{\|\mathbf{y} - \mathbf{z}\|}.$$

命题4.1. f 是凸函数当且仅当对于任何互不相同的三点 $\mathbf{x}, \mathbf{y}, \mathbf{z}$ 满足: 当 $\mathbf{x}, \mathbf{y} \in \text{dom } f, \mathbf{z} \in (\mathbf{x}, \mathbf{y})$ 时, 我们有 $\mathbf{z} \in \text{dom } f$, 并且

$$\frac{f(\mathbf{z}) - f(\mathbf{x})}{\|\mathbf{z} - \mathbf{x}\|} \leq \frac{f(\mathbf{y}) - f(\mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} \leq \frac{f(\mathbf{y}) - f(\mathbf{z})}{\|\mathbf{y} - \mathbf{z}\|}. \quad (4.3)$$

请注意, (4.3)中三个不等式中的任何一个都蕴含着另外两个.

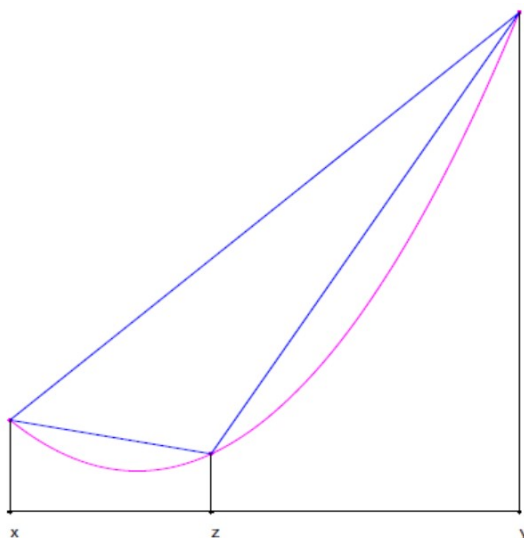


图 4.7: 凸函数的实际含意

命题4.2 (Jensen不等式). 设 f 是凸函数, $k \in \mathbb{Z}_+$ 给定. 如果

$$\mathbf{x}_i \in \text{dom} f, \theta_i \geq 0, i = 1, \dots, k, \sum_i \theta_i = 1,$$

那么

$$f\left(\sum_i \theta_i \mathbf{x}_i\right) \leq \sum_i \theta_i f(\mathbf{x}_i).$$

证明 点 $(\mathbf{x}_i, f(\mathbf{x}_i))$ 属于 $\text{epi} f$. 由于该集合是凸集, 因此

$$\left(\sum_i \theta_i \mathbf{x}_i, \sum_i \theta_i f(\mathbf{x}_i)\right) \in \text{epi} f.$$

再根据上镜图的定义可得

$$f\left(\sum_i \theta_i \mathbf{x}_i\right) \leq \sum_i \theta_i f(\mathbf{x}_i).$$

□

推广: 设 f 是凸函数, $\text{dom} f$ 是凸集, 并且 f 在 $\text{dom} f$ 上连续. 考虑 $\text{dom} f$ 上的随机变量 X , 那么

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

例4.1. (a) \mathbb{R} 上的凸函数: $x^2, x^4, x^6, \dots, \exp(x)$; \mathbb{R} 上的非凸函数: $x^3, \sin(x)$;

(b) \mathbb{R}_+ 上的凸函数: $x^p, p \geq 1; -x^p, 0 \leq p \leq 1; x \ln x$;

(c) \mathbb{R}^n 上的凸函数: 仿射函数 $f(\mathbf{x}) = \mathbf{f}^T \mathbf{x}$

(d) \mathbb{R}^n 上的范数 $\|\cdot\|$ 是凸函数:

$$\begin{aligned}\|(1-\theta)\mathbf{x} + \theta\mathbf{y}\| &\leq \|(1-\theta)\mathbf{x}\| + \|\theta\mathbf{y}\| \quad [\text{三角不等式}] \\ &= (1-\theta)\|\mathbf{x}\| + \theta\|\mathbf{y}\| \quad [\text{齐次性}]\end{aligned}$$

例4.2 (Jensen不等式的应用). 设 $p = \{p_i > 0\}_{i=1}^n$, $q = \{q_i > 0\}_{i=1}^n$ 是两个离散概率分布. 那么两分布之间的Kullback-Liebler距离

$$\sum_i p_i \ln \frac{p_i}{q_i} \geq 0.$$

事实上, 函数 $f(x) = -\ln x$ 是凸函数, 这里 $\text{dom} f = \{x \in \mathbb{R} : x > 0\}$. 置 $x_i = q_i/p_i$, $\theta_i = p_i$, 有

$$\begin{aligned}0 &= -\ln \left(\sum_i q_i \right) = f\left(\sum_i p_i x_i\right) \\ &\leq \sum_i p_i f(x_i) = \sum_i p_i (-\ln(q_i/p_i)) \\ &= \sum_i p_i \ln(p_i/q_i).\end{aligned}$$

有时约定凸函数 f 是定义在 \mathbb{R}^n 上的任何地方, 且取实数值和 $+\infty$ 是有益的. 在这种约定下, f 的(有效)定义域:

$$\text{dom} f = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \in \mathbb{R}\}, \mathbf{x} \notin \text{dom} f \Rightarrow f(\mathbf{x}) = +\infty, \quad (4.4)$$

并且凸性的定义变成

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \forall \theta \in [0, 1], \quad (4.5)$$

其中 $+\infty$ 与实数的运算法则是

$$\begin{aligned}+\infty &\leq +\infty, \\ a \in \mathbb{R} &\Rightarrow a + (+\infty) = (+\infty) + (+\infty) = +\infty, \\ 0 \cdot (+\infty) &= 0, \\ \theta > 0 &\Rightarrow \theta \cdot (+\infty) = +\infty.\end{aligned}$$

请注意未定义 $(+\infty) - (+\infty)$, $(-5) \cdot (+\infty)$ 等运算!

4.2 保凸运算

命题4.3. (a) [锥组合] 设 $f_i(\mathbf{x})$ 是 \mathbb{R}^n 上的凸函数, $\lambda_i \geq 0$, 那么函数 $\sum_i \lambda_i f_i(\mathbf{x})$ 是凸的.

(b) [自变量的仿射替换] 若 $f(\mathbf{x})$ 是 \mathbb{R}^n 上的凸函数, $\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{b}$ 是从 \mathbb{R}^m 到 \mathbb{R}^n 的仿射映射, 则函数 $g(\mathbf{y}) = f(\mathbf{A}\mathbf{y} + \mathbf{b})$ 在 \mathbb{R}^m 上是凸的.

(c) [上确界] 如果 $f_\alpha(\mathbf{x})$, $\alpha \in \mathcal{A}$, 是 \mathbb{R}^n 上的凸函数族, 那么函数 $\sup_{\alpha \in \mathcal{A}} f_\alpha(\mathbf{x})$ 是凸的.

(d) [叠加定理] 对 $i = 1, \dots, m$, 设 $f_i(\mathbf{x})$ 是 \mathbb{R}^n 上的凸函数, 函数 $F(y_1, \dots, y_m)$ 在 \mathbb{R}^m 上是凸并且单调递增, 那么函数

$$g(\mathbf{x}) = \begin{cases} F(f_1(\mathbf{x}), \dots, f_m(\mathbf{x})), & \mathbf{x} \in \text{dom} f_i, \forall i \\ +\infty, & \text{其它} \end{cases}$$

是凸的.

(e) [部分极小化] 设 $f(\mathbf{x}, \mathbf{y})$ 关于 $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n$ 是凸的, 并且

$$g(\mathbf{x}) = \inf_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}).$$

那么函数 $g(\mathbf{x})$ 在任何 g 不取 $-\infty$ 值的凸集 Q 上是凸的.

证明 这里仅给出命题(c)和(e)的证明. 因为

$$\text{epi} \left(\sup_{\alpha \in \mathcal{A}} f_\alpha(\cdot) \right) = \bigcap_{\alpha \in \mathcal{A}} \text{epi} \{f_\alpha(\cdot)\},$$

再由凸集的交集是凸集知(c)成立.

设 Q 是凸集, 且满足 g 在 Q 上不取 $-\infty$. 下面来检验凸性不等式

$$g((1-\theta)\mathbf{x}' + \theta\mathbf{x}'') \leq (1-\theta)g(\mathbf{x}') + \theta g(\mathbf{x}'') \quad [\theta \in [0, 1], \mathbf{x}', \mathbf{x}'' \in Q].$$

当 $\theta = 0$ 或 $\theta = 1$ 时无需检验, 因此设 $0 < \theta < 1$. 在该情况下, 当 $g(\mathbf{x}')$ 或 $g(\mathbf{x}'')$ 为 $+\infty$ 时也无需检验, 因此设 $g(\mathbf{x}') < +\infty$, $g(\mathbf{x}'') < +\infty$. 由于 $g(\mathbf{x}') < +\infty$, 对于任何 $\epsilon > 0$, 存在 \mathbf{y}' 使得 $f(\mathbf{x}', \mathbf{y}') \leq g(\mathbf{x}') + \epsilon$. 同理, 存在 \mathbf{y}'' 使得 $f(\mathbf{x}'', \mathbf{y}'') \leq g(\mathbf{x}'') + \epsilon$, 从而有

$$\begin{aligned} g((1-\theta)\mathbf{x}' + \theta\mathbf{x}'') &\leq f((1-\theta)\mathbf{x}' + \theta\mathbf{x}'', (1-\theta)\mathbf{y}' + \theta\mathbf{y}'') \\ &\leq (1-\theta)f(\mathbf{x}', \mathbf{y}') + \theta f(\mathbf{x}'', \mathbf{y}'') \\ &\leq (1-\theta)(g(\mathbf{x}') + \epsilon) + \theta(g(\mathbf{x}'') + \epsilon) \\ &= (1-\theta)g(\mathbf{x}') + \theta g(\mathbf{x}'') + \epsilon \end{aligned}$$

由于 $\epsilon > 0$ 是任意的, 从而得到

$$g((1-\theta)\mathbf{x}' + \theta\mathbf{x}'') \leq (1-\theta)g(\mathbf{x}') + \theta g(\mathbf{x}'').$$

□

4.3 如何检测凸性

以下命题表明凸性是一维性质.

命题4.4. (a) 集合 $X \subset \mathbb{R}^n$ 是凸集当且仅当由任何 (\mathbf{a}, \mathbf{h}) 确定的集合 $\{t : \mathbf{a} + t\mathbf{h} \in X\}$ 在数轴上是凸的.

(b) 函数 f 在 \mathbb{R}^n 上是凸的当且仅当由任一 (\mathbf{a}, \mathbf{h}) 确定的函数 $\phi(t) = f(\mathbf{a} + t\mathbf{h})$ 在数轴上是凸的.

下面讨论一元函数 ϕ 何时是凸的? 设 ϕ 在区间 (a, b) 上是凸的和有限的, 这时表述(4.3)即

$$\frac{\phi(z) - \phi(x)}{z - x} \leq \frac{\phi(y) - \phi(x)}{y - x} \leq \frac{\phi(y) - \phi(z)}{y - z}, \quad a < x < z < y < b,$$

它与 ϕ 凸完全相同. 假设 $\phi'(x), \phi'(y)$ 存在, 即 ϕ 在 (a, b) 上可导. 那么当 $z \rightarrow x + 0$ 和 $z \rightarrow y - 0$ 时取极限, 得到

$$\phi'(x) \leq \frac{\phi(y) - \phi(x)}{y - x} \leq \phi'(y),$$

也就是说, $\phi'(x)$ 在 (a, b) 上是单调非减的.

下面命题给出了单变量函数凸性的必要条件和充分条件.

命题4.5. (a) 函数 ϕ 的定义域应该是开区间 $\Delta = (a, b)$, 或者可能带有额外的一个或多个端点(倘若相应的端点是有限的).

(b) ϕ 在 (a, b) 上是连续的, 在(或许)除可数集合外是处处可微的, 并且导数是单调非减的.

(c) 在 (a, b) 的某个属于 $\text{dom}\phi$ 的端点上, 允许 ϕ 是“向上跳跃”, 但不允许向下跳跃.

(d) 单变量函数 ϕ 是凸函数的充分条件: $\text{dom}\phi$ 是凸集, ϕ 在 $\text{dom}\phi$ 上是连续的, 在 $\text{int}\text{dom}\phi$ 上二次可微, 并且 ϕ'' 是非负的.

的确, 应该证明: 在该条件下, 若 $x < z < y$ 在 $\text{dom}\phi$ 里, 那么

$$\frac{\phi(z) - \phi(x)}{z - x} \leq \frac{\phi(y) - \phi(z)}{y - z}.$$

根据拉格朗日定理, 存在某 $\xi \in (x, z)$ 使得左边比值是 $\phi'(\xi)$; 存在某 $\eta \in (z, y)$ 使得右边比值是 $\phi'(\eta)$. 因为 $\phi''(\cdot) \geq 0$ 和 $\eta > \xi$, 有 $\phi'(\eta) \geq \phi'(\xi)$. \square

多元函数 f 凸性的充分条件: $\text{dom} f$ 是凸集, f 在 $\text{dom} f$ 上是连续的, 在 $\text{intdom} f$ 上二次可微, 且海森矩阵 f'' 是半正定的.

例4.3. 作为一个富有启发性的例子, 证明函数

$$f(\mathbf{x}) = \ln \left(\sum_{i=1}^n \exp(x_i) \right)$$

在 \mathbb{R}^n 上是凸的. 事实上, 经演算可得

$$\begin{aligned} \mathbf{h}^T f'(\mathbf{x}) &= \frac{\sum_i \exp(x_i) h_i}{\sum_i \exp(x_i)}, \\ \mathbf{h}^T f''(\mathbf{x}) \mathbf{h} &= -\frac{\left(\sum_i \exp(x_i) h_i \right)^2}{\left(\sum_i \exp(x_i) \right)^2} + \frac{\sum_i \exp(x_i) h_i^2}{\sum_i \exp(x_i)} \\ &= -\left(\frac{\sum_i \exp(x_i) h_i}{\sum_i \exp(x_i)} \right)^2 + \frac{\sum_i \exp(x_i)}{\sum_i \exp(x_i)} h_i^2 \end{aligned}$$

置 $p_i = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$, 有

$$\begin{aligned} \mathbf{h}^T f''(\mathbf{x}) \mathbf{h} &= \sum_i p_i h_i^2 - \left[\sum_i p_i h_i \right]^2 \\ &= \sum_i p_i h_i^2 - \left[\sum_i \sqrt{p_i} (\sqrt{p_i} h_i) \right]^2 \\ &\geq \sum_i p_i h_i^2 - \left[\sum_i (\sqrt{p_i})^2 \right] \left[\sum_i (\sqrt{p_i} h_i)^2 \right] \\ &= \sum_i p_i h_i^2 - \left[\sum_i p_i h_i^2 \right] = 0 \end{aligned}$$

其中的不等式是因为Cauchy-Schwarz不等式, 最后的等式利用了 $\sum_i p_i = 1$ 的事实.

命题4.6. 当 $c_i > 0$ 时, 函数 $g(\mathbf{y}) = \ln \left(\sum_i c_i \exp(a_i^T \mathbf{y}) \right)$ 是凸函数.

事实上, 凸函数

$$\ln \left(\sum_i \exp(x_i) \right)$$

通过自变量仿射替换得到

$$g(\mathbf{y}) = \ln \left(\sum_i \exp(\ln c_i + a_i^T \mathbf{y}) \right).$$

4.4 凸函数的性质

命题4.7 (梯度不等式). 设 $\mathbf{x} \in Q$ 是函数 f 定义域的内点, 其中集合 Q 是凸的并且 f 在 Q 上也是凸的. 假设 f 在 \mathbf{x} 处可微. 那么

$$\forall \mathbf{y} \in Q : f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^T f'(\mathbf{x}). \quad (\text{GI})$$

证明 设 $\mathbf{y} \in Q$. 当 $\mathbf{y} = \mathbf{x}$ 或者 $f(\mathbf{y}) = +\infty$ 时, 无需证明. 因此, 假设 $f(\mathbf{y}) < \infty$ 且 $\mathbf{y} \neq \mathbf{x}$. 设 $\mathbf{z}_\theta = \mathbf{x} + \theta(\mathbf{y} - \mathbf{x})$, $0 < \theta < 1$. 那么 \mathbf{z}_θ 是区间 $[\mathbf{x}, \mathbf{y}]$ 的内点. 由于 f 是凸的, 我们有

$$\frac{f(\mathbf{y}) - f(\mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} \geq \frac{f(\mathbf{z}_\theta) - f(\mathbf{x})}{\|\mathbf{z}_\theta - \mathbf{x}\|} = \frac{f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\theta\|\mathbf{y} - \mathbf{x}\|}.$$

当 $\theta \rightarrow +0$ 时取极限, 得到

$$\frac{f(\mathbf{y}) - f(\mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} \geq \frac{(\mathbf{y} - \mathbf{x})^T f'(\mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|}.$$

由此得(GI). □

命题4.8 (凸函数的Lipschitz连续性). 设 f 是凸函数, K 包含于 f 定义域的相对内部, 并且是有界闭集. 那么 f 在 K 上是Lipschitz 连续的, 即存在常数 $L < \infty$ 满足

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2 \quad \forall \mathbf{x}, \mathbf{y} \in K.$$

请注意, 下述例子表明该命题中关于集合 K 的所有三个假设都是必需的.

例4.4 (反例). (a) $f(x) = -\sqrt{x}$, $\text{dom} f = \{x \in \mathbb{R} : x \geq 0\}$, $K = [0, 1]$. 此处 $K \subset \text{dom} f$ 是有界闭集, 但没有包含于 $\text{dom} f$ 的相对内部, 且 f 在 K 上不是Lipschitz连续的;

(b) $f(x) = x^2$, $\text{dom} f = K = \mathbb{R}$. 此处 K 是闭集且包含于 $\text{rint dom} f$, 但无界, 并且 f 在 K 上不是Lipschitz 连续的;

(c) $f(x) = \frac{1}{x}$, $\text{dom} f = \{x \in \mathbb{R} : x > 0\}$, $K = (0, 1]$. 此处 K 是有界的且包含于 $\text{rint dom} f$, 但不是闭集, 并且 f 在 K 上不是Lipschitz连续的.

下面讨论凸函数的最大值和最小值.

命题4.9 (单峰性). 设 f 是凸函数, \mathbf{x}_* 是 f 的局部极小点:

$$\mathbf{x}_* \in \text{dom} f \quad \& \quad \exists r > 0 : f(\mathbf{x}) \geq f(\mathbf{x}_*) \quad \forall (\mathbf{x} : \|\mathbf{x} - \mathbf{x}_*\| \leq r).$$

那么 \mathbf{x}_* 是 f 的全局极小点, 即 $f(\mathbf{x}) \geq f(\mathbf{x}_*) \quad \forall \mathbf{x}$.

证明 欲证明: 若 $\mathbf{x} \neq \mathbf{x}_*$ 且 $\mathbf{x} \in \text{dom} f$, 那么 $f(\mathbf{x}) \geq f(\mathbf{x}_*)$. 为此设 $\mathbf{z} \in (\mathbf{x}_*, \mathbf{x})$. 由凸性有

$$\frac{f(\mathbf{z}) - f(\mathbf{x}_*)}{\|\mathbf{z} - \mathbf{x}_*\|} \leq \frac{f(\mathbf{x}) - f(\mathbf{x}_*)}{\|\mathbf{x} - \mathbf{x}_*\|}.$$

由 $\|\mathbf{z} - \mathbf{x}_*\| \leq r$ 且当 $\mathbf{z} \in (\mathbf{x}_*, \mathbf{x})$ 足够靠近 \mathbf{x}_* 时, 有 $\frac{f(\mathbf{z}) - f(\mathbf{x}_*)}{\|\mathbf{z} - \mathbf{x}_*\|} \geq 0$, 因此 $\frac{f(\mathbf{x}) - f(\mathbf{x}_*)}{\|\mathbf{x} - \mathbf{x}_*\|} \geq 0$, 即 $f(\mathbf{x}) \geq f(\mathbf{x}_*)$. \square

命题4.10. 设 f 是凸函数. 那么全局极小点集合 X_* 是凸集.

证明 这是如下重要引理的直接推论.

引理4.11 (凸函数的水平集是凸的). 设 f 是凸函数. 那么 f 的水平集, 即集合

$$X_a = \{\mathbf{x} \in \text{dom} f : f(\mathbf{x}) \leq a\}$$

都是凸的, 此处 a 是任意实数.

证明 如果 $\mathbf{x}, \mathbf{y} \in X_a, \theta \in [0, 1]$, 那么由 $\text{dom} f$ 凸知 $(1 - \theta)\mathbf{x} + \theta\mathbf{y} \in \text{dom} f$, 也有

$$f((1 - \theta)\mathbf{x} + \theta\mathbf{y}) \leq (1 - \theta)f(\mathbf{x}) + \theta f(\mathbf{y}) \leq (1 - \theta)a + \theta a = a.$$

从而 $[\mathbf{x}, \mathbf{y}] \subset X_a$. \square

何时凸函数的极小点唯一呢?

定义4.2. 称函数 f 是严格凸的, 如果

$$f((1 - \theta)\mathbf{x} + \theta\mathbf{y}) < (1 - \theta)f(\mathbf{x}) + \theta f(\mathbf{y}) \quad \forall \mathbf{x} \neq \mathbf{y}, \forall \theta \in (0, 1).$$

请注意, 如果凸函数 f 的定义域是开集, 在定义域上二次连续可微, 并且

$$\mathbf{h}^T f''(\mathbf{x}) \mathbf{h} > 0 \quad \forall (\mathbf{x} \in \text{dom} f, \mathbf{h} \neq 0),$$

那么 f 是严格凸的.

命题4.12. 如果严格凸函数 f 存在极小点, 那么极小点唯一.

证明 假设 $X_* = \text{argmin} f$ 含有两个不同点 $\mathbf{x}', \mathbf{x}''$. 根据严格凸性,

$$f(\frac{1}{2}\mathbf{x}' + \frac{1}{2}\mathbf{x}'') < \frac{1}{2}[f(\mathbf{x}') + f(\mathbf{x}'')] = \inf_{\mathbf{x}} f,$$

这是不可能的. \square

定理4.13 (凸极小的最优性条件). 设函数 f 在点 \mathbf{x}_* 处可微, 在包含 \mathbf{x}_* 的凸集 $Q \subset \text{dom} f$ 上是凸函数. f 在 Q 上于 \mathbf{x}_* 处取得最小值的充分必要条件是

$$(\mathbf{x} - \mathbf{x}_*)^T f'(\mathbf{x}_*) \geq 0 \quad \forall \mathbf{x} \in Q. \quad (4.6)$$

证明 充分性: 假设(4.6)成立, 证明对于任何 $\mathbf{x} \in Q$, $f(\mathbf{x}) \geq f(\mathbf{x}_*)$. 当 $\mathbf{x} = \mathbf{x}_*$ 时无需证明, 因此令 $f(\mathbf{x}) < \infty$, $\mathbf{x} \neq \mathbf{x}_*$. 对于 $\mathbf{z}_\theta = \mathbf{x}_* + \theta(\mathbf{x} - \mathbf{x}_*)$, 有

$$\frac{f(\mathbf{z}_\theta) - f(\mathbf{x}_*)}{\|\mathbf{z}_\theta - \mathbf{x}_*\|} \leq \frac{f(\mathbf{x}) - f(\mathbf{x}_*)}{\|\mathbf{x} - \mathbf{x}_*\|} \quad \forall \theta \in (0, 1)$$

或者(其是一样的)

$$\frac{f(\mathbf{x}_* + \theta[\mathbf{x} - \mathbf{x}_*]) - f(\mathbf{x}_*)}{\theta \|\mathbf{x} - \mathbf{x}_*\|} \leq \frac{f(\mathbf{x}) - f(\mathbf{x}_*)}{\|\mathbf{x} - \mathbf{x}_*\|} \quad \forall \theta \in (0, 1).$$

当 $\theta \rightarrow +0$ 时, 左边比式收敛到

$$(\mathbf{x} - \mathbf{x}_*)^T f'(\mathbf{x}_*) / \|\mathbf{x} - \mathbf{x}_*\| \geq 0;$$

因此,

$$\frac{f(\mathbf{x}) - f(\mathbf{x}_*)}{\|\mathbf{x} - \mathbf{x}_*\|} \geq 0,$$

从而 $f(\mathbf{x}) \geq f(\mathbf{x}_*)$.

必要性: 给定 $\mathbf{x}_* \in \text{argmin}_{\mathbf{y} \in Q} f(\mathbf{y})$, 设 $\mathbf{x} \in Q$. 那么

$$0 \leq \frac{f(\mathbf{x}_* + \theta[\mathbf{x} - \mathbf{x}_*]) - f(\mathbf{x}_*)}{\theta} \quad \forall \theta \in (0, 1),$$

因此 $(\mathbf{x} - \mathbf{x}_*)^T f'(\mathbf{x}_*) \geq 0$. □

凸极小的最优性条件的等价新表述: 设函数 f 在点 \mathbf{x}_* 处可微, 在凸集 $Q \subset \text{dom} f$ 上是凸函数. 考虑 Q 在 $\mathbf{x}_* \in Q$ 处的射线锥(radial cone):

$$T_Q(\mathbf{x}_*) = \{\mathbf{h} : \exists t > 0 : \mathbf{x}_* + t\mathbf{h} \in Q\}.$$

请注意, $T_Q(\mathbf{x}_*)$ 实际上是由形如 $s(\mathbf{x} - \mathbf{x}_*)$ 的所有向量组成的锥, 这里 $\mathbf{x} \in Q$, $s \geq 0$. f 在 \mathbf{x}_* 处取得最小值当且仅当

$$\mathbf{h}^T f'(\mathbf{x}_*) \geq 0 \quad \forall \mathbf{h} \in T_Q(\mathbf{x}_*),$$

或者(其是一样的)当且仅当

$$f'(\mathbf{x}_*) \in \underbrace{N_Q(\mathbf{x}_*) = \{\mathbf{g} : \mathbf{g}^T \mathbf{h} \geq 0 \quad \forall \mathbf{h} \in T_Q(\mathbf{x}_*)\}}_{Q \text{ 在 } \mathbf{x}_* \text{ 处的法锥}}. \quad (\text{NC})$$

例4.5. (a) 设 $\mathbf{x}_* \in \text{int } Q$. 这里 $T_Q(\mathbf{x}_*) = \mathbb{R}^n$, 因此 $N_Q(\mathbf{x}_*) = \{\mathbf{0}\}$, (NC) 变成 Fermat 方程

$$f'(\mathbf{x}_*) = \mathbf{0}.$$

(b) 设 $\mathbf{x}_* \in \text{rint } Q$. 令

$$\text{aff } Q = \mathbf{x}_* + L,$$

此处 L 是 \mathbb{R}^n 的线性子空间. 该问题中

$$T_Q(\mathbf{x}_*) = L,$$

因此 $N_Q(\mathbf{x}_*) = L^\perp$. (NC) 变成条件

$$f'(\mathbf{x}_*) \text{ 与 } L \text{ 正交.}$$

等价地, 设 $\text{aff } Q = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}\}$, 其中 $\mathbf{A}^T = [\mathbf{a}_1 \ \cdots \ \mathbf{a}_m]$. 那么

$$L = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}, L^\perp = \{\mathbf{y} = \mathbf{A}^T \boldsymbol{\mu} : \boldsymbol{\mu} \in \mathbb{R}^m\},$$

并且最优性条件变成: $\exists \boldsymbol{\mu}^*$ 满足

$$\nabla|_{\mathbf{x}=\mathbf{x}_*}[f(\mathbf{x}) + (\boldsymbol{\mu}^*)^T(\mathbf{A}\mathbf{x} - \mathbf{b})] = \mathbf{0},$$

这等价于

$$f'(\mathbf{x}_*) + \sum_{i=1}^m \mu_i^* \mathbf{a}_i = \mathbf{0}.$$

(c) 设 $Q = \{\mathbf{x} : \mathbf{A}\mathbf{x} - \mathbf{b} \leq \mathbf{0}\}$ 是多面集. 此处

$$T_Q(\mathbf{x}_*) = \{\mathbf{h} : \mathbf{a}_i^T \mathbf{h} \leq 0 \ \forall i \in \mathcal{I}(\mathbf{x}_*)\},$$

其中

$$\mathcal{I}(\mathbf{x}_*) = \{i : \mathbf{a}_i^T \mathbf{x}_* - b_i = 0\}.$$

根据齐次 Farkas 引理(定理 3.8),

$$\begin{aligned} N_Q(\mathbf{x}_*) &\equiv \{\mathbf{y} : \mathbf{a}_i^T \mathbf{h} \leq 0, i \in \mathcal{I}(\mathbf{x}_*) \Rightarrow \mathbf{y}^T \mathbf{h} \geq 0\} \\ &= \{\mathbf{y} = - \sum_{i \in \mathcal{I}(\mathbf{x}_*)} \lambda_i \mathbf{a}_i : \lambda_i \geq 0\}, \end{aligned}$$

并且最优性条件变成: $\exists (\lambda_i^* \geq 0, i \in \mathcal{I}(\mathbf{x}_*))$ 使得

$$f'(\mathbf{x}_*) + \sum_{i \in \mathcal{I}(\mathbf{x}_*)} \lambda_i^* \mathbf{a}_i = \mathbf{0}$$

或者意思一样: $\exists \lambda^* \geq 0$ 使得

$$f'(\mathbf{x}_*) + \sum_{i=1}^m \lambda_i^* \mathbf{a}_i = \mathbf{0},$$

$$\lambda_i^* (\mathbf{a}_i^T \mathbf{x}_* - b_i) = 0, i = 1, \dots, m.$$

关键是在凸函数情况下, 上述条件是 \mathbf{x}_* 为 f 在 Q 上最小点的充分必要条件.

例4.6. 考虑求解问题

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \mathbf{c}^T \mathbf{x} + \sum_{i=1}^m x_i \ln x_i : \mathbf{x} \geq 0, \sum_i x_i = 1 \right\}.$$

目标是凸的, 定义域 $Q = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0\}$ 是凸集(并且甚至是多面集). 假设在点 $\mathbf{x}_* \in \text{rint } Q$ 处取得最小值, 那么最优性条件变成

$$\nabla \left[\mathbf{c}^T \mathbf{x} + \sum_i x_i \ln x_i + \mu \left[\sum_i x_i - 1 \right] \right] = 0.$$

这等价于

$$\ln x_i = -c_i - \mu - 1 \quad \forall i,$$

即

$$x_i = \exp\{-1 - \mu\} \exp(-c_i).$$

由 $\sum_i x_i = 1$ 得到

$$x_i = \frac{\exp(-c_i)}{\sum_j \exp(-c_j)}, i = 1, \dots, n.$$

因此该点满足最优性, 的确是极小点.

命题4.14 (凸函数的最大值问题). 设 f 是凸函数. 那么

- (a) 如果 f 在 $\text{dom } f$ 上的最大值在点 $\mathbf{x}^* \in \text{rint dom } f$ 处取到, 那么 f 在 $\text{dom } f$ 上是常数.
- (b) 如果 $\text{dom } f$ 是闭集且不包含任何直线, 并且 f 在 $\text{dom } f$ 取到最大值, 那么在最大点中有 $\text{dom } f$ 的极点.
- (c) 如果 $\text{dom } f$ 是多面集并且 f 在 $\text{dom } f$ 上有界, 那么 f 在 $\text{dom } f$ 能取到它的最大值.

4.5 凸函数的次梯度及其运算法则

设 f 是凸函数, $\bar{\mathbf{x}} \in \text{int dom } f$. 如果 f 在 $\bar{\mathbf{x}}$ 处可微, 那么根据梯度不等式知, 存在仿射函数

$$h(\mathbf{x}) = f(\bar{\mathbf{x}}) + (\mathbf{x} - \bar{\mathbf{x}})^T f'(\bar{\mathbf{x}})$$

使得

$$f(\mathbf{x}) \geq h(\mathbf{x}) \quad \forall \mathbf{x} \quad \text{并且} \quad f(\bar{\mathbf{x}}) = h(\bar{\mathbf{x}}). \quad (4.7)$$

当 f 在 $\bar{\mathbf{x}} \in \text{dom } f$ 处不可微时, 也可能存在满足式(4.7)性质的仿射函数. 式(4.7)蕴含着: 对于某确定的 \mathbf{g} , 函数

$$h(\mathbf{x}) = f(\bar{\mathbf{x}}) + (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{g}$$

满足(4.7)当且仅当 \mathbf{g} 满足

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{g} \quad \forall \mathbf{x}. \quad (4.8)$$

定义4.3. 设 f 是凸函数, $\bar{\mathbf{x}} \in \text{dom} f$. 称满足不等式(4.8)中的 \mathbf{g} 是函数 f 在 $\bar{\mathbf{x}}$ 处的次梯度(subgradient). 称函数 f 在 $\bar{\mathbf{x}}$ 处所有次梯度(如果存在)组成的集合是函数 f 在 $\bar{\mathbf{x}}$ 处的次微分(subdifferential), 记为 $\partial f(\bar{\mathbf{x}})$.

例4.7. (a) 根据梯度不等式可知, 如果凸函数 f 在 $\bar{\mathbf{x}}$ 处可微, 那么 $\nabla f(\bar{\mathbf{x}}) \in \partial f(\bar{\mathbf{x}})$. 此外, 如果 $\bar{\mathbf{x}} \in \text{intdom} f$, 那么 $\nabla f(\bar{\mathbf{x}})$ 是 $\partial f(\bar{\mathbf{x}})$ 的唯一元素.

(b) 设 $f(x) = |x|$ ($x \in \mathbb{R}$). 当 $\bar{x} \neq 0$ 时, f 在 \bar{x} 处可微, 那么 $\partial f(\bar{x}) = f'(\bar{x})$. 当 $\bar{x} = 0$ 时, 次梯度 \mathbf{g} 由

$$|x| \geq 0 + gx = gx \quad \forall x,$$

给出, 即 $\partial f(0) = [-1, 1]$.

请注意, 该问题中, f 在每个点 $x \in \mathbb{R}$ 处沿方向 $h \in \mathbb{R}$ 的方向导数

$$Df(x)[h] = \lim_{t \rightarrow +0} \frac{f(x + th) - f(x)}{t}$$

存在, 而且该导数

$$Df(x)[h] = \max_{\mathbf{g} \in \partial f(x)} \mathbf{g}^T h.$$

命题4.15 (次微分的性质). 设 f 是凸函数. 那么

- (a) 对于每个 $\mathbf{x} \in \text{dom} f$, 次微分 $\partial f(\mathbf{x})$ 是闭凸集;
- (b) 如果 $\mathbf{x} \in \text{rintdom} f$, 那么 $\partial f(\mathbf{x})$ 非空;
- (c) 如果 $\mathbf{x} \in \text{rintdom} f$, 那么 f 沿每个方向 $\mathbf{h} \in \mathbb{R}^n$ 的方向导数 $Df(\mathbf{x})[\mathbf{h}]$ 存在, 并且

$$Df(\mathbf{x})[\mathbf{h}] \equiv \lim_{t \rightarrow +0} \frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t} = \max_{\mathbf{g} \in \partial f(\mathbf{x})} \mathbf{g}^T \mathbf{h}.$$

- (d) 假设将 $\bar{\mathbf{x}} \in \text{dom} f$ 表述为 $\lim_{i \rightarrow \infty} \mathbf{x}_i$, $\mathbf{x}_i \in \text{dom} f$, 并且

$$f(\bar{\mathbf{x}}) \leq \liminf_{i \rightarrow \infty} f(\mathbf{x}_i).$$

如果序列 $\mathbf{g}_i \in \partial f(\mathbf{x}_i)$ 收敛到某已知向量 \mathbf{g} , 那么 $\mathbf{g} \in \partial f(\bar{\mathbf{x}})$.

- (e) 集值映射 $\mathbf{x} \mapsto \partial f(\mathbf{x})$ 在每个 $\bar{\mathbf{x}} \in \text{intdom} f$ 处是局部有界的, 即当 $\bar{\mathbf{x}} \in \text{intdom} f$ 时, 存在 $r > 0$ 和 $R < \infty$ 使得

$$\|\mathbf{x} - \bar{\mathbf{x}}\|_2 \leq r, \quad \mathbf{g} \in \partial f(\mathbf{x}) \Rightarrow \|\mathbf{g}\|_2 \leq R.$$

选择部分命题证明 首先看命题(b)“如果 $\bar{\mathbf{x}} \in \text{rintdom} f$, 那么 $\partial f(\bar{\mathbf{x}})$ 非空.”

不失一般性, 设 $\text{dom} f$ 是满维的, 因此 $\bar{\mathbf{x}} \in \text{intdom} f$. 考虑凸集

$$T = \text{epi} f = \{(\mathbf{x}, t) : \mathbf{x} \in \text{dom} f, f(\mathbf{x}) \leq t\}.$$

由于 f 是凸函数, 故它在 $\text{intdom} f$ 上是连续的, 因此 T 的内部非空. 点 $(\bar{\mathbf{x}}, f(\bar{\mathbf{x}}))$ 显然不属于 T 的内部, 因此 $S = \{(\bar{\mathbf{x}}, f(\bar{\mathbf{x}}))\}$ 与 T 是可分离的: 存在 $(\mathbf{a}, b) \neq 0$ 满足

$$\mathbf{a}^T \bar{\mathbf{x}} + b f(\bar{\mathbf{x}}) \leq \mathbf{a}^T \mathbf{x} + b t \quad \forall (\mathbf{x}, t) \in T \quad (4.9)$$

显然有 $b \geq 0$. 否则当 $\mathbf{x} = \bar{\mathbf{x}}, t > f(\bar{\mathbf{x}})$ 很大时, (4.9) 不成立. 进一步可以断言: $b > 0$. 事实上, 当 $b = 0$ 时, (4.9) 式蕴含着

$$\mathbf{a}^T \bar{\mathbf{x}} \leq \mathbf{a}^T \mathbf{x} \quad \forall \mathbf{x} \in \text{dom} f. \quad (4.10)$$

由于 $(\mathbf{a}, b) \neq 0, b = 0$, 有 $\mathbf{a} \neq 0$; 但是 (4.10) 与 $\bar{\mathbf{x}} \in \text{intdom} f$ 矛盾.

由于 $b > 0$, (4.9) 蕴含着如果 $\mathbf{g} = -b^{-1}\mathbf{a}$, 那么

$$-\mathbf{g}^T \bar{\mathbf{x}} + f(\bar{\mathbf{x}}) \leq -\mathbf{g}^T \mathbf{x} + f(\mathbf{x}) \quad \forall \mathbf{x} \in \text{dom} f,$$

即

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{g} \quad \forall \mathbf{x}.$$

□

在应用中, 需要计算各种凸函数的次梯度. 下面是次梯度的基本运算规则.

命题4.16 (次梯度的基本运算规则). (a) 如果 $\mathbf{g}_i \in \partial f_i(\mathbf{x}), \lambda_i \geq 0$, 那么

$$\sum_i \lambda_i \mathbf{g}_i \in \partial \left(\sum_i \lambda_i f_i \right) (\mathbf{x}).$$

(b) 设

$$f(\cdot) = \sup_{\alpha \in \mathcal{A}} f_\alpha(\cdot),$$

并且 \mathbf{x} 使得

$$\mathcal{A}_*(\mathbf{x}) := \{\alpha \in \mathcal{A} : f(\mathbf{x}) = f_\alpha(\mathbf{x})\} \neq \emptyset.$$

如果 $\mathbf{g}_\alpha \in \partial f_\alpha(\mathbf{x}), \alpha \in \mathcal{A}_*$, 那么向量

$$\mathbf{g}_\alpha, \alpha \in \mathcal{A}_*(\mathbf{x})$$

的任意凸组合是 f 在 \mathbf{x} 处的次梯度.

(c) 如果 $\mathbf{g}_i \in \partial f_i(\mathbf{x}), i = 1, \dots, m, F(y_1, \dots, y_m)$ 是凸的和单调的, 并且

$$0 \leq d, \mathbf{g}_i \in \partial F(f_1(\mathbf{x}), \dots, f_m(\mathbf{x})),$$

那么向量 $\sum_i d_i \mathbf{g}_i$ 是 $F(f_1(\cdot), \dots, f_m(\cdot))$ 在 \mathbf{x} 处的梯度.

5 拉格朗日对偶和鞍点最优性条件

数学规划(mathematical programming)问题是

$$f_* = \min_x \left\{ \begin{array}{l} g(\mathbf{x}) \equiv (g_1(\mathbf{x}), \dots, g_m(\mathbf{x}))^T \leq 0 \\ f(\mathbf{x}) : h(\mathbf{x}) = (h_1(\mathbf{x}), \dots, h_k(\mathbf{x}))^T = 0 \\ \mathbf{x} \in X \end{array} \right\} \quad (\text{MP})$$

其中 \mathbf{x} 是设计向量, $f(\mathbf{x})$ 是目标函数,

$$g(\mathbf{x}) \equiv (g_1(\mathbf{x}), \dots, g_m(\mathbf{x}))^T \leq 0$$

是不等式约束,

$$h(\mathbf{x}) = (h_1(\mathbf{x}), \dots, h_k(\mathbf{x}))^T = 0$$

是等式约束, $X \subset \mathbb{R}^n$ 是定义域. 假设目标函数和约束在 X 上都有明确定义.

如果 \mathbf{x} 满足所有约束条件, 就称该解是可行的. 称具有可行解的问题是可行的. 如果目标函数在可行解集上有(下)界, 称问题(P)是有界的. 最优值

$$f_* = \begin{cases} \inf_x \{f(\mathbf{x}) : \mathbf{x} \text{ 是可行的}\}, & \text{(P)是可行的;} \\ +\infty, & \text{其它.} \end{cases}$$

对于可行有界的问题, f_* 是实数, 对于可行无界的问题, f_* 是 $-\infty$, 对于不可行的问题, f_* 是 $+\infty$. (P)的最优解 \mathbf{x}_* 是满足 $f(\mathbf{x}_*) = f_*$ 的可行解. 称有最优解的问题是可行的. 如果要表述严格, 应将(P)中的min换成inf.

5.1 凸规划

称问题(P)是凸的, 如果 X 是 \mathbb{R}^n 的凸子集, $f(\cdot), g_1(\cdot), \dots, g_m(\cdot)$ 是 X 上的实值凸函数, 没有等式约束. 尽管可以允许有线性等式约束, 但这并不具备普适性.

研究Lagrange对偶性的工具是关于择一的凸定理. 现在考虑如何验证系统

$$\begin{aligned} f(\mathbf{x}) &< c \\ g_j(\mathbf{x}) &\leq 0, j = 1, \dots, m \\ \mathbf{x} &\in X \end{aligned} \quad (\text{I})$$

是不可解的. 答案是假定存在非负权重 $\lambda_j, j = 1, \dots, m$, 使得不等式

$$f(\mathbf{x}) + \sum_{j=1}^m \lambda_j g_j(\mathbf{x}) < c$$

在 X 上是无解的, 即

$$\exists \lambda_j \geq 0 : \inf_{\mathbf{x} \in X} \left[f(\mathbf{x}) + \sum_{j=1}^m \lambda_j g_j(\mathbf{x}) \right] \geq c.$$

那么(I)是不可解的.

定理5.1 (凸择一定理). 考虑关于 \mathbf{x} 的约束系统(I)和连同关于 λ 的约束系统

$$\begin{aligned} \inf_{\mathbf{x} \in X} \left[f(\mathbf{x}) + \sum_{j=1}^m \lambda_j g_j(\mathbf{x}) \right] &\geq c \\ \lambda_j &\geq 0, j = 1, \dots, m. \end{aligned} \quad (\text{II})$$

(a) [平凡部分]如果(II)是可解的, 那么(I)是不可解的.

(b) [非平凡部分] 如果(I)不可解, 系统(I)是凸的(即 X 是凸集; f, g_1, \dots, g_m 是 X 上的实值凸函数), 并且子系统

$$\begin{aligned} g_j(\mathbf{x}) &< 0, j = 1, \dots, m, \\ \mathbf{x} &\in X. \end{aligned}$$

是可解的[Slater条件], 那么(II)是可解的.

证明 非平凡部分: 假设(I)无解. 考虑 \mathbb{R}^{m+1} 上的两个集合:

$$\begin{aligned} T &:= \left\{ \mathbf{u} \in \mathbb{R}^{m+1} : \exists \mathbf{x} \in X \text{ s.t. } \begin{aligned} &f(\mathbf{x}) \leq u_0, \\ &g_1(\mathbf{x}) \leq u_1, \\ &\vdots \\ &g_m(\mathbf{x}) \leq u_m \end{aligned} \right\}, \\ S &:= \{ \mathbf{u} \in \mathbb{R}^{m+1} : u_0 < c, u_1 \leq 0, \dots, u_m \leq 0 \}. \end{aligned}$$

观察到 S, T 是非空凸集; S 与 T 不相交(否则(I)有解). 从而由凸集分离定理(定理3.18)知 S 和 T 是可分离的:

$$\exists (a_0, \dots, a_m) \neq 0 : \inf_{\mathbf{u} \in T} \mathbf{a}^T \mathbf{u} \geq \sup_{\mathbf{u} \in S} \mathbf{a}^T \mathbf{u}.$$

即 $\exists (a_0, \dots, a_m) \neq 0$ 使得

$$\begin{aligned} &\inf_{\mathbf{x} \in X} \inf_{u_0, u_1, \dots, u_m} \{ \mathbf{a}^T \mathbf{u} : u_0 \geq f(\mathbf{x}), u_1 \geq g_1(\mathbf{x}), \dots, u_m \geq g_m(\mathbf{x}) \} \\ &\geq \sup_{u_0, u_1, \dots, u_m} \{ \mathbf{a}^T \mathbf{u} : u_0 < c, u_1 \leq 0, \dots, u_m \leq 0 \}. \end{aligned}$$

从而

$$\inf_{\mathbf{x} \in X} [a_0 f(\mathbf{x}) + a_1 g_1(\mathbf{x}) + \cdots + a_m g_m(\mathbf{x})] \geq a_0 c \quad (5.1)$$

并且 $\mathbf{a} \geq 0$, 即存在 $\exists \mathbf{a} \geq 0, \mathbf{a} \neq 0$ 使得(5.1)成立.

进一步观察到 $a_0 > 0$. 的确, 否则 $0 \neq (a_1, \dots, a_m) \geq 0$ 且

$$\inf_{\mathbf{x} \in X} [a_1 g_1(\mathbf{x}) + \cdots + a_m g_m(\mathbf{x})] \geq 0,$$

这与 $\exists \bar{\mathbf{x}} \in X$: 对所有 j 有 $g_j(\bar{\mathbf{x}}) < 0$ 矛盾. 由 $a_0 > 0$ 和(5.1)得到

$$\inf_{\mathbf{x} \in X} \left[f(\mathbf{x}) + \underbrace{\sum_{j=1}^m \left[\frac{a_j}{a_0} \right]}_{\lambda_j \geq 0} g_j(\mathbf{x}) \right] \geq c.$$

□

5.2 拉格朗日对偶

考虑优化问题

$$\text{Opt}(P) = \min \{ f(\mathbf{x}) : g_j(\mathbf{x}) \leq 0, j \leq m, \mathbf{x} \in X \} \quad (P)$$

和与之关联的拉格朗日函数(Lagrangian function)

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{j=1}^m \lambda_j g_j(\mathbf{x})$$

以及拉格朗日对偶问题

$$\text{Opt}(D) = \max_{\boldsymbol{\lambda} \geq 0} \underline{L}(\boldsymbol{\lambda}), \quad \underline{L}(\boldsymbol{\lambda}) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\lambda}). \quad (D)$$

定理5.2 (凸规划的对偶定理). (a) 弱对偶性: 对于每个 $\boldsymbol{\lambda} \geq 0$, $\underline{L}(\boldsymbol{\lambda}) \leq \text{Opt}(P)$, 特别地有

$$\text{Opt}(D) \leq \text{Opt}(P).$$

(b) 强对偶性 如果(P)是凸的, 有下界, 并且满足Slater条件, 那么(D)是可解的, 并且

$$\text{Opt}(D) = \text{Opt}(P).$$

证明. 弱对偶性: $\text{Opt}(D) \leq \text{Opt}(P)$. 当(P)不可行(即 $\text{Opt}(P) = \infty$)时, 该结论显然成立, 无需证明. 如果 \mathbf{x} 是(P)的可行解, 并且 $\boldsymbol{\lambda} \geq 0$, 那么 $L(\mathbf{x}, \boldsymbol{\lambda}) \leq f(\mathbf{x})$,

随之对 $\lambda \geq 0$, 有

$$\begin{aligned}\underline{L}(\lambda) &\equiv \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda) \\ &\leq \inf_{\mathbf{x} \in X \text{ 是可行的}} L(\mathbf{x}, \lambda) \\ &\leq \inf_{\mathbf{x} \in X \text{ 是可行的}} f(\mathbf{x}) \\ &= \text{Opt}(P)\end{aligned}$$

从而得

$$\text{Opt}(D) = \sup_{\lambda \geq 0} \underline{L}(\lambda) \leq \text{Opt}(P).$$

强对偶性: 如果(P)是凸的, 有下界, 并且满足Slater条件. 那么(D)是可解的, 并且 $\text{Opt}(D) = \text{Opt}(P)$. 系统

$$f(\mathbf{x}) < \text{Opt}(P), \quad g_j(\mathbf{x}) \leq 0, j = 1, \dots, m, \mathbf{x} \in X$$

无解, 同时问题

$$g_j(\mathbf{x}) < 0, j = 1, \dots, m, \mathbf{x} \in X$$

有解. 根据凸择一定理5.1,

$$\exists \lambda^* \geq 0 : f(\mathbf{x}) + \sum_j \lambda_j^* g_j(\mathbf{x}) \geq \text{Opt}(P) \quad \forall \mathbf{x} \in X,$$

再结合 $\underline{L}(\cdot)$ 的定义, 有

$$\underline{L}(\lambda^*) \geq \text{Opt}(P). \quad (5.2)$$

结合弱对偶性, (5.2)表明

$$\text{Opt}(D) = \underline{L}(\lambda^*) = \text{Opt}(P).$$

请注意就等价关系而言, 拉格朗日函数把(P)与(D)的关系牢记在心. 这种关系充其量为等价关系. 的确,

$$\text{Opt}(D) = \sup_{\lambda \geq 0} \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda)$$

是由拉格朗日函数给出的. 现在考虑函数

$$\bar{L}(\mathbf{x}) = \sup_{\lambda \geq 0} L(\mathbf{x}, \lambda) = \begin{cases} f(\mathbf{x}), & g_j(\mathbf{x}) \leq 0, j \leq m; \\ +\infty, & \text{otherwise.} \end{cases}$$

问题(P)显然等价于在 $\mathbf{x} \in X$ 上最小化 $\bar{L}(\mathbf{x})$ 的问题, 即

$$\text{Opt}(P) = \inf_{\mathbf{x} \in X} \sup_{\lambda \geq 0} L(\mathbf{x}, \lambda).$$

5.3 鞍点与最优性条件

设 $X \subset \mathbb{R}^n$, $\Lambda \subset \mathbb{R}^m$ 是非空集合, $F(\mathbf{x}, \boldsymbol{\lambda})$ 是 $X \times \Lambda$ 上的实值函数, 这个函数产生两个优化问题:

$$\text{Opt}(P) = \inf_{\mathbf{x} \in X} \overbrace{\sup_{\boldsymbol{\lambda} \in \Lambda} F(\mathbf{x}, \boldsymbol{\lambda})}^{\bar{F}(\mathbf{x})} \quad (P)$$

$$\text{Opt}(D) = \sup_{\boldsymbol{\lambda} \in \Lambda} \underbrace{\inf_{\mathbf{x} \in X} F(\mathbf{x}, \boldsymbol{\lambda})}_{\underline{F}(\boldsymbol{\lambda})} \quad (D)$$

博弈解释. 玩家 I 选择 $\mathbf{x} \in X$, 玩家 II 选择 $\boldsymbol{\lambda} \in \Lambda$. 对于玩家选取的 $\mathbf{x}, \boldsymbol{\lambda}$, 玩家 I 向玩家 II 支付 $F(\mathbf{x}, \boldsymbol{\lambda})$. 为了最优化他们的财富, 玩家们该如何博弈?

如果玩家 I 先选择 \mathbf{x} , 玩家 II 知道该如何选择, 此时他将选择 $\boldsymbol{\lambda}$ 以最大化自己利润, 而玩家 I 的损失将是 $\bar{F}(\mathbf{x})$. 为了最小化该损失, 玩家 I 将求解问题(P), 以确保自己的损失是 $\text{Opt}(P)$ 或更少.

如果玩家 II 先选择 $\boldsymbol{\lambda}$, 玩家 I 知道该如何选择, 此时他选择 \mathbf{x} , 则玩家 I 会最小化其损失, 而玩家 II 的利润将是 $\underline{F}(\boldsymbol{\lambda})$. 为了最大化利润, 玩家 II 将求解问题(D), 确保其利润是 $\text{Opt}(D)$ 或更多.

观察到第二种情况似乎对玩家 I 更有利, 因此自然而然地猜测: 他的预期损失在该情况下小于等于他在第一种情况下所预期的损失, 即

$$\text{Opt}(D) \equiv \sup_{\boldsymbol{\lambda} \in \Lambda} \inf_{\mathbf{x} \in X} F(\mathbf{x}, \boldsymbol{\lambda}) \leq \inf_{\mathbf{x} \in X} \sup_{\boldsymbol{\lambda} \in \Lambda} F(\mathbf{x}, \boldsymbol{\lambda}) \equiv \text{Opt}(P).$$

这的确是事实. 假定 $\text{Opt}(P) < \infty$ (否则该不等式关系是显然成立的),

$$\begin{aligned} \forall (\epsilon > 0) : \exists \mathbf{x}_\epsilon \in X : \sup_{\boldsymbol{\lambda} \in \Lambda} F(\mathbf{x}_\epsilon, \boldsymbol{\lambda}) &\leq \text{Opt}(P) + \epsilon \\ \Rightarrow \forall \boldsymbol{\lambda} \in \Lambda : \underline{F}(\boldsymbol{\lambda}) = \inf_{\mathbf{x} \in X} F(\mathbf{x}, \boldsymbol{\lambda}) &\leq F(\mathbf{x}_\epsilon, \boldsymbol{\lambda}) \leq \text{Opt}(P) + \epsilon \\ \Rightarrow \text{Opt}(D) \equiv \sup_{\boldsymbol{\lambda} \in \Lambda} \underline{F}(\boldsymbol{\lambda}) &\leq \text{Opt}(P) + \epsilon \\ \Rightarrow \text{Opt}(D) &\leq \text{Opt}(P). \end{aligned}$$

当玩家们同时做选择时, 他们该如何决策呢? 能够回答该问题的一种“良好情况”—— F 具有鞍点.

定义5.1. 称点 $(\mathbf{x}_*, \boldsymbol{\lambda}_*) \in X \times \Lambda$ 是 F 的鞍点, 如果

$$F(\mathbf{x}, \boldsymbol{\lambda}_*) \geq F(\mathbf{x}_*, \boldsymbol{\lambda}_*) \geq F(\mathbf{x}_*, \boldsymbol{\lambda}) \quad \forall (\mathbf{x} \in X, \boldsymbol{\lambda} \in \Lambda).$$

用博弈的术语说, 鞍点就是**平衡点**——在该点, 倘若对手保持自己的选择不变化的话, 玩家无法增加自己的财富.

命题5.3. F 有鞍点当且仅当(P)和(D)都是可解的, 且二者的最优值相等. 此时, F 的所有鞍点恰好是点对 $(\mathbf{x}_*, \boldsymbol{\lambda}^*)$, 其中 \mathbf{x}_* 是(P)的最优解, $\boldsymbol{\lambda}^*$ 是(D)的最优解.

证明 \Rightarrow : 假设 $(\mathbf{x}_*, \boldsymbol{\lambda}^*)$ 是 F 的鞍点, 让我们证明 \mathbf{x}_* 求解(P), $\boldsymbol{\lambda}^*$ 求解(D), 且 $\text{Opt}(P) = \text{Opt}(D)$.

的确由鞍点定义得

$$F(\mathbf{x}, \boldsymbol{\lambda}^*) \geq F(\mathbf{x}_*, \boldsymbol{\lambda}^*) \geq F(\mathbf{x}_*, \boldsymbol{\lambda}) \quad \forall (\mathbf{x} \in X, \boldsymbol{\lambda} \in \Lambda)$$

随之有

$$\begin{aligned} \text{Opt}(P) &\leq \overline{F}(\mathbf{x}_*) = \sup_{\boldsymbol{\lambda} \in \Lambda} F(\mathbf{x}_*, \boldsymbol{\lambda}) = F(\mathbf{x}_*, \boldsymbol{\lambda}^*), \\ \text{Opt}(D) &\geq \underline{F}(\boldsymbol{\lambda}^*) = \inf_{\mathbf{x} \in X} F(\mathbf{x}, \boldsymbol{\lambda}^*) = F(\mathbf{x}_*, \boldsymbol{\lambda}^*). \end{aligned}$$

由于 $\text{Opt}(P) \geq \text{Opt}(D)$, 从而证明了下述不等式链:

$$\text{Opt}(P) \leq \overline{F}(\mathbf{x}_*) = F(\mathbf{x}_*, \boldsymbol{\lambda}^*) = \underline{F}(\boldsymbol{\lambda}^*) \leq \text{Opt}(D)$$

中的不等号都取等号. 这样, \mathbf{x}_* 求解(P), $\boldsymbol{\lambda}^*$ 求解(D), 并且 $\text{Opt}(P) = \text{Opt}(D)$.

\Leftarrow : 假设(P)和(D)有最优解 \mathbf{x}_* 和 $\boldsymbol{\lambda}^*$, 并且 $\text{Opt}(P) = \text{Opt}(D)$, 证明 $(\mathbf{x}_*, \boldsymbol{\lambda}^*)$ 是鞍点.

由 \mathbf{x}_* 和 $\boldsymbol{\lambda}^*$ 的最优性, 有

$$\begin{aligned} \text{Opt}(P) &= \overline{F}(\mathbf{x}_*) = \sup_{\boldsymbol{\lambda} \in \Lambda} F(\mathbf{x}_*, \boldsymbol{\lambda}) \geq F(\mathbf{x}_*, \boldsymbol{\lambda}^*), \\ \text{Opt}(D) &= \underline{F}(\boldsymbol{\lambda}^*) = \inf_{\mathbf{x} \in X} F(\mathbf{x}, \boldsymbol{\lambda}^*) \leq F(\mathbf{x}_*, \boldsymbol{\lambda}^*). \end{aligned} \tag{5.3}$$

由于 $\text{Opt}(P) = \text{Opt}(D)$, 那么(5.3)中所有不等号都取等号, 因此

$$\sup_{\boldsymbol{\lambda} \in \Lambda} F(\mathbf{x}_*, \boldsymbol{\lambda}) = F(\mathbf{x}_*, \boldsymbol{\lambda}^*) = \inf_{\mathbf{x} \in X} F(\mathbf{x}, \boldsymbol{\lambda}^*).$$

□

定理5.4 (凸规划鞍点形式的最优性条件). 设 $\mathbf{x}_* \in X$.

- (a) 最优性的充分条件. 倘若能将 \mathbf{x}_* 扩展 $\boldsymbol{\lambda}^* \geq 0$ 得到拉格朗日函数在 $X \times \{\boldsymbol{\lambda} \geq 0\}$ 上的鞍点:

$$L(\mathbf{x}, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}_*, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}_*, \boldsymbol{\lambda}) \quad \forall (\mathbf{x} \in X, \boldsymbol{\lambda} \geq 0), \tag{5.4}$$

那么 \mathbf{x}_* 是(P)的最优解.

- (b) 最优性的必要条件. 如果 \mathbf{x}_* 是(P)的最优解, (P)是凸的, 并且满足Slater条件, 那么可将 \mathbf{x}_* 扩展 $\boldsymbol{\lambda}^* \geq 0$ 后得到拉格朗日函数在 $X \times \{\boldsymbol{\lambda} \geq 0\}$ 上的鞍点.

证明 \Rightarrow : “假设 $\mathbf{x}_* \in X, \exists \boldsymbol{\lambda}^* \geq 0$ 满足(5.4):

$$L(\mathbf{x}, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}_*, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}_*, \boldsymbol{\lambda}) \quad \forall (\mathbf{x} \in X, \boldsymbol{\lambda} \geq 0),$$

那么 \mathbf{x}_* 是(P)的最优解.”

显然,

$$\sup_{\boldsymbol{\lambda} \geq 0} L(\mathbf{x}_*, \boldsymbol{\lambda}) = \begin{cases} +\infty, & \mathbf{x}_* \text{ 是可行的;} \\ f(\mathbf{x}_*), & \text{其它.} \end{cases}$$

因此, $\boldsymbol{\lambda}^* \geq 0$ 且 $L(\mathbf{x}_*, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}_*, \boldsymbol{\lambda}) \quad \forall \boldsymbol{\lambda} \geq 0$ 等价于

$$g_j(\mathbf{x}_*) \leq 0 \quad \forall j \quad \text{和} \quad \lambda_j^* g_j(\mathbf{x}_*) = 0 \quad \forall j.$$

结果有 $L(\mathbf{x}_*, \boldsymbol{\lambda}^*) = f(\mathbf{x}_*)$, 随之由

$$L(\mathbf{x}, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}_*, \boldsymbol{\lambda}^*) \quad \forall \mathbf{x} \in X$$

得到

$$L(\mathbf{x}, \boldsymbol{\lambda}^*) \geq f(\mathbf{x}_*) \quad \forall \mathbf{x}. \quad (5.5)$$

由于对于 $\boldsymbol{\lambda} \geq 0$ 及所有可行的 \mathbf{x} , 有 $f(\mathbf{x}) \geq L(\mathbf{x}, \boldsymbol{\lambda})$, 因此(5.5)蕴含着

$$\mathbf{x} \text{ 是可行的} \Rightarrow f(\mathbf{x}) \geq f(\mathbf{x}_*).$$

\Leftarrow : 假设 \mathbf{x}_* 是凸问题(P)的最优解, 并且该凸问题满足Slater条件. 那么 $\exists \boldsymbol{\lambda}^* \geq 0$:

$$L(\mathbf{x}, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}_*, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}_*, \boldsymbol{\lambda}) \quad \forall (\mathbf{x} \in X, \boldsymbol{\lambda} \geq 0).$$

由拉格朗日对偶定理(定理5.2), $\exists \boldsymbol{\lambda}^* \geq 0$:

$$f(\mathbf{x}_*) = \underline{L}(\boldsymbol{\lambda}^*) \equiv \inf_{\mathbf{x} \in X} \left[f(\mathbf{x}) + \sum_j \lambda_j^* g_j(\mathbf{x}) \right]. \quad (5.6)$$

由于 \mathbf{x}_* 是可行解, 从而有

$$\inf_{\mathbf{x} \in X} \left[f(\mathbf{x}) + \sum_j \lambda_j^* g_j(\mathbf{x}) \right] \leq f(\mathbf{x}_*) + \sum_j \lambda_j^* g_j(\mathbf{x}_*) \leq f(\mathbf{x}_*).$$

根据式(5.6), 上式中最后的“ \leq ”取“ $=$ ”, 即 $\lambda_j^* g_j(\mathbf{x}_*) = 0 \quad \forall j, \boldsymbol{\lambda}^* \geq 0$, 由此得

$$f(\mathbf{x}_*) = L(\mathbf{x}_*, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}_*, \boldsymbol{\lambda}) \quad \forall \boldsymbol{\lambda} \geq 0. \quad (5.7)$$

由(5.6)有

$$L(\mathbf{x}, \boldsymbol{\lambda}^*) \geq f(\mathbf{x}_*) = L(\mathbf{x}_*, \boldsymbol{\lambda}^*).$$

再结合(5.7)得

$$F(\mathbf{x}, \boldsymbol{\lambda}^*) \geq F(\mathbf{x}_*, \boldsymbol{\lambda}^*) \geq F(\mathbf{x}_*, \boldsymbol{\lambda}), (\mathbf{x}, \boldsymbol{\lambda}) \in X \times \{\boldsymbol{\lambda} \geq 0\}.$$

所以 $(\mathbf{x}_*, \boldsymbol{\lambda}^*)$ 是 L 的鞍点. □

定理5.5 (凸规划的KKT最优性条件). 设(P)是凸规划, \mathbf{x}_* 是它的可行解, 函数 f, g_1, \dots, g_m 在 \mathbf{x}_* 处可微, 那么

(a) KKT条件. 存在拉格朗日乘子 $\boldsymbol{\lambda}^* \geq 0$ 使得

$$\nabla f(\mathbf{x}_*) + \sum_{j=1}^m \lambda_j^* \nabla g_j(\mathbf{x}_*) \in N_X^*(\mathbf{x}_*) \quad (5.8)$$

$$\lambda_j^* g_j(\mathbf{x}_*) = 0, \quad j \leq m \quad [\text{互补松弛性}] \quad (5.9)$$

为 \mathbf{x}_* 是最优解的充分条件.

(b) 如果(P)满足受限Slater条件, 即 $\exists \bar{\mathbf{x}} \in \text{rint } X$: 对于所有约束有 $g_j(\bar{\mathbf{x}}) \leq 0$, 对于所有非线性约束有 $g_j(\bar{\mathbf{x}}) < 0$, 那么KKT条件为 \mathbf{x}_* 是最优解的充分必要条件.

证明 \Rightarrow : 设(P)是凸的, \mathbf{x}_* 是(P)的可行解, 并且 f, g_j 在 \mathbf{x}_* 处可微. 假设此时KKT条件也成立, 即(5.8)和(5.9)成立, 那么 \mathbf{x}_* 是最优的.

的确, 互补松弛性和 $\boldsymbol{\lambda}^* \geq 0$ 确保(5.7)成立. 进一步, $L(\mathbf{x}, \boldsymbol{\lambda}^*)$ 在 $\mathbf{x} \in X$ 上凸, 且在 $\mathbf{x}_* \in X$ 处可微, 因此(5.8)意味着

$$L(\mathbf{x}, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}_*, \boldsymbol{\lambda}^*) \quad \forall \mathbf{x} \in X.$$

这样, 可将 \mathbf{x}_* 扩充成拉格朗日函数的鞍点, 从而 \mathbf{x}_* 是(P)的最优解.

\Leftarrow : [在Slater条件下] 设(P)是凸的且满足Slater条件, \mathbf{x}_* 是最优的, 并且 f, g_j 在 \mathbf{x}_* 处可微, 那么(5.8)和(5.9)成立.

根据鞍点最优性条件, 由 \mathbf{x}_* 的最优性得到: $\exists \boldsymbol{\lambda}^* \geq 0$ 使得 $(\mathbf{x}_*, \boldsymbol{\lambda}^*)$ 是 $L(\mathbf{x}, \boldsymbol{\lambda})$ 在 $X \times \{\boldsymbol{\lambda} \geq 0\}$ 上的鞍点, 这等价于

$$\lambda_j^* g_j(\mathbf{x}_*) = 0 \quad \forall j$$

和

$$\min_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\lambda}^*) = L(\mathbf{x}_*, \boldsymbol{\lambda}^*). \quad (5.10)$$

由于函数 $L(\mathbf{x}, \boldsymbol{\lambda}^*)$ 在 X 上关于 \mathbf{x} 是凸的, 在 $\mathbf{x}_* \in X$ 处可微, 关系式(5.10)意味着(5.8)成立. \square

例5.1 (应用举例). 假定 $a_i > 0, p \geq 1$, 考虑求解问题

$$\min_{\mathbf{x}} \left\{ \sum_i \frac{a_i}{x_i} : \mathbf{x} > 0, \sum_i x_i^p \leq 1 \right\}.$$

假定 $\mathbf{x}_* > 0$ 是该问题的解, 且满足 $\sum_i (x_i^*)^p = 1$, 由KKT条件得

$$\nabla_{\mathbf{x}} \left\{ \sum_i \frac{a_i}{x_i} + \lambda \left(\sum_i x_i^p - 1 \right) \right\} = 0 \Leftrightarrow \frac{a_i}{x_i^2} = p\lambda x_i^{p-1}$$

$$\sum_i x_i^p = 1$$

随之得 $x_i = c(\lambda) a_i^{\frac{1}{p+1}}$. 由于 $\sum_i x_i^p$ 应该等于 1, 得到

$$x_i^* = \frac{a_i^{\frac{1}{p+1}}}{\left(\sum_j a_j^{\frac{p}{p+1}} \right)^{\frac{1}{p}}}.$$

该点是可行的, 问题是凸的, 所以该点满足KKT条件, 所以 \mathbf{x}^* 是最优的!

5.4 鞍点的存在性

引理5.6 (MiniMax引理). 设 $f_i(\mathbf{x})$, $i = 1, \dots, m$, 是凸紧集 $X \subset \mathbb{R}^n$ 上的凸连续函数, 那么存在 $\theta^* \geq 0$, $\sum_i \theta_i^* = 1$ 使得

$$\min_{\mathbf{x} \in X} \max_{1 \leq i \leq m} f_i(\mathbf{x}) = \min_{\mathbf{x} \in X} \sum_i \theta_i^* f_i(\mathbf{x}).$$

请注意当 $\theta \geq 0$, $\sum_i \theta_i = 1$ 时, 有

$$\max_{1 \leq i \leq m} f_i(\mathbf{x}) \geq \sum_i \theta_i f_i(\mathbf{x}).$$

从而有

$$\min_{\mathbf{x} \in X} \max_i f_i(\mathbf{x}) \geq \min_{\mathbf{x} \in X} \sum_i \theta_i f_i(\mathbf{x}).$$

证明 记 $X_+ = \{(t, \mathbf{x}) : \mathbf{x} \in X\}$. 考虑优化问题

$$\min_{t, \mathbf{x}} \{t : f_i(\mathbf{x}) - t \leq 0, \quad i \leq m, \quad (t, \mathbf{x}) \in X_+\}. \quad (5.11)$$

易见该问题的最优值

$$t_* = \min_{\mathbf{x} \in X} \max_i f_i(\mathbf{x}).$$

显然该优化问题是凸的, 可解的, 且满足Slater条件, 从而由鞍点最优性条件(定理5.4): 存在 $\lambda^* \geq 0$ 和(5.11)的最优解 (\mathbf{x}_*, t_*) 一起使得 $(\mathbf{x}_*, t_*; \lambda^*)$ 是拉格朗日函数

在 $X^+ \times \{\lambda \geq 0\}$ 上的鞍点:

$$\min_{\mathbf{x} \in X, t} \left\{ t + \sum_i \lambda_i^* (f_i(\mathbf{x}) - t) \right\} = t_* + \sum_i \lambda_i^* (f_i(\mathbf{x}_*) - t_*) \quad (5.12)$$

$$\max_{\lambda \geq 0} \left\{ t_* + \sum_i \lambda_i (f_i(\mathbf{x}_*) - t_*) \right\} = t_* + \sum_i \lambda_i^* (f_i(\mathbf{x}_*) - t_*) \quad (5.13)$$

(5.13) 蕴含着 $t_* + \sum_i \lambda_i^* (f_i(\mathbf{x}_*) - t_*) = t_*$. (5.12) 意味着 $\sum_i \lambda_i^* = 1$. 这样, $\lambda^* \geq 0$, $\sum_i \lambda_i^* = 1$, 并且

$$\begin{aligned} \min_{\mathbf{x} \in X} \sum_i \lambda_i^* f_i(\mathbf{x}) &= \min_{\mathbf{x} \in X, t} \left\{ t + \sum_i \lambda_i^* (f_i(\mathbf{x}) - t) \right\} \\ &= t_* + \sum_i \lambda_i^* (f_i(\mathbf{x}_*) - t_*) \quad \text{因为(5.12)} \\ &= t_* \\ &= \min_{\mathbf{x} \in X} \max_i f_i(\mathbf{x}). \end{aligned}$$

□

定理5.7 (Sion-Kakutani). 设 $X \subset \mathbb{R}^n$, $\Lambda \subset \mathbb{R}^m$ 是非空闭凸集, $F(\mathbf{x}, \lambda) : X \times \Lambda \rightarrow \mathbb{R}$ 是连续函数, 其在 $\mathbf{x} \in X$ 上是凸的, 在 $\lambda \in \Lambda$ 上是凹的. 假定 X 是紧的, 并且存在 $\bar{\mathbf{x}} \in X$ 使得所有集合

$$\Lambda_a : \{\lambda \in \Lambda : F(\bar{\mathbf{x}}, \lambda) \geq a\}$$

都有界(比如 Λ 是有界的). 那么 F 在 $X \times \Lambda$ 上有鞍点.

证明 欲证明问题

$$\text{Opt}(P) = \inf_{\mathbf{x} \in X} \overbrace{\sup_{\lambda \in \Lambda} F(\mathbf{x}, \lambda)}^{\bar{F}(\mathbf{x})} \quad (5.14)$$

和

$$\text{Opt}(D) = \sup_{\lambda \in \Lambda} \underbrace{\inf_{\mathbf{x} \in X} F(\mathbf{x}, \lambda)}_{\underline{F}(\lambda)} \quad (5.15)$$

都是可解的, 且二者的最优值相等.

1⁰. 由于 X 是紧的且 $F(\mathbf{x}, \lambda)$ 在 $X \times \Lambda$ 上是连续的, 因此函数 $\underline{F}(\lambda)$ 在 Λ 上是连续的. 此外, 集合

$$\Lambda^a = \{\lambda \in \Lambda : \underline{F}(\lambda) \geq a\}$$

包含于集合

$$\Lambda_a = \{\lambda \in \Lambda : F(\bar{\mathbf{x}}, \lambda) \geq a\},$$

因此是有界的. 最后, Λ 是闭的, 因此具有有界水平集 Λ^a 的连续函数 $F(\cdot)$ 在闭集 Λ 上取到它的最大值, 因此, (5.15) 是可解的. 设 λ^* 是 (5.15) 的最优解.

2⁰. 考虑集合

$$X(\lambda) = \{x \in X : F(x, \lambda) \leq \text{Opt}(D)\}.$$

它们是紧集 X 的闭凸子集. 下面用反证法证明这些集合中的任何有限个的交集非空. 假设

$$X(\lambda^1) \cap \cdots \cap X(\lambda^N) = \emptyset.$$

因此 $\forall x \in X$ 有

$$\max_{j=1, \dots, N} F(x, \lambda^j) > \text{Opt}(D).$$

根据 MinMax 引理 (引理 5.6), 存在权重 $\theta_j \geq 0, \sum_j \theta_j = 1$, 使得

$$\min_{x \in X} \sum_j \theta_j F(x, \lambda^j) > \text{Opt}(D).$$

从而

$$\min_{x \in X} F(x, \underbrace{\sum_j \theta_j \lambda^j}_{\tilde{\lambda}}) > \text{Opt}(D).$$

然而, 这是不可能的.

3⁰. 因为当紧集的任何有限个闭凸子集的交集是非空的时, 由提高版的 Helley 定理 (定理 3.5) 知所有这些集合的交集是非空的:

$$\exists x_* \in X : F(x_*, \lambda) \leq \text{Opt}(D) \quad \forall \lambda.$$

由于 $\text{Opt}(P) \geq \text{Opt}(D)$, 因此这是可能的当且仅当 x_* 是 (5.14) 的最优解, 并且 $\text{Opt}(P) = \text{Opt}(D)$. □

6 数学规划的最优性条件

面临的情景是考虑数学规划问题

$$\min_{\mathbf{x}} \left\{ \begin{array}{l} (g_1(\mathbf{x}), \dots, g_m(\mathbf{x})) \leq 0 \\ f(\mathbf{x}) : (h_1(\mathbf{x}), \dots, h_k(\mathbf{x})) = 0 \\ \mathbf{x} \in X \end{array} \right\}.$$

感兴趣的问题是假设已经给出该问题的可行解 \mathbf{x}_* , 那么 \mathbf{x}_* 是最优解的条件(必要条件、充分条件、充分必要条件)是什么?

事实是除了凸规划外, 还没有可验证的关于全局最优性的局部充分条件. 然而存在关于局部(也因此关于全局)最优性的可验证局部必要条件和关于局部最优性的局部充分条件.

另外的**事实**是在关于局部最优性的现有条件都假设 $\mathbf{x}_* \in \text{int}X$, 从 \mathbf{x}_* 的局部最优性角度讲, 这与 $X = \mathbb{R}^n$ 的描述完全一样. 这样, 可以简化面临的情景. 已知数学规划问题

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) : \begin{array}{l} (g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_m(\mathbf{x})) \leq 0 \\ (h_1(\mathbf{x}), \dots, h_k(\mathbf{x})) = 0 \end{array} \right\}, \quad (6.1)$$

和它的可行解 \mathbf{x}_* , 关心 \mathbf{x}_* 是局部最优解(存在 $r > 0$ 使得对于每个可行且满足 $\|\mathbf{x} - \mathbf{x}_*\| \leq r$ 的 \mathbf{x} , 有 $f(\mathbf{x}) \geq f(\mathbf{x}_*)$)的必要/充分条件.

6.1 一阶最优性条件

讨论一阶条件的默认假设是在 \mathbf{x}_* 的邻域内, 目标函数和所有的约束都连续可微. **一阶最优性条件**由目标函数和约束在 \mathbf{x}_* 处的值和梯度表述. 除凸规划情况外, 仅能给出一阶必要条件.

假设 \mathbf{x}_* 是(6.1)的局部最优解, **思路**是在 \mathbf{x}_* 附近用线性规划

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} f(\mathbf{x}_*) + (\mathbf{x} - \mathbf{x}_*)^T f'(\mathbf{x}_*) \\ & \text{subject to } \overbrace{g_j(\mathbf{x}_*)}^0 + (\mathbf{x} - \mathbf{x}_*)^T g'_j(\mathbf{x}_*) \leq 0, \quad j \in \mathcal{J}(\mathbf{x}_*) \\ & \quad \underbrace{h_i(\mathbf{x}_*)}_0 + (\mathbf{x} - \mathbf{x}_*)^T h'_i(\mathbf{x}_*) = 0, \quad 1 \leq i \leq k \\ & \quad \left[\mathcal{J}(\mathbf{x}_*) = \{j : g_j(\mathbf{x}_*) = 0\} \right] \end{aligned}$$

近似(6.1). 请注意, 由于所有 $g_j(\cdot)$ 在 \mathbf{x}_* 处连续, 非积极的(那些 $g_j(\mathbf{x}_*) < 0$)不等式约束不影响 \mathbf{x}_* 的局部最优性, 因此在这个线性规划中没有出现. 因为去掉目标中

的常数 $f(\mathbf{x}_*)$ 不影响整个问题的最优解, 从而上述线性规划问题等价于

$$\min_{\mathbf{x}} \left\{ \begin{array}{l} (\mathbf{x} - \mathbf{x}_*)^T g'_j(\mathbf{x}_*) \leq 0, \quad j \in \mathcal{J}(\mathbf{x}_*) \\ (\mathbf{x} - \mathbf{x}_*)^T f'(\mathbf{x}_*) : (\mathbf{x} - \mathbf{x}_*)^T h'_i(\mathbf{x}_*) = 0, \quad i = 1, \dots, k \\ \mathcal{J}(\mathbf{x}_*) = \{j : g_j(\mathbf{x}_*) = 0\} \end{array} \right\} \quad (\text{LP})$$

自然而然地会猜测, 如果 \mathbf{x}_* 是(6.1)的局部最优解, 那么它也是(LP)的局部最优解. 线性规划是带仿射约束的凸规划, KKT 条件是最优性的充分必要条件:

\mathbf{x}_* 是(LP)的最优解

\Updownarrow

$$\exists(\lambda_j^* \geq 0, \quad j \in \mathcal{J}(\mathbf{x}_*), \quad \mu_i) : f'(\mathbf{x}_*) + \sum_{j \in \mathcal{J}(\mathbf{x}_*)} \lambda_j^* g'_j(\mathbf{x}_*) + \sum_{i=1}^k \mu_i h'_i(\mathbf{x}_*) = 0$$

\Updownarrow

$$\exists(\lambda_j^* \geq 0, \quad \mu_i^*) : f'(\mathbf{x}_*) + \sum_j \lambda_j^* g'_j(\mathbf{x}_*) + \sum_i \mu_i^* h'_i(\mathbf{x}_*) = 0$$

$$\lambda_j^* g_j(\mathbf{x}_*) = 0, \quad j = 1, \dots, m$$

命题6.1. 设 \mathbf{x}_* 是(6.1)的局部最优解. 假设在从(6.1)变成线性问题(LP)时仍然是局部最优的. 那么在 \mathbf{x}_* 处, KKT条件成立:

$$\begin{aligned} \exists(\lambda_j^* \geq 0, \quad \mu_i^*) : f'(\mathbf{x}_*) + \sum_j \lambda_j^* g'_j(\mathbf{x}_*) + \sum_i \mu_i^* h'_i(\mathbf{x}_*) &= 0 \\ \lambda_j^* g_j(\mathbf{x}_*) &= 0, \quad j = 1, \dots, m. \end{aligned} \quad (6.2)$$

为了使该命题有用, 需要给出“ \mathbf{x}_* 在从(6.1)变成(LP)时仍然是局部最优的”可验证的充分条件.

这种条件最自然的形式是正则性: 在 \mathbf{x}_* 处的所有积极约束在 \mathbf{x}_* 处的梯度是线性无关的. 当然, 根据定义, 所有等式约束在每个可行解处都是积极的.

命题6.2. 设 \mathbf{x}_* 是(6.1)的局部最优正则解. 那么 \mathbf{x}_* 是(LP)的最优解, 因此 \mathbf{x}_* 满足KKT条件(6.2).

证明基于分析中的一个重要事实——一种版本的隐函数定理.

定理6.3 (隐函数定理). 设 $\mathbf{x}_* \in \mathbb{R}^n$, $p_1(\mathbf{x}), \dots, p_L(\mathbf{x})$ 是实值函数, 满足在 \mathbf{x}_* 的邻域内 p_ℓ 是 $k \geq 1$ 阶连续可微, $p_\ell(\mathbf{x}_*) = 0$, 向量

$$\nabla p_1(\mathbf{x}_*), \dots, \nabla p_L(\mathbf{x}_*)$$

是线性无关的. 那么存在定义在原点的邻域 V 内的变量代换

$$\mathbf{y} \mapsto \mathbf{x} = \Phi(\mathbf{y})$$

将 V 一一映射到 \mathbf{x}_* 的邻域 B 上, 并且满足 $\mathbf{x}_* = \Phi(0)$, $\Phi : V \rightarrow B$ 及其逆映射 $\Phi^{-1} : B \rightarrow V$ 都是 k 阶连续可微的, 在坐标系 \mathbf{y} 中, 函数 p_ℓ 恰好变成坐标:

$$\mathbf{y} \in V \Rightarrow p_\ell(\Phi(\mathbf{y})) \equiv y_\ell, \quad \ell = 1, \dots, L.$$

设 \mathbf{x}_* 是(6.1)的正则局部最优解. 假设与需要证明的结论相反, \mathbf{x}_* 不是(LP)的最优解, 然后由此导出矛盾.

1⁰. 由于 $\mathbf{x} = \mathbf{x}_*$ 不是(LP)的最优解, 存在可行解 $\mathbf{x}' = \mathbf{x}_* + \mathbf{d}$ 满足

$$(\mathbf{x}' - \mathbf{x}_*)^T f'(\mathbf{x}_*) = \mathbf{d}^T f'(\mathbf{x}_*) < 0.$$

因此

$$\mathbf{d}^T f'(\mathbf{x}_*) < 0, \quad \underbrace{\mathbf{d}^T h'_i(\mathbf{x}_*) = 0}_{\forall i}, \quad \underbrace{\mathbf{d}^T g'_j(\mathbf{x}_*) \leq 0}_{\forall j \in \mathcal{J}(\mathbf{x}_*)}.$$

2⁰. 不失一般性, 假设 $\mathcal{J}(\mathbf{x}_*) = \{1, \dots, \ell\}$. 根据隐函数定理, 存在变量的连续可微局部代换

$$\mathbf{x} = \Phi(\mathbf{y}) \quad [\Phi(0) = \mathbf{x}_*],$$

在原点的某个邻域内满足

$$h_i(\Phi(\mathbf{y})) \equiv y_i, \quad g_j(\Phi(\mathbf{y})) \equiv y_{k+j}, \quad j = 1, \dots, \ell,$$

且该变量代换在 \mathbf{x}_* 的某个邻域内具有连续可微的逆变换 $\mathbf{y} = \Psi(\mathbf{x})$. 由于 $\Psi(\Phi(\mathbf{y})) \equiv \mathbf{y}$, 从而有

$$\Psi'(\mathbf{x}_*)\Phi'(0) = \mathbf{I}, \quad (6.3)$$

随之

$$\exists \mathbf{p} : \Phi'(0)\mathbf{p} = \mathbf{d}.$$

现在的形势是找到了光滑的局部变量代换 $\mathbf{x} = \Phi(\mathbf{y})$ ($\mathbf{y} = 0$ 对应于 $\mathbf{x} = \mathbf{x}_*$)和方向 \mathbf{p} , 使得在 $\mathbf{y} = 0$ 的邻域内有

- (a) $h_i(\Phi(\mathbf{y})) \equiv y_i, \quad i \leq k;$
- (b) $g_j(\Phi(\mathbf{y})) \equiv y_{k+j}, \quad j \leq \ell; \quad [\mathcal{J}(\mathbf{x}_*) = \{1, \dots, \ell\}]$
- (c) $[\Phi'(0)\mathbf{p}]^T h'_i(\mathbf{x}_*) = 0, \quad i \leq k;$
- (d) $[\Phi'(0)\mathbf{p}]^T g'_j(\mathbf{x}_*) \leq 0, \quad j \leq \ell;$
- (e) $[\Phi'(0)\mathbf{p}]^T f'(\mathbf{x}_*) < 0.$

考虑可微曲线

$$\mathbf{x}(t) = \Phi(t\mathbf{p}),$$

经演算有

$$\begin{aligned} t\mathbf{p}_i &\equiv h_i(\Phi(t\mathbf{p})) \Rightarrow p_i = [\Phi'(0)\mathbf{p}]^T h'_i(\mathbf{x}_*) = 0 \\ t\mathbf{p}_{k+j} &\equiv g_j(\Phi(t\mathbf{p})) \Rightarrow p_{k+j} = [\Phi'(0)\mathbf{p}]^T g'_j(\mathbf{x}_*) \leq 0. \end{aligned}$$

由此得

$$\underbrace{h_i(\mathbf{x}(t)) = t\mathbf{p}_i = 0}_{\forall i}, \quad \underbrace{g_j(\mathbf{x}(t)) = t\mathbf{p}_{k+j} \leq 0}_{\forall j \in \mathcal{J}(\mathbf{x}_*)}$$

从而, 对于所有小的 $t \geq 0$, $\mathbf{x}(t)$ 是可行的. 但是:

$$\frac{d}{dt} \Big|_{t=0} f(\mathbf{x}(t)) = [\Phi'(0)\mathbf{p}]^T f'(\mathbf{x}_*) < 0,$$

随之对于所有足够小的 $t > 0$ 有

$$f(\mathbf{x}(t)) < f(\mathbf{x}(0)) = f(\mathbf{x}_*),$$

这与 \mathbf{x}_* 的局部最优性矛盾.

6.2 二阶最优性条件

考虑目标函数是连续可微的无约束最小化问题

$$\min_{\mathbf{x}}. \quad (\text{UCP})$$

那么KKT条件简化成**Fermat**原理:

$$\text{如果 } \mathbf{x}_* \text{ 是 } (\text{UCP}) \text{ 的局部最优解, 那么 } \nabla f(\mathbf{x}_*) = 0.$$

Fermat原理是无约束最小化问题中二阶最优性必要条件的“一阶”部分: 如果 \mathbf{x}_* 是(UCP)的局部最优解, 并且 f 在 \mathbf{x}_* 的邻域内是二次可微的. 那么

$$\nabla f(\mathbf{x}_*) = 0 \ \& \ \nabla^2 f(\mathbf{x}_*) \succeq 0 \Leftrightarrow \mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} \geq 0 \ \forall \mathbf{d}.$$

事实上, 设 \mathbf{x}_* 是(UCP)的局部最优解; 那么对于合适的 $r_d > 0$,

$$\begin{aligned} 0 \leq t \leq r_d &\Rightarrow 0 \leq f(\mathbf{x}_* + t\mathbf{d}) - f(\mathbf{x}_*) \\ &= \underbrace{t \mathbf{d}^T \nabla f(\mathbf{x}_*)}_{=0} + \frac{1}{2} t^2 \mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} + t^2 \underbrace{R_d(t)}_{\rightarrow 0, t \rightarrow 0} \\ &\Rightarrow \frac{1}{2} \mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} + R_d(t) \geq 0 \Rightarrow \mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} \geq 0. \end{aligned}$$

可将无约束最小化问题的二阶必要最优性条件加强成无约束最小化问题的二阶充分最优性条件：设 f 在 \mathbf{x}_* 的某邻域内是二次可微的。如果

$$\nabla f(\mathbf{x}_*) = 0, \quad \nabla^2 f(\mathbf{x}_*) \succ 0 \Leftrightarrow \mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} > 0 \quad \forall \mathbf{d} \neq 0$$

那么 \mathbf{x}_* 是(UCP)的严格局部最优解。

证明 由于对于所有 $\mathbf{d} > 0$, $\mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} > 0$, 因此存在 $\alpha > 0$ 使得对于所有 \mathbf{d} ,

$$\mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} \geq \alpha \mathbf{d}^T \mathbf{d}.$$

根据可微性, 对于任何 $t > 0$, $\|\mathbf{d}\| = 1$, 有

$$f(\mathbf{x}_* + t\mathbf{d}) - f(\mathbf{x}_*) = t \underbrace{\mathbf{d}^T \nabla f(\mathbf{x}_*)}_{=0} + \frac{t^2}{2} \underbrace{\mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d}}_{\geq \alpha \mathbf{d}^T \mathbf{d} = \alpha} + t^2 \underbrace{R_{\mathbf{d}(t)}}_{\rightarrow 0, t \rightarrow 0}$$

由 $R_{\mathbf{d}(t)} \rightarrow 0$ 知存在 $\delta > 0$, 使得 $0 < t < \delta$ 时, $|R_{\mathbf{d}(t)}| < \frac{\alpha}{4}$. 综上, 知对任意的 \mathbf{d} 使得 $\|\mathbf{d}\| = 1$, $0 < t < \delta$, 有

$$f(\mathbf{x}_* + t\mathbf{d}) - f(\mathbf{x}_*) \geq \frac{t^2}{2}(\alpha - \frac{\alpha}{2}) > 0$$

因此 \mathbf{x}_* 是 f 的局部最小点. □

已知数学规划问题(6.1), 与其关联的拉格朗日函数

$$L(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_j \lambda_j g_j(\mathbf{x}) + \sum_i \mu_i h_i(\mathbf{x}).$$

在关于约束优化问题(6.1)的二阶最优性条件中, 拉格朗日函数的海森矩阵扮演着 $\nabla^2 f(\mathbf{x}_*)$ 的角色。

定理6.4 (二阶必要最优性条件). 设 \mathbf{x}_* 是(6.1)的正则可行解, 且函数 f, g_j, h_i 在 \mathbf{x}_* 的某邻域内是二次连续可微的。如果 \mathbf{x}_* 是局部最优的, 那么存在唯一的拉格朗日乘子 $\lambda_j^* \geq 0, \mu_i^*$, 使得KKT条件(6.2)成立, 并且对正交于所有等式约束以及 \mathbf{x}_* 处所有的积极不等式约束在 \mathbf{x}_* 处梯度的每个 \mathbf{d} , 有

$$\mathbf{d}^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}_*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \mathbf{d} \geq 0.$$

证明 1⁰. 显然, \mathbf{x}_* 处的非积极约束既不影响 \mathbf{x}_* 的局部最优性, 也不影响欲证明的结论。因此, 把问题简化成所有约束在 \mathbf{x}_* 处都是积极的。

2⁰. 应用隐函数定理, 可找到局部变量代换和逆变换

$$\mathbf{x} = \Phi(\mathbf{y}) \Leftrightarrow \mathbf{y} = \Psi(\mathbf{x}) \quad [\Phi(\mathbf{0}) = \mathbf{x}_*, \Psi(\mathbf{x}_*) = \mathbf{0}],$$

满足

$$h_i(\Phi(\mathbf{y})) \equiv y_i, \quad i \leq k, \quad g_j(\Phi(\mathbf{y})) \equiv y_{m+j}, \quad j \leq m,$$

并且 Φ 和 Ψ 均是局部二次连续可微的. 从而, 以 \mathbf{y} 为变量, 问题(6.1)变成

$$\min_{\mathbf{y} \in \mathbb{R}^n} \left\{ \phi(\mathbf{y}) \equiv f(\Phi(\mathbf{y})) : \begin{array}{l} y_i = 0, \quad i \leq k \\ y_{k+j} = 0, \quad j \leq m \end{array} \right\}, \quad (6.4)$$

与(6.4)关联的拉格朗日函数

$$M(\mathbf{y}; \boldsymbol{\lambda}, \boldsymbol{\mu}) = \phi(\mathbf{y}) + \sum_{i=1}^k \mu_i y_i + \sum_{j=1}^m \lambda_j y_{k+j}.$$

思路: 由于 Φ 是从 \mathbf{x}_* 的一个邻域到 $\mathbf{y}_* = \mathbf{0}$ 的一个邻域的光滑一一映射, 因此 \mathbf{x}_* 是(6.1)的局部最优解当且仅当 $\mathbf{y}_* = \mathbf{0}$ 是(6.4)的局部最优解. 从而, 计划为“ $\mathbf{y}_* = \mathbf{0}$ 是(6.4)的局部最优解”建立必要/充分条件; “转换成”变量 \mathbf{x} 后, 这些条件将蕴含着“ \mathbf{x}_* 是(6.1)的局部最优解”的必要/充分条件.

3⁰. 由于 $\mathbf{x}_* = \Phi(\mathbf{0})$ 是(6.1)的局部最优解, 因此, $\mathbf{y}_* = \mathbf{0}$ 是(6.4)的局部最优解. 特别地, 如果 \mathbf{e}_i 是第 i 个标准正交向量, 那么对于合适的 $\epsilon > 0$:

$$\begin{aligned} s > m + k &\Rightarrow \text{当 } \epsilon \geq t \geq -\epsilon \text{ 时, } \mathbf{y}(t) = te_s \text{ 是(6.4)的可行解} \\ &\Rightarrow \frac{\partial \phi(\mathbf{0})}{\partial y_s} = \frac{d}{dt} \Big|_{t=0} \phi(\mathbf{y}(t)) = 0. \end{aligned}$$

和

$$\begin{aligned} j \leq m &\Rightarrow \text{当 } 0 \leq t \leq \epsilon \text{ 时, } \mathbf{y}(t) = -te_{k+j} \text{ 是(6.4)的可行解} \\ &\Rightarrow -\frac{\partial \phi(\mathbf{0})}{\partial y_{m+j}} = -\frac{d}{dt} \Big|_{t=0} \phi(\mathbf{y}(t)) \geq 0 \\ &\Rightarrow \lambda_j^* \equiv -\frac{\partial \phi(\mathbf{0})}{\partial y_{m+j}} \geq 0 \end{aligned}$$

置 $\mu_i^* = -\frac{\partial \phi(\mathbf{0})}{\partial y_i}$, $i = 1, \dots, k$, 得到

$$\boldsymbol{\lambda}^* \geq \mathbf{0} \quad \& \quad \nabla_{\mathbf{y}} M(\mathbf{0}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \mathbf{0}. \quad (\text{KKT})$$

现在的局势是 $\mathbf{y}_* = \mathbf{0}$ 是(6.4)的局部最优解, 并且 $\exists \boldsymbol{\lambda}^* \geq \mathbf{0}, \boldsymbol{\mu}^*$ 使得与之关联的拉格朗日函数满足:

$$0 = \frac{\partial M(\mathbf{0}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)}{\partial y_i} \equiv \begin{cases} \frac{\partial \phi(\mathbf{0})}{\partial y_i} + \mu_i^*, & i \leq k \\ \frac{\partial \phi(\mathbf{0})}{\partial y_i} + \lambda_{i-k}^*, & k < i \leq m + k \\ \frac{\partial \phi(\mathbf{0})}{\partial y_i}, & i > m + k \end{cases} \quad (\text{KKT})$$

请注意, 条件 $\nabla_{\mathbf{y}} M(\mathbf{0}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \mathbf{0}$ 唯一的定义了 $\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*$.

4⁰. 由上分析知道对于(6.4), 二阶必要最优性条件的一阶部分仍然成立. 下面证明该条件的二阶部分, 即

$$\forall (\mathbf{d} : \mathbf{d}^T \nabla_{\mathbf{y}} y_i = 0, \quad i \leq m + k) : \mathbf{d}^T \nabla_{\mathbf{y}}^2 M(\mathbf{0}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \mathbf{d} \geq 0. \quad (6.5)$$

这是显而易见的. 因为

$$M(\mathbf{y}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \phi(\mathbf{y}) + \sum_{i=1}^k \mu_i^* y_i + \sum_{j=1}^m \lambda_j^* y_{m+j},$$

从而有

$$\nabla_{\mathbf{y}}^2 M(\mathbf{0}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \nabla^2 \phi(\mathbf{0}).$$

所以需要证明: 对于来自线性子空间

$$L = \{\mathbf{d} \in \mathbb{R}^n : d_1 = \cdots = d_{m+k} = 0\}$$

的每个向量 \mathbf{d} , 有 $\mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} \geq 0$. 但是该子空间是(6.4)的可行子空间, 因此 ϕ (被限制在 L 上)应该在原点处取得无约束局部最小值. 根据无约束最小化的二阶必要最优性条件, 可得

$$\mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} \geq 0 \quad \forall \mathbf{d} \in L.$$

5⁰. 已经知道, 如果 \mathbf{x}_* 是(6.1)的局部最优解, 那么存在唯一的 $\boldsymbol{\lambda}^* \geq \mathbf{0}$, $\boldsymbol{\mu}^*$ 满足(KKT), 并且有(6.5)成立. 接下来证明

$$\nabla_{\mathbf{x}} L(\mathbf{x}_*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \mathbf{0} \tag{6.6}$$

和

$$\left. \begin{array}{l} \mathbf{p}^T g'_j(\mathbf{x}_*) = 0, \quad j \leq m \\ \mathbf{p}^T h'_i(\mathbf{x}_*) = 0, \quad i \leq k \end{array} \right\} \Rightarrow \mathbf{p}^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}_*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \mathbf{p} \geq 0. \tag{6.7}$$

设

$$\mathcal{L}(\mathbf{x}) = L(\mathbf{x}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*), \quad \mathcal{M}(\mathbf{y}) = M(\mathbf{y}; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*).$$

那么 $\mathcal{L}(\mathbf{x}) = \mathcal{M}(\Psi(\mathbf{x}))$. 从而有

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_*) = [\Psi'(\mathbf{x}_*)]^T \nabla_{\mathbf{y}} \mathcal{M}(\mathbf{y}_*) = \mathbf{0}.$$

此即(6.6). 设 \mathbf{p} 满足(6.7)中的前提条件, 并设 $\mathbf{d} = [\Phi'(\mathbf{0})]^{-1} \mathbf{p}$, 那么

$$\begin{aligned} \underbrace{\frac{d}{dt} \Big|_{t=0} t d_i}_{\frac{d}{dt} \Big|_{t=0} h_i(\Phi(t\mathbf{d}))} &= [h'_i(\mathbf{x}_*)]^T \underbrace{[\Phi'(\mathbf{0})] \mathbf{d}}_{\mathbf{p}} \Rightarrow d_i = \mathbf{p}^T h'_i(\mathbf{x}_*) = 0, \quad i \leq k, \\ \underbrace{\frac{d}{dt} \Big|_{t=0} g_j(\Phi(t\mathbf{d}))}_{\frac{d}{dt} \Big|_{t=0} g_{k+j}(\Phi(t\mathbf{d}))} &= [g'_j(\mathbf{x}_*)]^T \underbrace{[\Phi'(\mathbf{0})] \mathbf{d}}_{\mathbf{p}} \Rightarrow d_{k+j} = \mathbf{p}^T g'_j(\mathbf{x}_*) = 0, \quad j \leq m. \end{aligned}$$

从而有

$$\begin{aligned}
\mathbf{p}^T \nabla^2 \mathcal{L}(\mathbf{x}_*) \mathbf{p} &= \frac{d^2}{dt^2} \Big|_{t=0} \mathcal{L}(\mathbf{x}_* + t\mathbf{p}) = \frac{d^2}{dt^2} \Big|_{t=0} \mathcal{M}(\Psi(\mathbf{x}_* + t\mathbf{p})) \\
&= \frac{d}{dt} \Big|_{t=0} [\mathbf{p}^T [\Psi'(\mathbf{x}_* + t\mathbf{p})]^T \nabla \mathcal{M}(\Psi(\mathbf{x}_* + t\mathbf{p}))] \\
&= \mathbf{p}^T [\Psi'(\mathbf{x}_*)]^T \nabla^2 \mathcal{M}(\mathbf{0}) \overbrace{[\Psi'(\mathbf{x}_*)\mathbf{p}]}^{=[\Phi'(\mathbf{0})]^{-1}\mathbf{p}=\mathbf{d}} + \mathbf{p}^T \left[\frac{d}{dt} \Big|_{t=0} \Psi'(\mathbf{x}_* + t\mathbf{p}) \right]^T \underbrace{\nabla \mathcal{M}(\mathbf{0})}_{=\mathbf{0}} \\
&= \mathbf{d}^T \nabla^2 \mathcal{M} \mathbf{d} \geq 0 \quad \text{由于 } d_j = 0, 1 \leq j \leq k+m.
\end{aligned}$$

其中第四个等式中, 由(6.3)有 $\Psi'(\mathbf{x}_*) = [\Phi'(\mathbf{0})]^{-1}$. 这样, 只要 \mathbf{p} 与 \mathbf{x}_* 处所有积极约束的梯度正交, 就有 $\mathbf{p}^T \nabla^2 \mathcal{L} \mathbf{p} \geq 0$.

定理6.5 (二阶充分最优性条件). 设 \mathbf{x}_* 是(6.1)的正则可行解, 并且函数 f, g_j, h_i 在 \mathbf{x}_* 的邻域内是二次连续可微的, 如果存在拉格朗日乘子 $\lambda_j^* \geq 0, \mu_i^*$, 使得KKT条件(6.2)成立, 并且对于每个 $\mathbf{0} \neq \mathbf{d}$ 正交于所有等式约束的梯度并正交于 $\lambda_j^* > 0$ 的 \mathbf{x}_* 处的积极不等式约束在 \mathbf{x}_* 处的梯度, 有

$$\mathbf{d}^T \nabla_x^2 L(\mathbf{x}_*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \mathbf{d} > 0.$$

那么 \mathbf{x}_* 是(6.1)的局部最优解.

请注意最优性的充分条件和必要条件之间的差异在于它们的“二阶”部分, 并且是双重的: 次要差异是必要条件说明 $\nabla_x^2 L(\mathbf{x}_*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ 沿线性子空间 T 的半正定性:

$$\forall \mathbf{d} \in T = \left\{ \mathbf{d} \in \mathbb{R}^n : \overbrace{\mathbf{d}^T h'_i(\mathbf{x}_*)}^{\forall i \leq k} = 0, \overbrace{\mathbf{d}^T g'_j(\mathbf{x}_*)}^{\forall j \in \mathcal{J}(\mathbf{x}_*)} = 0 \right\} : \mathbf{d}^T \nabla_x^2 L(\mathbf{x}_*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \mathbf{d} \geq 0.$$

而充分条件要求 $\nabla_x^2 L(\mathbf{x}_*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ 沿线性子空间 T^+ 的半正定性:

$$\forall \mathbf{0} \neq \mathbf{d} \in T^+ = \left\{ \mathbf{d} \in \mathbb{R}^n : \overbrace{\mathbf{d}^T h'_i(\mathbf{x}_*)}^{\forall i \leq k} = 0, \overbrace{\mathbf{d}^T g'_j(\mathbf{x}_*)}^{\forall j \leq m: \lambda_j^* > 0} = 0 \right\} : \mathbf{d}^T \nabla_x^2 L(\mathbf{x}_*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \mathbf{d} > 0.$$

主要差异是提及的两个线性子空间是不同的, 并且 $T \subset T^+$; 两个子空间相等当且仅当 \mathbf{x}_* 处所有积极不等式约束具有正的拉格朗日乘子 λ_j^* .

例6.1. 下面的例子说明这种二阶必要条件和二阶充分条件之间的“间隙”是本质的. 考虑

$$\min_{x_1, x_2} \{ f(\mathbf{x}) = x_2^2 - x_1^2 : g_1(\mathbf{x}) \equiv x_1 \leq 0 \} \quad [\mathbf{x}_* = (0, 0)].$$

这里, 必要的二阶最优性条件“严格”满足:

$$L(\mathbf{x}; \lambda) = x_2^2 - x_1^2 + \lambda x_1,$$

随之

$$\begin{aligned}\lambda^* = 0 &\Rightarrow \nabla_{\mathbf{x}} L(\mathbf{x}_*; \lambda^*) = 0, \\ T &= \{\mathbf{d} \in \mathbb{R}^2 : \mathbf{d}^T g'_1(0) = 0\} = \{\mathbf{d} : d_1 = 0\}, \\ 0 \neq \mathbf{d} \in T &\Rightarrow \mathbf{d}^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}_*; \lambda^*) \mathbf{d} = 2d_2^2 > 0\end{aligned}$$

然而 \mathbf{x}_* 不是局部解. 该例中 $T^+ = \mathbb{R}^2$.

证明充分的二阶最优性条件. 1^0 . 与在讨论二阶必要的最优性条件时所用的处理方法一样, 将问题简化成所有不等式约束在 \mathbf{x}_* 处是积极的, 即(6.4). 对于(6.4)的情况, 充分条件是: $\exists \lambda^* \geq 0, \mu^*$ 使得

$$\begin{aligned}\nabla_{\mathbf{y}}|_{\mathbf{y}=\mathbf{0}} \left\{ \phi(\mathbf{y}) + \sum_{i=1}^k \mu_i^* y_i + \sum_{j=1}^m \lambda_j^* y_{m+j} \right\} &= \mathbf{0} \\ d_j = 0, j \in \mathcal{J}, \mathbf{d} \neq \mathbf{0} &\Rightarrow \mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} > 0 \\ [\mathcal{J} = \{1, \dots, k\} \cup \{k+j : j \leq m, \lambda_j^* > 0\}] &\end{aligned} \quad (6.8)$$

不失一般性, 假设 $\{j : \lambda_j^* > 0\} = \{1, \dots, q\}$, 那么由(6.8)看出

$$\begin{aligned}\frac{\partial \phi(\mathbf{0})}{\partial y_{k+j}} &< 0, j = 1, \dots, q \\ \frac{\partial \phi(\mathbf{0})}{\partial y_{k+j}} &= 0, j = q+1, \dots, m \\ \frac{\partial \phi(\mathbf{0})}{\partial y_i} &= 0, i = m+k+1, \dots, n \\ \mathbf{0} \neq \mathbf{d} \in T^+ &= \{\mathbf{d} \in \mathbb{R}^n : d_i = 0, i = 1, \dots, k, k+1, \dots, k+q\} \\ &\Rightarrow \mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} > 0\end{aligned} \quad (6.9)$$

目标源自 $\mathbf{y}_* = \mathbf{0}$ 是(6.4)的局部最优解的假设.

2^0 . (6.4)的可行集是闭锥

$$K = \{\mathbf{d} \in \mathbb{R}^n : d_i = 0, i = 1, \dots, k, d_{k+j} \leq 0, j = 1, \dots, m\}. \quad (6.10)$$

引理: 对于 $\mathbf{0} \neq \mathbf{d} \in K$, 有 $\mathbf{d}^T \nabla \phi(\mathbf{0}) \geq 0$, 并且 $\mathbf{d}^T \nabla \phi(\mathbf{0}) = 0$ 蕴含着 $\mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} > 0$.

证明 对于 $\mathbf{d} \in K$, 有

$$\mathbf{d}^T \nabla \phi(\mathbf{0}) = \sum_{i=1}^n \frac{\partial \phi(\mathbf{0})}{\partial y_i} d_i.$$

根据(6.9)和(6.10), 求和式中间的 q 项是非负的, 其余的是 0 , 因此和总是大于等于 0 . 对于 $\mathbf{d} \neq \mathbf{0}$, 要使和消失, 唯一的可能是 $\mathbf{d} \in T^+$, 并且在这种情况下, $\mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} > 0$. “图解”说明如下:

$$\begin{array}{l}
\frac{\partial \phi(\mathbf{0})}{\partial y_i} = 0, 1 \leq i \leq k \\
\frac{\partial \phi(\mathbf{0})}{\partial y_{k+j}} < 0, k+1 \leq j \leq q \\
\frac{\partial \phi(\mathbf{0})}{\partial y_{k+j}} = ?, q+1 \leq \ell \leq m \\
\frac{\partial \phi(\mathbf{0})}{\partial y_i} = 0, k+m+1 \leq i \leq n \\
\mathbf{0} \neq \mathbf{d} \in T^+ = \left\{ \mathbf{d} \in \mathbb{R}^n : \begin{array}{l} d_i = 0, 1 \leq i \leq k \\ d_{k+j} = 0, 1 \leq j \leq q \end{array} \right\} \Rightarrow \mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} > 0 \\
K = \left\{ \mathbf{d} \in \mathbb{R}^n : \begin{array}{l} d_i = 0, 1 \leq i \leq k \\ d_{k+j} \leq 0, 1 \leq j \leq m \end{array} \right\}
\end{array}$$

因此

$$\begin{array}{l}
K \ni \mathbf{d} = [\underbrace{\quad}_{=0}^k \quad \underbrace{\quad}_{\leq 0}^q \quad \underbrace{\quad}_{\leq 0}^{m-q} \quad \underbrace{\quad}_{???}^{n-m-k}] \\
\nabla \phi(\mathbf{0}) = [\underbrace{\quad}_{???}^k \quad \underbrace{\quad}_{< 0}^q \quad \underbrace{\quad}_{=0}^{m-q} \quad \underbrace{\quad}_{=0}^{n-m-k}] \\
\hline
\{K \cap [\nabla \phi(\mathbf{0})]^\perp\} \ni \mathbf{d} = [\underbrace{\quad}_{=0}^k \quad \underbrace{\quad}_{=0}^q \quad \underbrace{\quad}_{\leq 0}^{m-q} \quad \underbrace{\quad}_{???}^{n-m-k}] \\
\hline
T_+ \ni \mathbf{d} = [\underbrace{\quad}_{=0}^k \quad \underbrace{\quad}_{=0}^q \quad \underbrace{\quad}_{???}^{m-q} \quad \underbrace{\quad}_{???}^{n-m-k}]
\end{array}$$

发现

$$\mathbf{d} \in K \Rightarrow \mathbf{d}^T \nabla \phi(\mathbf{0}) \geq 0 \quad (6.11)$$

和

$$\mathbf{0} \neq \mathbf{d} \in K \ \& \ \mathbf{d}^T \nabla \phi(\mathbf{0}) = 0 \Rightarrow \mathbf{d} \in T_+ \Rightarrow \mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} > 0. \quad (6.12)$$

目前形势是(6.4)即问题

$$\min_{\mathbf{y} \in K} \phi(\mathbf{y}), \quad (6.13)$$

其中 K 是闭凸锥, ϕ 在原点的邻域内是二次连续可微的, 并且满足(6.11)和(6.12).

下面证明 $\mathbf{0}$ 是(6.13)的局部最优解.

设

$$M = \{\mathbf{d} \in K : \|\mathbf{d}\|_2 = 1\},$$

和

$$M_0 = \{\mathbf{d} \in M : \mathbf{d}^T \nabla \phi(\mathbf{0}) = 0\}.$$

由(6.12)知对于 $\mathbf{d} \in M_0$, $\mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} > 0$. 由于 K 是闭的, M 和 M_0 都是紧集, 从而存在 M_0 的邻域 V 和 $\alpha > 0$ 使得

$$\mathbf{d} \in V \Rightarrow \mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} \geq \alpha.$$

集合 $V_1 = M \setminus V$ 是紧集, 并且当 $\mathbf{d} \in V_1$ 时, $\mathbf{d}^T \nabla \phi(\mathbf{0}) > 0$; 因此, 存在 $\beta > 0$ 使得

$$\mathbf{d} \in V_1 \Rightarrow \mathbf{d}^T \nabla \phi(\mathbf{0}) \geq \beta.$$

从而, 将集合 $M = \{\mathbf{d} \in K : \|\mathbf{d}\|_2 = 1\}$ 剖分成两个子集 $V_0 = V \cap M$ 和 V_1 使得

$$\begin{aligned} \mathbf{d} \in V_0 &\Rightarrow \mathbf{d}^T \nabla \phi(\mathbf{0}) \geq 0, \quad \mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} \geq \alpha > 0 \\ \mathbf{d} \in V_1 &\Rightarrow \mathbf{d}^T \nabla \phi(\mathbf{0}) > \beta > 0. \end{aligned}$$

现在的目标是证明 $\mathbf{0}$ 是 ϕ 在 K 上的局部最小点, 也即

$$\exists r > 0 : \phi(\mathbf{0}) \leq \phi(t\mathbf{d}) \quad \forall (\mathbf{d} \in M, 0 \leq t \leq r).$$

设 $\mathbf{d} \in M, t \geq 0$. 当 $\mathbf{d} \in V_0$ 时, 有

$$\phi(t\mathbf{d}) - \phi(\mathbf{0}) = t\mathbf{d}^T \nabla \phi(\mathbf{0}) + \frac{1}{2}t^2 \mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} - t^2 \underbrace{R(t)}_{\rightarrow 0, t \rightarrow +0} \geq \frac{1}{2}t^2(\alpha - 2R(t)).$$

从而 $\exists r_0 > 0$:

$$\phi(t\mathbf{d}) - \phi(\mathbf{0}) \geq \frac{1}{4}t^2\alpha \geq 0 \quad \forall t \leq r_0.$$

当 $\mathbf{d} \in V_1$ 时, 令 $C = \max\{\frac{1}{2}\mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} : \mathbf{d} \in M\}$, 那么有

$$\phi(t\mathbf{d}) - \phi(\mathbf{0}) \geq t\mathbf{d}^T \nabla \phi(\mathbf{0}) + \frac{1}{2}t^2 \mathbf{d}^T \nabla^2 \phi(\mathbf{0}) \mathbf{d} - t^2 \underbrace{R(t)}_{\rightarrow 0, t \rightarrow +0} \geq \beta t - Ct^2 - t^2 R(t).$$

从而 $\exists r_1 > 0$:

$$\phi(t\mathbf{d}) - \phi(\mathbf{0}) \geq \frac{\beta}{2}t \geq 0 \quad \forall t \leq r_1.$$

因此, 对于所有的 $t \leq \min\{r_0, r_1\}$, $\mathbf{d} \in M$, 有 $\phi(t\mathbf{d}) - \phi(\mathbf{0}) \geq 0$.

6.3 敏感性分析

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) : \begin{array}{l} (g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_m(\mathbf{x})) \leq 0 \\ (h_1(\mathbf{x}), \dots, h_k(\mathbf{x})) = 0 \end{array} \right\} \quad (P)$$

\Downarrow

$$L(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_j \lambda_j g_j(\mathbf{x}) + \sum_i \mu_i h_i(\mathbf{x})$$

定义6.1. 设 \mathbf{x}_* 是(6.1)的可行解, 并且函数 f, g_j, h_i 在 \mathbf{x}_* 的邻域内是 $\ell \geq 2$ 次连续可微的. 称 \mathbf{x}_* 是(6.1)的非退化局部最优解, 如果 \mathbf{x}_* 是正则解(即 \mathbf{x}_* 处积极约束的梯

度线性无关), 在 \mathbf{x}_* 处二阶充分最优性条件成立: $\exists(\boldsymbol{\lambda}^* \geq \mathbf{0}, \boldsymbol{\mu}^*)$ 使得

$$\left. \begin{aligned} \nabla_{\mathbf{x}} L(\mathbf{x}_*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) &= \mathbf{0} \\ \lambda_j^* g_j(\mathbf{x}_*) &= 0, \quad j = 1, \dots, m \\ \mathbf{d}^T \nabla g_j(\mathbf{x}_*) &= 0, \quad \forall (j : \lambda_j^* > 0) \\ \mathbf{d}^T \nabla h_i(\mathbf{x}_*) &= 0, \quad \forall i \\ \mathbf{d} &\neq \mathbf{0}, \end{aligned} \right\} \Rightarrow \mathbf{d}^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}_*; \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \mathbf{d} > 0,$$

并且 \mathbf{x}_* 处所有的积极不等式约束的拉格朗日乘子都是正的:

$$g_j(\mathbf{x}_*) = 0 \Rightarrow \lambda_j^* > 0.$$

定理6.6 (敏感度分析). 设 \mathbf{x}_* 是(6.1)的非退化局部最优解. 将(6.1)嵌入一族参数化问题

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) : \begin{aligned} &g_1(\mathbf{x}) \leq a_1, \dots, g_m(\mathbf{x}) \leq a_m \\ &h_1(\mathbf{x}) = b_1, \dots, h_k(\mathbf{x}) = b_k \end{aligned} \right\} \quad (P[\mathbf{a}, \mathbf{b}])$$

其中 $\mathbf{a} \in \mathbb{R}^m, \mathbf{b} \in \mathbb{R}^k$. 这样(6.1)就是 $(P[0, 0])$. 那么存在 \mathbf{x}_* 的邻域 $V_{\mathbf{x}^*}$, 并在参数 (\mathbf{a}, \mathbf{b}) 空间存在点 $(\mathbf{a}, \mathbf{b}) = (\mathbf{0}, \mathbf{0})$ 的邻域 $V_{(\mathbf{a}, \mathbf{b})}$ 满足:

- (a) $\forall (\mathbf{a}, \mathbf{b}) \in V_{(\mathbf{a}, \mathbf{b})}$, 在 $V_{\mathbf{x}^*}$ 中存在 $(P[\mathbf{a}, \mathbf{b}])$ 的唯一KKT点 $\mathbf{x}_*(\mathbf{a}, \mathbf{b})$, 并且该点是 $(P[\mathbf{a}, \mathbf{b}])$ 的非退化局部最优解; 此外, $\mathbf{x}_*(\mathbf{a}, \mathbf{b})$ 是优化问题

$$\text{Opt}_{\text{loc}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{x}} \left\{ f(\mathbf{x}) : \begin{aligned} &g_1(\mathbf{x}) \leq a_1, \dots, g_m(\mathbf{x}) \leq a_m \\ &h_1(\mathbf{x}) = b_1, \dots, h_k(\mathbf{x}) = b_k \\ &\mathbf{x} \in V_{\mathbf{x}^*} \end{aligned} \right\} \quad (P_{\text{loc}}[\mathbf{a}, \mathbf{b}])$$

的最优解.

- (b) $\mathbf{x}_*(\mathbf{a}, \mathbf{b})$ 和相应的拉格朗日乘子 $\boldsymbol{\lambda}^*(\mathbf{a}, \mathbf{b}), \boldsymbol{\mu}^*(\mathbf{a}, \mathbf{b})$ 都是 $(\mathbf{a}, \mathbf{b}) \in V_{(\mathbf{a}, \mathbf{b})}$ 的 $\ell - 1$ 次连续可微函数, 并且

$$\begin{aligned} \frac{\partial \text{Opt}_{\text{loc}}(\mathbf{a}, \mathbf{b})}{\partial a_j} &= \frac{\partial f(\mathbf{x}_*(\mathbf{a}, \mathbf{b}))}{\partial a_j} = -\lambda_j^*(\mathbf{a}, \mathbf{b}), \\ \frac{\partial \text{Opt}_{\text{loc}}(\mathbf{a}, \mathbf{b})}{\partial b_i} &= \frac{\partial f(\mathbf{x}_*(\mathbf{a}, \mathbf{b}))}{\partial b_i} = -\mu_i^*(\mathbf{a}, \mathbf{b}). \end{aligned}$$

6.4 应用举例

例6.2 (特征值的存在性). 考虑优化问题

$$\text{Opt} = \min_{\mathbf{x} \in \mathbb{R}^n} \{ f(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} : h(\mathbf{x}) = 1 - \mathbf{x}^T \mathbf{x} = 0 \} \quad (6.14)$$

其中 $A = A^T$ 是 $n \times n$ 矩阵. 问题显然是可解的. 设 x_* 是它的最优解, 如何刻画 x_* ?

首先断言 x_* 是(6.14)的正则解. 为此应该证明 x_* 处积极约束的梯度线性无关. 而这里只有一个约束, 它在可行集上的梯度非零. 由于 x_* 是正则的全局(因此也是局部)最优解, 因此在 x_* 处必要的二阶最优性条件应该成立: $\exists \mu^*$ 使得

$$\begin{aligned} \nabla_x \left[\overbrace{x^T A x + \mu^*(1 - x^T x)}^{L(x; \mu^*)} \right] &= 0 \Leftrightarrow 2(A - \mu^* I)x_* = 0 \\ \underbrace{d^T \nabla_x h(x_*) = 0}_{\Leftrightarrow d^T x_* = 0} &\Rightarrow \underbrace{d^T \nabla_x^2 L(x_*; \mu^*) d}_{\Leftrightarrow d^T (A - \mu^* I) d \geq 0} \geq 0 \end{aligned}$$

这样, 如果 x_* 是最优解, 那么 $\exists \mu^*$:

$$Ax_* = \mu^* x_* \quad (6.15)$$

$$d^T x_* = 0 \Rightarrow d^T (A - \mu^* I) d \geq 0 \quad (6.16)$$

(6.15)说明 $x_* \neq 0$ 是 A 的特征值 μ^* 对应的特征向量; 特别地, 由线性代数知识, 任何对称矩阵总有实特征向量. (6.16)与(6.15)一起说明对于所有的 y , $y^T (A - \mu^* I) y \geq 0$. 事实上, 每个 $y \in \mathbb{R}^n$ 可以表述为 $y = tx_* + d$, 其中 $d^T x_* = 0$. 现在有

$$\begin{aligned} y^T [A - \mu^* I] y &= (tx_* + d)^T [A - \mu^* I] (tx_* + d) \\ &= t^2 x_*^T \underbrace{[A - \mu^* I] x_*}_{=0} + 2td^T \underbrace{[A - \mu^* I] x_*}_{=0} + \underbrace{d^T [A - \mu^* I] d}_{\geq 0} \geq 0 \end{aligned}$$

请注意在所讨论的情况下, 二阶必要的最优性条件可等价地写作: $\exists \mu^*$ 使得

$$\begin{aligned} [A - \mu^* I] x_* &= 0 \\ y^T [A - \mu^* I] y &\geq 0 \quad \forall y. \end{aligned} \quad (6.17)$$

下面说明对于“可行解 x_* 是全局最优的”而言, 它不仅是必要条件, 也是充分条件.

为了证明充分性, 设 x_* 是可行的, μ^* 使得(6.17)成立. 对于每个可行解 x , 有

$$0 \leq x^T [A - \mu^* I] x = x^T A x - \mu^* x^T x = x^T A x - \mu^*,$$

因此, $x^T A x \geq \mu^*$. 对于 $x = x_*$, 有

$$0 = x_*^T [A - \mu^* I] x_* = x_*^T A x_* - \mu^* x_*^T x_* = x_*^T A x_* - \mu^*,$$

因此 $x_*^T A x_* = \mu^*$. 这样, x_* 是(6.14)的全局最优解, μ^* 是(6.14)的最优值.

例6.3 (扩展S-引理). 设 A, B 是对称矩阵, 并且设 B 使得

$$\exists \bar{x} : \bar{x}^T B \bar{x} > 0. \quad (6.18)$$

那么不等式

$$\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0 \quad (6.19)$$

是不等式

$$\mathbf{x}^T \mathbf{B} \mathbf{x} \geq 0 \quad (6.20)$$

的结果当且仅当(6.19)是(6.20)的“线性结果”：存在 $\lambda \geq 0$ 使得

$$\mathbf{x}^T [\mathbf{A} - \lambda \mathbf{B}] \mathbf{x} \geq 0 \quad \forall \mathbf{x}, \quad (6.21)$$

即(6.19)是(6.20)的加权和(权重 $\lambda \geq 0$)，也等同于(6.21)成立。

证明概要： 唯一非平凡的叙述“如果(6.19) 是(6.20)的结果, 那么存在 $\lambda \geq 0$ 使得...” 为了证明该叙述, 假设(6.19)是(6.20)的结果。

形势是

$$\exists \bar{\mathbf{x}} : \bar{\mathbf{x}}^T \mathbf{B} \bar{\mathbf{x}} > 0; \underbrace{\mathbf{x}^T \mathbf{B} \mathbf{x} \geq 0}_{(6.20)} \Rightarrow \underbrace{\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0}_{(6.19)}$$

考虑优化问题

$$\text{Opt} = \min_{\mathbf{x}} \{ \mathbf{x}^T \mathbf{A} \mathbf{x} : h(\mathbf{x}) \equiv 1 - \mathbf{x}^T \mathbf{B} \mathbf{x} = 0 \}.$$

根据(6.18), 该问题是可行的, 且 $\text{Opt} \geq 0$. 假设存在最优解 \mathbf{x}_* , 那么和上面的讨论一样, \mathbf{x}_* 是正则的, 并且在 \mathbf{x}_* 处二阶必要条件成立: $\exists \mu^*$ 使得

$$\begin{aligned} \nabla_{\mathbf{x}}|_{\mathbf{x}=\mathbf{x}_*} [\mathbf{x}^T \mathbf{A} \mathbf{x} + \mu^* [1 - \mathbf{x}^T \mathbf{B} \mathbf{x}]] &= 0 \Leftrightarrow [\mathbf{A} - \mu^* \mathbf{B}] \mathbf{x}_* = 0 \\ \underbrace{\mathbf{d}^T \nabla_{\mathbf{x}}|_{\mathbf{x}=\mathbf{x}_*} h(\mathbf{x}) = 0}_{\Leftrightarrow \mathbf{d}^T \mathbf{B} \mathbf{x}_* = 0} &\Rightarrow \mathbf{d}^T [\mathbf{A} - \mu^* \mathbf{B}] \mathbf{d} \geq 0 \end{aligned}$$

由上面的第一个等式有 $0 = \mathbf{x}_*^T [\mathbf{A} - \mu^* \mathbf{B}] \mathbf{x}_*$, 再由 \mathbf{x}_* 的可行性得 $\mu_* = \text{Opt} \geq 0$. 重新将 $\mathbf{y} \in \mathbb{R}^n$ 表述为 $t\mathbf{x}_* + \mathbf{d}$, $\mathbf{d}^T \mathbf{B} \mathbf{x}_* = 0$ (即 $t = \mathbf{x}_*^T \mathbf{B} \mathbf{y}$), 得到

$$\mathbf{y}^T [\mathbf{A} - \mu^* \mathbf{B}] \mathbf{y} = t^2 \underbrace{\mathbf{x}_*^T [\mathbf{A} - \mu^* \mathbf{B}] \mathbf{x}_*}_{=0} + 2t \underbrace{\mathbf{d}^T [\mathbf{A} - \mu^* \mathbf{B}] \mathbf{x}_*}_{=0} + \underbrace{\mathbf{d}^T [\mathbf{A} - \mu^* \mathbf{B}] \mathbf{d}}_{\geq 0} \geq 0,$$

因此, $\mu^* \geq 0$, 并且对于所有的 \mathbf{y} , 有 $\mathbf{y}^T [\mathbf{A} - \mu^* \mathbf{B}] \mathbf{y} \geq 0$ 成立. \square

7 最优化算法简介

目标是求解决数学规划问题

$$\min_x \left\{ \begin{array}{l} f(x) : \\ g_j(x) \leq 0 \quad j = 1, \dots, m \\ h_i(x) = 0 \quad i = 1, \dots, k \end{array} \right\} \quad (\text{P})$$

的近似数值解. 本课程考虑的传统MP算法假设事先不知道问题(P)的解析结构(和不知道如何使用已知的结构). 这些算法是面向黑盒的: 即求解(P)时, 算法生成了一个迭代序列 x_1, x_2, \dots , 使得 x_{t+1} 仅取决于所收集的问题(P)沿着以前迭代 x_1, \dots, x_t 的局部信息.

在 x_t 处获得的关于问题(P)的信息通常包括目标和约束在 x_t 的值、一阶和二阶导数.

7.1 算法概述

7.1.1 草垛中找针有多困难呢?

在某些情况下, 面向黑匣子算法的局部信息确实很差, 因此逼近问题的全局解就变成在多维草垛中找针. 如图7.1.1所示, 让我们来看一个边长为2m的3D草垛, 设针是一个高20mm, 横截面半径是1mm的圆柱体. 如何在草垛里找到针呢?

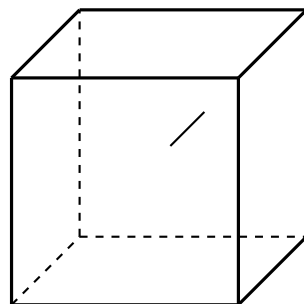


图 7.8: 草垛和针

优化设置是我们想要最小化在“针外部”恒取零, 而在其内部取负值的光滑函数 f . 请注意, 如果仅提供有关该函数的局部信息, 那么除非获得的迭代序列触碰到了针, 否则我们将获得无关紧要的信息. 因此, 易于说明以合理置信度碰到针所需的迭代次数不会比随机产生迭代时小得多. 在这种情况下, 一个迭代碰到针的概率小到 $7.8 \cdot 10^{-9}$, 也就是说, 要想有一个合理的置信度, 我们需要产生数亿次迭代. 随着问题维数的增长, 困难将剧烈地放大. 例如, 保留草垛和针的线性大小, 并将草垛的维数从3增加到20, 迭代碰到针的概率就小

到 $8.9 \cdot 10^{-67}$! 在“草垛中的针”问题中, 易于找到局部最优解. 然而, 稍微修改一下问题, 就可能给后一项任务带来灾难性的困难.

在无约束最小化问题中, 找到使目标函数梯度变得很小的点, 即找到“几乎”满足一阶必要最优条件的点并不太难. 在约束最小化问题中, 即使仅找到一个可行解也可能很难. 然而连续优化的经典算法虽然在最坏情况下没有提供有意义的保证, 但能够相当有效地处理应用中出现的典型优化问题.

请注意在优化中, 确实存在利用问题结构, 并允许在合理时间内逼近全局解的算法. 这种类型的传统方法——单纯形法及其变体——不会超出线性规划和线性约束凸二次规划的范围. 在1990年代, 人们发现在一种有效利用问题结构的新方法(内点法), 但是所得到的算法并没有超出凸规划的范围.

除了非常特殊和相对简单的问题类别(如线性规划或线性约束二次规划)外, 优化算法无法保证在有限时间内找到精确的局部或全局解. 对这些算法, 我们期望得到的最好结果是: 算法产生的近似解收敛到精确解.

即使在确实存在“有限”求解方法的情况下(线性规划中的单纯形法), 也无法确定这些方法的合理复杂度的界, 因此在现实中, 一种方法在有限步得到精确解的能力既不必要, 也不足以解释该方法. 除了凸规划之外, 传统优化方法无法保证收敛到全局最优解. 的确, 在非凸情况下, 无法通过局部信息来判断给定点是否是全局最优的:

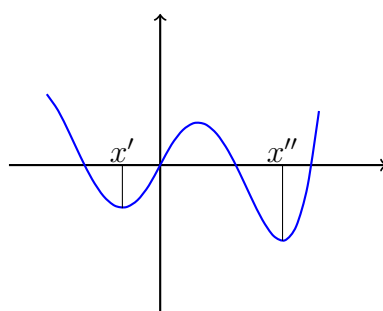


图 7.9: 局部解与全局解

以图7.1.1为例, 在 x' 附近“观察”问题, 我们一定找不到真正的全局最优解是 x'' 的线索.

为了保证近似全局解, 似乎不可避免地要“扫描” x 的一组密集值, 以确保不会错过全局最优解. 从理论上讲, 存在这种可能性. 但是, “穷举搜索”方法的复杂性随决策向量的维数呈指数级增长, 从而使得这些方法完全不切实际. 传统的优化方法不包含穷举搜索, 因此, 不能保证收敛到全局解. 将传统优化方法应用于一般问题(不一定是凸问题)的典型理论结果听起来像是:

假定问题(P)具有以下性质:

• • •

那么方法 X 生成的近似解序列是有界的, 并且序列的所有极限点都是问题的KKT点.

或者

假设 x_* 是问题(P)的非退化局部解. 那么方法 X 从足够接近 x_* 的位置开始迭代, 产生的序列收敛到 x_* .

7.1.2 MP算法的分类与收敛速度

MP算法有两种主要的传统分类. 按应用领域主要分作无约束优化算法和约束优化算法. 按算法使用的信息主要分作仅使用目标和约束函数值的零阶方法、同时使用目标和约束的函数值和它们的一阶导数的一阶方法、使用目标和约束的函数值及一阶和二阶导数的二阶方法.

有必要量化MP算法的收敛性质. 传统上, 这是通过如下定义的渐近收敛速率来完成的:

Step 1. 我们引入一个(根据正在求解的问题而定义的)合适的误差度量——近似解的非负函数 $\text{Error}_P(x)$, 此函数在我们想要逼近的问题(P)的解集 X_* 处恰好是零.

例7.1. (i) 到集合 X_* 的距离:

$$\text{Error}_P(x) = \inf_{x_* \in X_*} \|x - x_*\|_2$$

(ii) 用目标和约束的残差

$$\text{Error}_P(x) = \max \{f(x) - \text{Opt}(P), [g_1(x)]_+, \dots, [g_m(x)]_+, |h_1(x)|, \dots, |h_k(x)|\}$$

Step 2. 假设我们已经建立了方法的收敛性, 也就是说, 我们知道如果 x_t^* 是由方法应用于给定问题族(P)产生的近似解, 那么有

$$\text{Error}_P(t) \equiv \text{Error}_P(x_t^*) \rightarrow 0, t \rightarrow \infty.$$

接着我们粗略地量化非负实数序列 $\text{Error}_P(t)$ 收敛到0的速度.

特别地, 我们称方法是次线性(sublinearly)收敛的, 如果误差序列不会比几何级数更快地收敛到零. 比如 $1/t$ 或 $1/t^2$; 称方法是线性(linearly)收敛的, 如果存在 $C < \infty$ 和 $q \in (0, 1)$ 使得

$$\text{Error}_P(t) \leq Cq^t,$$

称 q 是收敛比(convergence ratio), 比如

$$\text{Error}_P(t) \asymp e^{-at}$$

呈现收敛比是 e^{-a} 的线性收敛. 速度是 $q \in (0, 1)$ 的线性收敛方法的充分条件是

$$\lim_{t \rightarrow \infty} \frac{\text{Error}_P(t+1)}{\text{Error}_P(t)} < q.$$

称方法是超线性(superlinearly)收敛的, 如果误差列比几何级数更快地收敛到零:

$$\forall q \in (0, 1), \exists C : \text{Error}_P(t) \leq Cq^t.$$

比如

$$\text{Error}_P(t) \asymp e^{-at^2}$$

对应的是超线性收敛. 方法是超线性收敛的充分条件是

$$\lim_{t \rightarrow \infty} \frac{\text{Error}_P(t+1)}{\text{Error}_P(t)} = 0.$$

方法是 $p > 1$ 阶收敛的, 如果

$$\exists C : \text{Error}_P(t+1) \leq C(\text{Error}_P(t))^p.$$

称收敛阶为 2 的方法是二次的(quadratic). 比如

$$\text{Error}_P(t) \asymp e^{-ap^t}$$

p 阶收敛到0.

非正式解释: 当方法收敛时, 随着 $t \rightarrow 0$, $\text{Error}_P(t)$ 收敛到0, 即最终 $\text{Error}_P(t)$ 的十进制表示形式中, 小数点之前为零, 小数点之后的零越来越多; 称对应近似解中小数点后零的个数是**有效数字(accuracy digit)**. 传统收敛速度分类是以渐近地为现有近似解增加一位新有效数字需要多少步为基础.

对于次线性收敛, 有效数字的“价格”随数字的位置而增加. 例如, 收敛速度是 $O(1/t)$ 的情况, 就步数而言, 每位新有效数字要比前一位贵10倍. 线性收敛, 每位有效数字价格相同, 与

$$\frac{1}{\ln(\frac{1}{\text{convergence ratio}})}$$

成正比. 等价地, 方法每步都添加固定 r (对于不太接近0的 q , $r \approx 1 - q$)位有效数字.

对于超线性收敛, 每位后续有效数字最终都将变得比其的前一位者更便宜——随着有效数字的增加, 价格将变为0. 等价地, 每个额外步会增加越来越多的有效数字.

对于 $p > 1$ 阶收敛, 有效数字的价格不仅随着数字位置 k 的增加而变为0, 而且如几何级数般足够快. 等价地, 最终方法每个额外步将使有效数字变成 p 倍的.

对于传统方法, 方法的收敛性质越好, 那么该方法在上述分类中的等级越高. 给定一族问题, 传统上认为每个问题是线性收敛的方法比次线性收敛的更快, 超线性收敛的方法比线性收敛的更快, 等等.

请注意通常我们能够证明存在参数 C 和 q 量化线性收敛:

$$\text{Error}_P(t) \leq Cq^t$$

或者 $p > 1$ 阶收敛

$$\text{Error}_P(t+1) \leq C(\text{Error}_P(t))^p$$

但无法找到这些参数的数值——它们可能取决于我们正在求解的特定问题的“不可观察”特征. 因此, 传统的收敛性质“量化”是定性的和渐进的.

7.1.3 可解的MP——凸规划

我们已经看到, 当应用于一般的MP问题时, 优化方法具有许多严格的理论限制, 包括的主要限制如下: (a) 除非使用(在高维优化中完全不现实的)穷尽搜索, 否则无法保证接近全局解. (b) 收敛性质的量化具有渐近和定性的特点. 因此最自然的问题类似:

We should solve problems of such and such structure with such and such sizes and the data varying in such and such ranges. How many steps of method X are sufficient to solve problems within such and such accuracy?

通常没有理论上有效的答案.

尽管存在理论上的局限性, 实际上, 传统MP算法能解决许多(即使不是全部)现实世界中的MP问题, 这包括具有成千上万个变量和约束的MP问题.

此外, 存在“可解情况”——实际效率拥有坚实理论保证——凸规划的情况. 这是一个典型的“凸规划”结论: 假设我们正在求解一个凸规划问题

$$\text{Opt} = \min_x \{f(x) : g_j(x) \leq 0, j \leq m, |x_i| \leq 1, i \leq n\},$$

其中目标和约束是规范化的, 即满足

$$|x_i| \leq 1, i \leq n \Rightarrow |f(x)| \leq 1, |g_j(x)| \leq 1, j \leq m.$$

给定 $\epsilon \in (0, 1)$, 可找到问题的 ϵ 解 x^ϵ :

$$\underbrace{|x_i^\epsilon| \leq 1}_{\forall i \leq n} \& \underbrace{g_j(x^\epsilon) \leq \epsilon}_{\forall j \leq m} \& f(x^\epsilon) - \text{Opt} < \epsilon$$

至多需要 $2n^2 \ln(\frac{2n}{\epsilon})$ 步, 并且每步只需计算一次 f, g_1, \dots, g_m 在一点上的值及一阶导数和额外的 $100(m+n)n$ 次算术运算.

7.2 线搜索

“线搜索”是用于一维“简单约束”优化的技术之通称. 特别地, 对于问题

$$\min_x \{f(x) : a \leq x \leq b\}, \quad (\text{LS})$$

其中 $[a, b]$ 是坐标轴上的给定线段(有时, 我们将允许 $b = +\infty$), f 是在 (a, b) 上至少一次连续可微, 在线段 $[a, b]$ (在射线 $[a, \infty)$ 上, 当 $b = \infty$) 上连续的函数. 许多针对多维优化的算法将使用线搜索作为子程序.

7.2.1 零阶线搜索与黄金搜索

在零阶线搜索中,人们利用问题(LS)中的目标 f 的值,而不使用其导数.为了确保问题是适定的,假设目标函数是单峰的(unimodal),即在 $[a, b]$ 上具有唯一的局部极小值点 x_* .等价于存在唯一的点 $x_* \in [a, b]$ 使得 $f(x)$ 在 $[a, x_*]$ 上是严格递减的,在 $[x_*, b]$ 上是严格递增的.

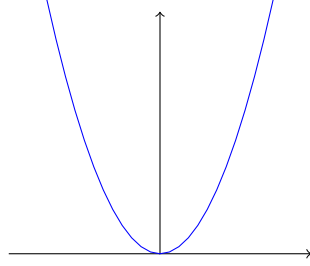


图 7.10: 单峰函数

主要观察如下: 设 f 是 $[a, b]$ 上的单峰函数, 假设对于某 x', x'' , $a < x' < x'' < b$, 我们知道 $f(x')$ 和 $f(x'')$. 如果 $f(x'') \geq f(x')$, 那么 $f(x) > f(x'')$, $x > x''$, 因此最小值属于 $[a, x'']$. 类似的, 如果 $f(x'') < f(x')$, 那么 $f(x) > f(x')$, $x < x'$, 因此最小值属于 $[x', b]$.

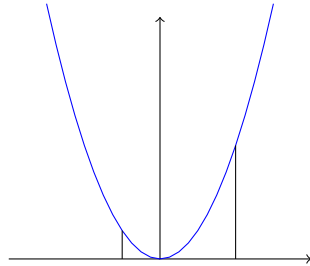


图 7.11: 对于单峰函数的主要观察

在这两种情况下, 在 x', x'' 处两次计算 f 的值能将初始“搜索域”缩小到较小的区间($[a, x'']$ 或 $[x', b]$). 选择 x', x'' 将 $[a_0, b_0] = [a, b]$ 三个相等的段, 计算 $f(x'), f(x'')$ 并互相比大小, 我们得到新的段 $[a_1, b_1] \subset [a_0, b_0]$ 使得(a)新的段是定位器——它包含解 x_* . (b) 新定位器的长度是初始定位器 $[a_0, b_0] = [a, b]$ 长度的 $2/3$ 倍.

与在原始定位器上一样, 目标函数在新定位器上是单峰的, 并且我们可以迭代我们的构造. 在第 $N \geq 1$ 步(计算 $2N$ 次 f), 我们将定位器的长度缩小 $(2/3)^N$ 倍, 即根据讨论, 我们得到收敛比 $q = \sqrt{2/3} = 0.8165 \dots$ 的线性收敛算法.

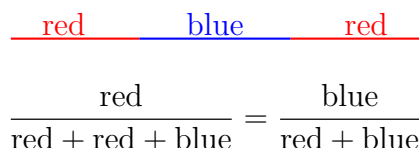
我们可以做得更好吗？答案是肯定的.

$$\left. \begin{array}{l} \text{null } [a_{t-1}, b_{t-1}] \\ x'_t < x''_t \end{array} \right\} \Rightarrow f(x'_t), f(x''_t) \Rightarrow \begin{cases} \text{null } [a_t, b_t] = [a_{t-1}, x''_t] \\ [a_t, b_t] = [x'_t, b_{t-1}] \end{cases}$$

观察到我们每一步计算 f 的两个点之一会成为新定位器的端点，而另一个是新定位器的内点，因此我们可以将这个内点在下一步用作计算 f 值的两个点中的一个！使用这种方法，只有最开始的步需要计算2次函数值计算，而后续步则只需计算1次！我们让所有搜索点都以固定比例 $x' - a = b - x'' = \theta(b - a)$ 分划自定位器的方式来实现这一想法：方程

$$\theta \equiv \frac{x' - a}{b - a} = \frac{x'' - x'}{b - x'} \equiv \frac{1 - 2\theta}{1 - \theta} \Rightarrow \theta = \frac{3 - \sqrt{5}}{2}$$

给出了这个比例.



$$\frac{\text{red}}{\text{red} + \text{red} + \text{blue}} = \frac{\text{blue}}{\text{red} + \text{blue}}$$

图 7.12: 黄金搜索

至此，我们已得到了黄金搜索，其根据

$$\frac{x' - a}{b - a} = \frac{b - x''}{b - a} = \frac{3 - \sqrt{5}}{2}$$

在当前定位器 $[a_{t-1}, b_{t-1}]$ 中确定第 t 步的搜索点 x_{t-1}, x_t . 在此方法中，一步将误差(定位器的长度)减小了 $1 - \frac{3-\sqrt{5}}{2} = \frac{\sqrt{5}-1}{2}$ 倍. 因此收敛比约等于

$$\frac{\sqrt{5} - 1}{2} \approx 0.6180 \dots$$

7.2.2 一阶线搜索之二分法

假设 f 在 (a, b) 上是可微的且严格单峰的，也就是说 f 是单峰的，并且当 $x_* \in (a, b), a < x < x_*$ 时， $f'(x) < 0$ ；当 $x_* < x < b$ 时 $f'(x) > 0$.

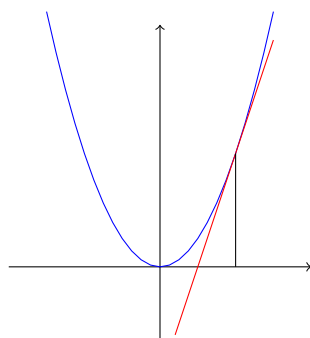


图 7.13: 二分法

设 f 和 f' 都可得, 此时我们选择二分法. **主要观察:** 给定 $x_1 \in [a, b] \equiv [a_0, b_0]$, 设我们计算了 $f'(x_1)$. 如果 $f'(x_1) > 0$, 那么由严格单峰性质, 在 x_1 的右侧有 $f(x) > f(x_1)$. 因此, $x_* \in [a, x_1]$; 类似的, 如果 $f'(x_1) \leq 0$, 那么对 $x < x_1$ 有 $f(x) > f(x_1)$. 因此, $x_* \in [x_1, b]$. 在这两种情况, 我们可以用更小的定位器 $[a_1, b_1]$ 代替原始定位器 $[a, b] = [a_0, b_0]$, 然后重复这个过程.

在二分法中, 第 t 步选 $[a_{t-1}, b_{t-1}]$ 的中点作为 x_t , 然后计算 $f'(x_t)$, 因此每步操作都会将定位器的长度减少 2 倍. 显然, 根据讨论, 二分法以收敛比 0.5 线性收敛:

$$a_t - x_* \leq 2^{-t}(b_0 - a_0).$$

7.2.3 非精确线搜索之回溯Armijo线搜索算法

多维最小化的许多算法使用“线搜索”作为子程序, 方式如下: 给定当前迭代 $x_t \in \mathbb{R}^n$, 算法定义一个搜索方向 $d_t \in \mathbb{R}^n$, 要求它是 f 的下降方向:

$$d_t^T \nabla f(x_t) < 0.$$

然后援引线搜索在 $\gamma \geq 0$ 上极小化一元函数

$$\phi(\gamma) = f(x_t + \gamma d_t);$$

所得到的 $\gamma = \gamma^t$ 确定了沿方向 d_t 的步长, 因此外部算法的新迭代是

$$x_{t+1} = x_t + \gamma^t d_t.$$

在这种情况的许多形势下, 没必要精确地最小化 γ . 只要 ϕ 有“本质的”减少就足够了. Armijo规则给出了“本质的”减少之定义(及获取)的标准方式: 设 $\phi(\gamma)$ 在 $\gamma \geq 0$ 上是连续可微函数, 且满足 $\phi'(0) < 0$. 设 $\rho \in (0, 1)$, $\eta > 1$ 是参数(通常选 $\rho = 0.2$, $\eta = 2$ 或者 $\eta = 10$).

我们称步长 $\gamma > 0$ 是合适的, 如果

$$\phi(\gamma) \leq \phi(0) + \rho\gamma\phi'(0), \quad (7.1)$$

称 γ 是几乎最大的, 如果 η 倍的步长不是合适的

$$\phi(\eta\gamma) > \phi(0) + \rho\eta\gamma\phi'(0). \quad (7.2)$$

称步长 $\gamma > 0$ 通过Armijo测试(“本质的”减少 ϕ), 如果它既是合适的又是几乎最大的.

重要的事实是假设在射线 $\gamma > 0$ 上 ϕ 是有下界的. 那么通过Armijo 规则的步长肯定存在, 并且能被有效地找出. Armijo-可接受步长 $\gamma > 0$ 满足(7.1)和(7.2).

找Armijo-可接受步长的算法:

开始: 选择 $\gamma_0 > 0$, 并检查其是否满足(7.1). 如果满足, 转分支 A, 否则转分支 B.

分支 A: γ_0 满足(7.1). 依次测试值为 $\eta\gamma_0, \eta^2\gamma_0, \eta^3\gamma_0, \dots$ 的 γ , 在当前值首次不满足(7.1)时, 那么 γ 的前一个值通过Armijo 测试.

分支 B: γ_0 不满足(7.1). 依次测试值为 $\eta^{-1}\gamma_0, \eta^{-2}\gamma_0, \eta^{-3}\gamma_0, \dots$ 的 γ , 当当前值满足(7.1)时, 那么这个 γ 值通过Armijo测试.

算法验证: 显然, 如果算法终止, 那么结果确实通过Armijo测试, 因此我们需要证实的是算法最终终止. 分支 A 明显是有限的: 这里我们沿着序列 $\gamma_i = \eta^i\gamma_0 \rightarrow \infty$ 验证不等式(7.1). 当不等式首次满足时终止计算. 由于 $\phi'(0) < 0$ 且 ϕ 有下界, 那么上述情况一定会发生. 分支 B 明显是有限的: 这里我们沿着序列 $\gamma_i = \eta^{-i}\gamma_0 \rightarrow +0$ 验证不等式(7.1), 并且当不等式首次满足时终止计算. 由于 $\rho \in (0, 1)$ 且 $\phi'(0) < 0$, 故

$$\phi(\gamma) = \phi(0) + \gamma[\phi'(0) + \underbrace{R(\gamma)}_{\rightarrow 0, \gamma \rightarrow +0}]$$

从而不等式(7.1)对于所有足够小的正值 γ 都是满足的. 因为当 i 充分大时, γ_i 一定会变得“足够小”. 因此, 分支 B 也是有限的.

8 无约束最小化方法：梯度下降与牛顿法

无约束极小化问题是

$$f_* = \min_x f(x), \quad (\text{UC})$$

其中 f 是适定的并且在整个 \mathbb{R}^n 上是连续可微的. 请注意提出的大多数构造都能够直接推广到“本质上无约束的情况”—— f 在 \mathbb{R}^n 中的开集定义域 D 上是连续可微的并且使得水平集 $\{x \in D : f(x) \leq a\}$ 是闭的.

8.1 梯度下降

梯度下降是用于无约束最小化的最简单的一阶方法. 它的思想如下. 设 x 是当前迭代, 并且不是 f 的临界点: $f'(x) \neq 0$. 有

$$f(x + th) = f(x) + th^T f'(x) + t\|h\|_2 R_x(th) \quad [\text{当 } s \rightarrow 0, R_x(s) \rightarrow 0].$$

由于 $f'(x) \neq 0$, 那么单位负梯度方向 $g = -f'(x)/\|f'(x)\|_2$ 是 f 的下降方向:

$$\frac{d}{dt}\bigg|_{t=0} f(x + tg) = g^T f'(x) = -\|f'(x)\|_2.$$

因此沿着方向 g 的移动 $x \mapsto x + tg$ 使 f 以“速度” $\|f'(x)\|_2$ 局部地减小. 请注意就局部下降速度而言, g 是最佳的下降方向: 对于任何其它单位方向 h , 有

$$\frac{d}{dt}\bigg|_{t=0} f(x + th) = h^T f'(x) > -\|f'(x)\|_2.$$

在一般的梯度下降中, 用从 x_{t-1} 出发, 沿着可以使目标函数减小的负梯度方向的步更新当前迭代 x_t :

$$x_t = x_{t-1} - \gamma_t f'(x_{t-1})$$

其中 γ_t 是满足

$$f'(x_{t-1}) \neq 0 \Rightarrow f(x_t) < f(x_{t-1})$$

的正步长.

有两种**标准实现**.

最速下降GD: $\gamma_t = \operatorname{argmin}_{\gamma \geq 0} f(x_{t-1} - \gamma f'(x_{t-1}))$. 除了 f 是二次的情况外, 这种步长选取方式略微理想化.

Armijo GD: 已知参数 $\rho \in (0, 1)$ 和 $\eta > 1$. 找 $\gamma_t > 0$ 使得

$$f(x_{t-1} - \gamma_t f'(x_{t-1})) \leq f(x_{t-1}) - \rho \gamma_t \|f'(x_{t-1})\|_2^2, \quad (8.1)$$

$$f(x_{t-1} - \eta \gamma_t f'(x_{t-1})) > f(x_{t-1}) - \rho \eta \gamma_t \|f'(x_{t-1})\|_2^2. \quad (8.2)$$

倘若 $f'(x_{t-1}) \neq 0$ 并且当 $\gamma \geq 0$ 时 $f(x_{t-1} - \gamma f'(x_{t-1}))$ 有下界, 这种确定步长的方法是可实现的.

请注意, 由构造可知, 当 $f'(x_{t-1}) = 0$ 时有 $x_t = x_{t-1}$. 所以GD不能逃离临界点.

8.1.1 收敛性

定理8.1 (全局收敛). 假设与初始点 x_0 关联的 f 之水平集

$$G = \{x : f(x) \leq f(x_0)\}$$

是紧的, 并且 f 在 G 的某邻域内是连续可微的. 那么对于SGD和AGD来说:

- (i) 方法开始于从 x_0 出发的轨迹 x_0, x_1, \dots 是适定的并且永不离开 G (因此是有界的);
- (ii) 方法是单调的: 除非方法得到临界点 x_t , 因此有 $x_t = x_{t+1} = x_{t+2} = \dots$; 否则 $f(x_0) \geq f(x_1) \geq \dots$, 并且不等式是严格成立的.
- (iii) 轨迹的每个极限点均是 f 的临界点.

证明概要: 1⁰. 如果 $f'(x_0) = 0$, 方法永远停留在 x_0 并且陈述是显然的. 现在假设 $f'(x_0) \neq 0$. 那么函数

$$\phi_0(\gamma) = f(x_0 - \gamma f'(x_0))$$

有下界, 并且集合 $\{\gamma \geq 0 : \phi_0(\gamma) \leq \phi_0(0)\}$ 随同集合 G 一起是紧集, 因此 $\phi_0(\gamma)$ 在射线 $\gamma \geq 0$ 上取到其最小值并且 $\phi'_0(0) < 0$. 由此可见 GD 的第一步是适定的, 并且 $f(x_1) < f(x_0)$. 集合 $\{x : f(x) \leq f(x_1)\}$ 是 G 的闭子集, 因此是紧的. 我们让 x_1 作为 x_0 重复我们的推理, 我们断定轨迹是适定的, 从没有离开 G , 并且除非得到临界点, 目标函数是严格减小的.

2⁰. “轨迹的所有极限点均是 f 的临界点”:

事实: 设 $x \in G$ 并且 $f'(x) \neq 0$. 那么存在 $\epsilon > 0$ 和 x 的邻域 U 使得对于每个 $x' \in U$, 方法从 x' 开始的步 $x' \rightarrow x'_+$ 使 f 至少减小 ϵ . 鉴于此事实, 设 x 是 $\{x_i\}$ 的极限点; 假设 $f'(x) \neq 0$, 下面由此导出矛盾. 由该事实, 存在 x 的邻域 U 使得

$$x_i \in U \Rightarrow f(x_{i+1}) \leq f(x_i) - \epsilon.$$

由于轨迹无限多次访问 U 并且该方法是单调的, 因此我们断定 $f(x_i) \rightarrow -\infty, i \rightarrow \infty$. 然而由于 G 是紧的, 因此 f 在 G 上有下界. 从而这是不可能的. \square

梯度下降法的极限点. 在全局收敛定理的假设下, GD的极限点存在, 并且它们都是 f 的临界点. 那它们的极限点都有哪些类型呢? 首先 f 的非退化极大值点不可能是GD的极限点, 除非方法由这个极大值点开始. 其次, f 的鞍点“极不可能”成为候选极限点. 实践经验表明, 极限点是 f 的局部极小值点. 最后, f 的非退化全局极小点 x_* (如果有的话), 如同GD的“吸引点”, 当初始点距离该极小点足够近时, 方法收敛到 x_* .

8.1.2 收敛速度

一般而言，我们仅能保证收敛到 f 的临界点集合. 与这个集合相关联的自然误差度量是

$$\delta^2(x) = \|f'(x)\|_2^2.$$

定义8.1. 设 U 是 \mathbb{R}^n 中的开集， $L \geq 0$ 且 f 是定义在 U 上的函数. 我们说 f 在 U 上是 $C^{1,1}(L)$ 的，如果 f 在 U 上是连续可微的并且具有常数为 L 的局部Lipschitz连续的梯度：

$$[x, y] \in U \Rightarrow \|f'(x) - f'(y)\|_2 \leq L\|x - y\|_2.$$

我们说 f 在集合 $Q \subset \mathbb{R}^n$ 是 $C^{1,1}(L)$ 的，如果存在开集 $U \supset Q$ 使得 f 在 U 上是 $C^{1,1}(L)$ 的.

请注意，假设 f 在 U 上是二阶连续可微的，那么 f 在 U 上是 $C^{1,1}(L)$ 的当且仅当 f 的海森矩阵的范数不超过 L ：

$$\forall (x \in U, d \in \mathbb{R}^n) : |d^T f''(x) d| \leq L\|d\|_2^2.$$

定理8.2 (收敛速度). 除过全局收敛定理的假设之外，假设 f 在 $G = \{x : f(x) \leq f(x_0)\}$ 上是 $C^{1,1}(L)$ 的. 那么对于SGD，可以得到

$$\min_{0 \leq k \leq t} \delta^2(x_k) \leq \frac{2[f(x_0) - f_*]L}{t+1}, t = 0, 1, 2, \dots$$

对于AGD，可以得到

$$\min_{0 \leq k \leq t} \delta^2(x_k) \leq \frac{\eta}{2\rho(1-\rho)} \cdot \frac{[f(x_0) - f_*]L}{t+1}, t = 0, 1, 2, \dots$$

先给出一个重要引理，由它可得推出定理的结论.

引理8.3. 对 $x \in G$ ， $0 \leq s \leq 2/L$ ，可以得到

$$x - sf'(x) \in G, \quad (8.3)$$

$$f(x - sf'(x)) \leq f(x) - \delta^2(x)s + \frac{L\delta^2(x)}{2}s^2. \quad (8.4)$$

证明 当 $g \equiv -f'(x) = 0$ 时不需要证明. 设 $g \neq 0$ ， $s_* = \max\{s \geq 0 : x + sg \in G\}$ ， $\delta^2 = \delta^2(x) = g^T g$ ，函数

$$\phi(s) = f(x - sf'(x)) : [0, s_*] \rightarrow \mathbb{R}$$

是连续可微的并且满足 (a) $\phi'(0) = -g^T g \equiv -\delta^2$; (b) $\phi(s_*) = f(x_0)$; (c) $|\phi'(s) - \phi'(0)| = |g^T [f'(x + sg) - f'(x)]| \leq Ls\delta^2$. 因此

$$\phi(s) \leq \phi(0) - \delta^2 s + \frac{L\delta^2}{2}s^2 \quad (8.5)$$

此即(8.4). 的确, 置

$$\theta(s) = \phi(s) - \left[\phi(0) - \delta^2 s + \frac{L\delta^2}{2} s^2 \right]$$

可以得到

$$\theta(0) = 0, \theta'(s) = \phi'(s) - \phi'(0) - Ls\delta^2 \underset{\text{由(c)}}{\leq} 0.$$

将 $s = s^*$ 代入(8.5), 并由(b)式, 可以得到

$$f(x_0) \overset{\text{由(b)}}{\leq} \phi(0) - \delta^2 s_* + \frac{L\delta^2}{2} s_*^2 \leq f(x_0) - \delta^2 s_* + \frac{L\delta^2}{2} s_*^2 \Rightarrow s_* \geq 2/L.$$

□

引理 \Rightarrow 定理: 先考虑**SGD**. 由引理, 可以得到

$$\begin{aligned} f(x_t) - f(x_{t+1}) &= f(x_t) - \min_{\gamma \geq 0} f(x_t - \gamma f'(x_t)) \\ &\geq f(x_t) - \min_{0 \leq s \leq 2/L} [f(x_t) - \delta^2(x_t)s + \frac{L\delta^2(x_t)}{2} s^2] = \frac{\delta^2(x_t)}{2L} \end{aligned}$$

由此得

$$f(x_0) - f_* \geq \sum_{k=0}^t [f(x_k) - f(x_{k+1})] \geq \sum_{k=0}^t \frac{\delta^2(x_k)}{2L} \geq \frac{t+1}{2L} \min_{0 \leq k \leq t} \delta^2(x_k).$$

从而

$$\min_{0 \leq k \leq t} \delta^2(x_k) \leq \frac{2L(f(x_0) - f_*)}{t+1}.$$

现在考虑**AGD**. 可以断言 $\gamma_{t+1} > \frac{2(1-\rho)}{L\eta}$. 的确, 如果断言不成立, 有

$$\begin{aligned} f(x_t - \gamma_{t+1}\eta f'(x_t)) &\leq f(x_t) - \gamma_{t+1}\eta\delta^2(x_t) + \frac{L\delta^2(x_t)}{2}\eta^2\gamma_{t+1}^2 \\ &= f(x_t) - \underbrace{\left[1 - \frac{L}{2}\eta\gamma_{t+1}\right]}_{\geq \rho} \eta\gamma_{t+1}\delta^2(x_t) \\ &\leq f(x_t) - \rho\eta\gamma_{t+1}\delta^2(x_t) \end{aligned}$$

由于 γ_{t+1} 满足(8.2), 所以这是不可能的. 我们已经证明 $\gamma_{t+1} > \frac{2(1-\rho)}{L\eta}$. 由Armijo法则(8.1), 得

$$f(x_t) - f(x_{t+1}) \geq \rho\gamma_{t+1}\delta^2(x_t) \geq \frac{2\rho(1-\rho)}{L\eta}\delta^2(x_t).$$

其余证明与SGD相同.

□

8.1.3 凸的情况

除了全局收敛定理的假设外, 假设 f 是凸的. 因为凸函数的所有临界点是它的全局极小点, 所以凸的情况下, 当 $t \rightarrow \infty$ 时, SGD和AGD收敛到 f 的全局极小值点集, 当 $t \rightarrow \infty, f(x_t) \rightarrow f_*$, 并且轨迹的所有极限点均是 f 的全局极小点. 在凸 $C^{1,1}(L)$ 的情况下, 可用残差 $f(x_t) - f_*$ 量化全局收敛速度.

定理8.4. 设 $G = \{x : f(x) \leq f(x_0)\}$ 是紧凸集, f 在 G 上是凸的和 $C^{1,1}(L)$ 的. 考虑AGD, 并且设 $\rho \geq 0.5$, 那么方法产生的轨迹收敛到 f 的全局极小点 x_* , 并且

$$f(x_t) - f_* \leq \frac{\eta L \|x_0 - x_*\|_2^2}{4(1-\rho)t}, t = 1, 2, \dots$$

定义8.2. 设 M 是 \mathbb{R}^n 上的凸集并且 $0 < \ell \leq L < \infty$. 称函数 f 是 M 上以参数 ℓ, L 是强凸的, 如果 f 在 M 上是 $C^{1,1}(L)$ 的, 并且对于 $x, y \in M$, 可以得到

$$[x - y]^T [f'(x) - f'(y)] \geq \ell \|x - y\|_2^2 \quad (8.6)$$

称比值 $Q_f = L/\ell$ 是 f 的条件数.

关于强凸函数的解释: 如果 f 在凸集 M 上是 $C^{1,1}(L)$ 的, 那么

$$x, y \in M \Rightarrow |f(y) - [f(x) + (y - x)^T f'(x)]| \leq \frac{L}{2} \|x - y\|_2^2.$$

如果 f 在凸集 M 上满足(8.6), 那么

$$\forall x, y \in M : f(y) \geq f(x) + (y - x)^T f'(x) + \frac{\ell}{2} \|y - x\|_2^2.$$

特别地, f 在 M 上是凸的. 综上, f 在凸集 M 上以参数 ℓ, L 是强凸的, 那么 $\forall x, y \in M$, 有

$$f(x) + (y - x)^T f'(x) + \frac{\ell}{2} \|y - x\|_2^2 \leq f(y) \leq f(x) + (y - x)^T f'(x) + \frac{L}{2} \|y - x\|_2^2.$$

请注意, 假设 f 在凸集 M 的某个邻域上是二次连续可微的. 那么 f 在 M 上是 (ℓ, L) -强凸的当且仅当对于所有的 $x \in M$ 和所有的 $d \in \mathbb{R}^n$ 可以得到

$$\ell \|d\|_2^2 \leq d^T f''(x) d \leq L \|d\|_2^2,$$

这等价于

$$\lambda_{\min}(f''(x)) \geq \ell, \lambda_{\max}(f''(x)) \leq L.$$

特别地, 当 A 是正定矩阵时, 二次函数

$$f(x) = \frac{1}{2} x^T A x - b^T x + c \quad (\text{QF})$$

在整个空间以参数 $\ell = \lambda_{\min}(A), L = \lambda_{\max}(A)$ 强凸的.

定理8.5. 在强凸的情况下，AGD呈现线性全局收敛速度. 特别地，设集合 $G = \{x : f(x) \leq f(x_0)\}$ 是闭的和凸的，并且 f 在集合 G 上是参数 ℓ, L 强凸的，那么 G 是紧的，并且 f 的全局极小点 x_* 存在且唯一； $\rho \geq 1/2$ 的AGD 线性收敛到 x_* ：

$$\|x_t - x_*\|_2 \leq \theta^t \|x_0 - x_*\|_2,$$

其中

$$\theta = \sqrt{\frac{Q_f - (2 - \rho^{-1})(1 - \varepsilon)\eta^{-1}}{Q_f + (\rho^{-1} - 1)\eta^{-1}}} = 1 - O(Q_f^{-1}).$$

除此之外，

$$f(x_t) - f_* \leq \theta^{2t} Q_f [f(x_0) - f_*].$$

强凸二次情况下的SGD. 假设(QF)中的 f 是强凸二次函数，即 $A = A^T \succ 0$. 在这种情况下，SGD 变成可实现的，由循环

$$\begin{aligned} g_t &= f'(x_t) = Ax_t - b \\ \text{null } \gamma_{t+1} &= \frac{g_t^T g_t}{g_t^T A g_t} \\ x_{t+1} &= x_t - \gamma_{t+1} g_t \end{aligned}$$

给出，并且可以保证

$$\underbrace{f(x_{t+1}) - f_*}_{E_{t+1}} \leq \left[1 - \frac{(g_t^T g_t)^2}{[g_t^T A g_t][g_t^T A^{-1} g_t]} \right] E_t \leq \left(\frac{Q_f - 1}{Q_f + 1} \right)^2 E_t$$

随之

$$f(x_t) - f_* \leq \left(\frac{Q_f - 1}{Q_f + 1} \right)^{2t} [f(x_0) - f_*], t = 1, 2, \dots$$

请注意，如果我们知道SGD收敛到 f 的一个非退化局部极小点 x_* . 那么在温和的规律性假设下，该方法的渐近行为将好像 f 是强凸二次型

$$f(x) = \text{const} + \frac{1}{2}(x - x_*)^T f''(x_*)(x - x_*)$$

一样. 图8.14给出了 f 在一个非退化局部极小点附近的迭代轨迹.

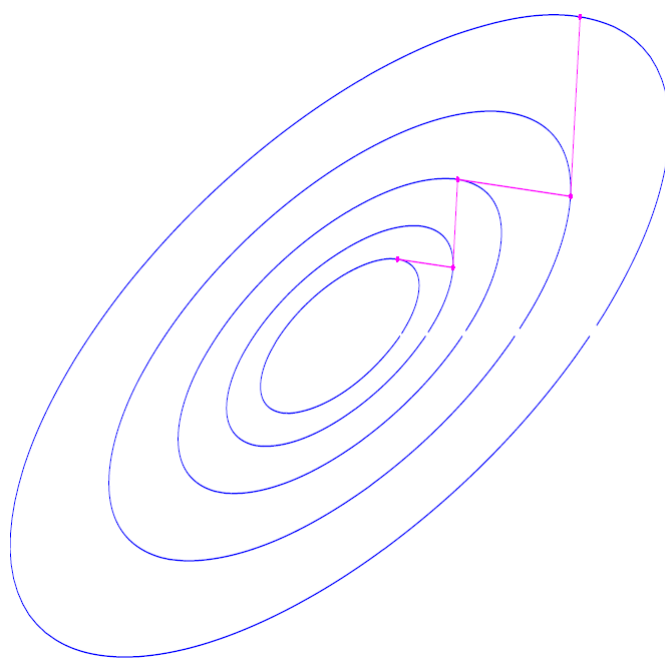


图 8.14: GD在非退化局部极小点附近的迭代轨迹

将SGD应用于 $Q_f = 1000$ 的二次形, 其中 $f(x_0) = 2069.4$, $f(x_{999}) = 0.0232$, 所得结果见图8.15.

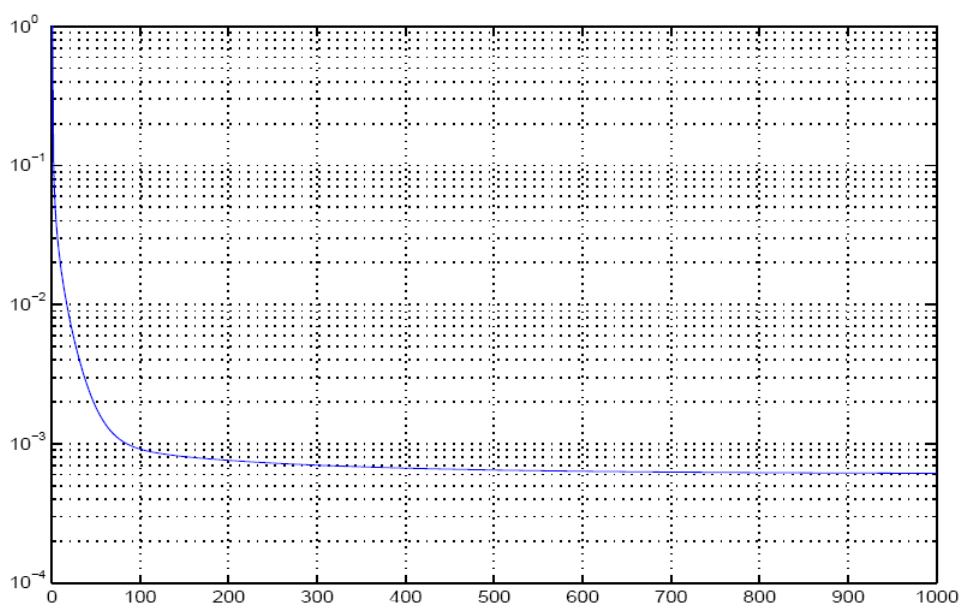


图 8.15: $\frac{f(x_t) - f_*}{(f(x_0) - f_*) \left(\frac{Q_f - 1}{Q_f + 1}\right)^{2t}}$ 的图

8.1.4 小结

在温和的正则性和有界性假设下, SGD和AGD都收敛到目标函数的临界点集合. 在目标函数是 $C^{1,1}(L)$ -光滑的情况下, 就误差度量 $\delta^2(x) = \|f'(x)\|_2^2$ 而言,

方法呈现非渐进的 $O(1/t)$ 收敛速度.

在相同的正则性假设下, 凸情况下方法收敛到目标的全局极小点集合. 在 $C^{1,1}(L)$ -凸的情况下, 就目标残差 $f(x) - f_*$ 而言, AGD呈现非渐进的 $O(1/t)$ 的收敛速度.

在强凸情况下, 就目标残差 $f(x) - f_*$ 和距离 $\|x - x_*\|_2$ 而言, AGD都呈现出非渐进的线性收敛. 收敛比是 $1 - O(1/Q_f)$, 其中 Q_f 是目标的条件数. 换句话说, 为了获得额外的有效数字, 需要 $O(Q_f)$ 步.

GD的优点是简单、并在关于待极小化函数的温和假设下具有可接受的全局收敛性质. GD的缺点是 " 标架依赖性 " 方法不是仿射不变的! 你正在从 $x_0 = 0$ 开始, 用GD求解问题 $\min_x f(x)$. 你的第一个搜索点将是

$$x_1 = -\gamma_1 f'(0).$$

我求解同样的问题, 并且新变量 $y : x = Ay$. 我的问题是从 $y_0 = 0$ 开始, $\min_y g(y) := f(Ay)$. 我的第一个搜索点是

$$y_1 = -\hat{\gamma}_1 g'(0) = -\hat{\gamma}_1 A^T f'(0)$$

用 x 变量表示, 我的搜索点是

$$\hat{x}_1 = Ay_1 = -\hat{\gamma}_1 AA^T f'(0).$$

如果 AA^T 与单位矩阵不成比例, 我的搜索点通常会与你的搜索点不同!

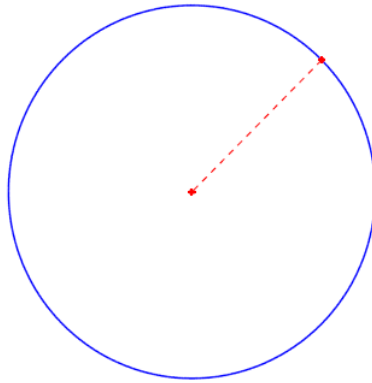


图 8.16: 将SGD应用于 $f(x) = \frac{1}{2}x^T x$

例8.1. 将 $x_1 = y_1, x_2 = y_2/3$ 代入 $f(x) = \frac{1}{2}x^T x$, 问题变成了

$$\min_y g(y) = \frac{1}{2} \left[y_1^2 + \frac{1}{9}y_2^2 \right].$$

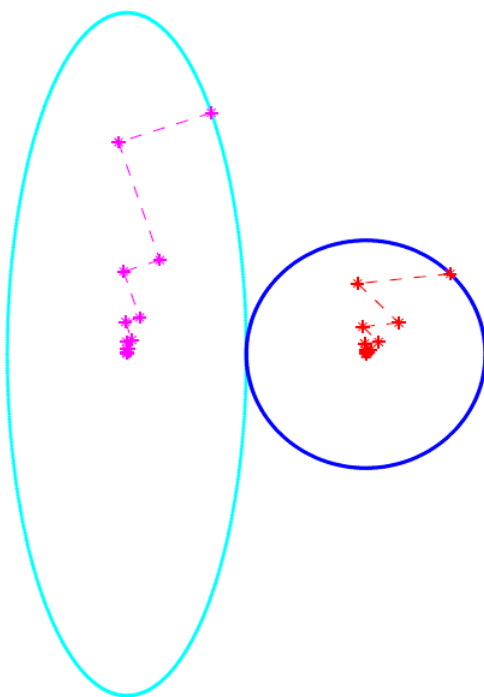


图 8.17: 左侧: 将SGD应用于 g 得到的轨迹; 右侧: x -坐标系中的同一轨迹.

表 8.1: 将SGD应用于 g 得到的函数值/目标残差序列

t	1	3	5	7	9
$g(y_t)$	0.5000	0.0761	0.0116	0.0018	0.0003

“框架依赖性”是几乎所有一阶优化方法的共同缺点，这就是为什么即使对于最有利的强凸目标，它们的收敛速度也对问题的条件数很敏感. GD对条件数“超敏感”：当最小化强凸函数 f 时，GD的收敛比是 $1 - O(1/Q_f)$ ，然而对于更好的方法，它的收敛比是 $1 - O(1/Q_f^{1/2})$.

8.2 牛顿法

考虑目标函数是二次连续可微的无约束问题(UC). 假设二阶信息是可得的，我们在当前迭代 x 的附近利用二阶泰勒展式

$$f(y) \approx f(x) + (y - x)^T f'(x) + \frac{(y - x)^T f''(x)(y - x)}{2}$$

近似 f . 在牛顿法中，新的迭代是二次近似的极小点. 如果存在，极小点由下式给出：

$$\begin{aligned} \nabla_y [f(x) + (y - x)^T f'(x) + \frac{(y - x)^T f''(x)(y - x)}{2}] &= 0 \\ \Leftrightarrow f''(x)(y - x) &= -f'(x) \\ \Leftrightarrow y &= x - [f''(x)]^{-1} f'(x) \end{aligned}$$

我们已经得到了基本牛顿法

$$x_{t+1} = x_t - [f''(x_t)]^{-1} f'(x_t). \quad (\text{Nwt})$$

当矩阵 $f''(x_t)$ 是奇异的时, 步 t 是无定义的.

现在考虑另一种动机. 假设我们寻找费马方程

$$f'(x) = 0$$

的解. 给定解的当前近似 x_t , 我们在 x_t 的附近线性化方程的左侧, 因此得到线性化的费马方程

$$f'(x_t) + f''(x_t)[x - x_t] = 0$$

并把这个方程的解, 即 $x_t - [f''(x_t)]^{-1} f'(x_t)$, 作为我们的新迭代.

定理8.6 (局部二次收敛). 设 x_* 是 f 的非退化局部极小点, 因此 $f''(x_*) \succ 0$, 并且设 f 在 x_* 的一个邻域内是三次连续可微的. 那么从足够接近 x_* 的地方开始的循环(Nwt)是适定的, 并且二次收敛到 x_* .

证明 设 U 是以 x_* 为中心的球, 并且 f 的三阶导数在 U 上有界 β_1 . 对于 $y \in U$, 可以得到

$$\begin{aligned} \|\nabla f(y) + \nabla^2 f(y)(x_* - y)\|_2 &\equiv \|\nabla f(y) + \nabla^2 f(y)(x_* - y) - \nabla f(x_*)\|_2 \\ &\leq \beta_1 \|y - x_*\|_2^2. \end{aligned} \quad (8.7)$$

因为 $f''(x)$ 在 $x = x_*$ 处是连续的并且 $f''(x_*)$ 是非奇异的, 那么存在以 x_* 为中心的球 $U' \subset U$ 使得

$$y \in U' \Rightarrow \|[f''(y)]^{-1}\| \leq \beta_2 \quad (8.8)$$

目前的形势是存在 $r > 0$ 和正常数 β_1, β_2 使得当 $\|y - x_*\| < r$ 时, 有(8.7)和(8.8)成立. 设方法的迭代 x_t 接近 x_* :

$$x_t \in V = \left\{ x : \|x - x_*\|_2 \leq \rho \equiv \min \left\{ \frac{1}{2\beta_1\beta_2}, r \right\} \right\}$$

可以得到

$$\begin{aligned} \|x_{t+1} - x_*\| &= \|x_t - x_* - [f''(x_t)]^{-1} f'(x_t)\|_2 \\ \text{null} \quad &= \|[f''(x_t)]^{-1} [-f''(x_t)(x_* - x_t) - f'(x_t)]\|_2 \\ &\leq \beta_1 \beta_2 \|x_t - x_*\|_2^2 \leq 0.5 \|x_t - x_*\|_2. \end{aligned}$$

可以断定方法在第 t 步后仍是适定的, 并且二次收敛到 x_* . □

牛顿法令人瞩目的性质是仿射不变性(“标架独立性”). 设 $x = Ay + b$ 是变量的可逆仿射变换. 那么

$$\begin{aligned} f(x) &\Leftrightarrow g(y) := f(Ay + b) \\ \text{null} \\ \bar{x} &= A\bar{y} + b \Leftrightarrow \bar{y} \end{aligned}$$

在 y -变量空间的牛顿迭代

$$\begin{aligned} \bar{y}_+ &= \bar{y} - [g''(\bar{y})]^{-1}g'(\bar{y}) \\ &= \bar{y} - [A^T f''(\bar{x})A]^{-1}[A^T f'(\bar{x})] \\ &= \bar{y} - A^{-1}[f''(\bar{x})]^{-1}f'(\bar{x}), \end{aligned}$$

与 x -变量空间对应的点恰好等于 x -变量空间的牛顿迭代:

$$A\bar{y}_+ + b = [A\bar{y} + b] - [f''(\bar{x})]^{-1}f'(\bar{x}) = \bar{x} - [f''(\bar{x})]^{-1}f'(\bar{x}).$$

基本牛顿法的困境. 从足够接近 f 的非退化局部极小点 x_* 的开始基本牛顿法(Nwt)二次收敛到 x_* . 然而, 即使对于优好的强凸函数 f , 如果初始点不是太接近 f (唯一的) 的局部 \equiv 全局极小点, 方法可能发散:

$$f(x) = \sqrt{1+x^2} \Rightarrow x_{t+1} = -x_t^3.$$

从而当 $|x_0| < 1$, 方法二次(甚至 3 阶)收敛到 $x_* = 0$; 当 $|x_0| > 1$, 方法迅速发散... 当 f 不是强凸的时, 牛顿方向

$$-[f''(x)]^{-1}f'(x)$$

可能是没有定义的或者不是 f 的下降方向. 由于这些缺点, 需要修正基本牛顿法以确保全局收敛. 修正包括: 结合线搜索、在牛顿方向是无定义或者其不是 f 的下降方向时修正牛顿方向.

结合线搜索. 假设水平集 $G = \{x : f(x) \leq f(x_0)\}$ 是闭的和凸的, 并且 f 在 G 上是强凸的. 那么对于 $x \in G$, 牛顿方向

$$e(x) = -[f''(x)]^{-1}f'(x)$$

是 f 的下降方向, 除过 x 是 f 的临界点的情况(或者, 这与强凸情况中的全局极小点相同):

$$f'(x) \neq 0 \Rightarrow e^T(x)f'(x) = -[f'(x)]^T \underbrace{[f''(x)]^{-1}}_{\succ 0} f'(x) < 0.$$

在牛顿法的线搜索版本中, 使用 $e(x)$ 作为搜索方向而不是位移:

$$x_{t+1} = x_t + \gamma_{t+1}e(x_t) = x_t - \gamma_{t+1}[f''(x_t)]^{-1}f'(x_t),$$

其中 $\gamma_{t+1} > 0$ 是通过沿牛顿方向精确地最小化 f 或通过 Armijo 线搜索确定的步长.

定理8.7. 设水平集 $G = \{x : f(x) \leq f(x_0)\}$ 是凸的和紧的, 并且 f 在 G 上是强凸的. 那么由精确步长与满足Armijo法则的步长对应的线搜索牛顿法收敛到 f 的唯一全局极小值点. 在恰当的线搜索实现下, 收敛是二次的.

牛顿法小结. **优点**是二次渐近收敛的, 但前提是我们设法使轨迹接近非退化局部极小点. **缺点一**是需要计算和求海森矩阵逆矩阵使得计算成本相对较高. **缺点二**是在非强凸情况下, 牛顿方向可能是不适定的, 或者不是目标的下降方向, 因此必须“矫正/治愈”该方法.

9 无约束最小化方法：牛顿法的修正

牛顿法的修正旨在避免它(非凸目标的困难和计算成本相对高)的缺点的同时保留它的主要优点——快的渐近收敛. 有四个主要的修正组：三次正则化牛顿法、基于二阶信息的修正牛顿法、共轭梯度法和拟牛顿法这两个基于一阶信息的修正.

9.1 三次正则化牛顿法

关注的问题是

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

其中 $X \subset \mathbb{R}^n$ 是内部非空的闭凸集, f 在 X 上是三次连续可微的. 假设所给初始点 $x_0 \in \text{int}X$ 使得集合

$$X_0 = \{x \in X : f(x) \leq f(x_0)\}$$

是有界的并且包含在 X 的内部.

先描述三次正则化牛顿法的思想. 为了理解方法, 考虑 $X = \mathbb{R}^n$ 且 f 的三阶导数在 X 上是有界的, 因此 f 在任意点沿任何单位方向的三阶方向导数不超过 $L \in (0, \infty)$. 此时可以得到

$$\forall x, h : f(x+h) \leq \bar{f}_x(h),$$

其中

$$\bar{f}_x(h) = f(x) + h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x) h + \frac{L}{6} \|h\|^3$$

请注意, 对于较小的 h , $\bar{f}_x(h)$ 作为 $f(x+h)$ 的近似基本上和 f 在 x 处的二阶泰勒一样好, 还兼具了对所有 h , $\bar{f}_x(h)$ 是 $f(x_h)$ 上界的优点. 从而当从 x 到 $x^+ = x + h_*$, 其中

$$h_* \in \underset{h}{\operatorname{argmin}} \bar{f}_x(h),$$

我们确保有

$$f(x^+) \leq \bar{f}_x(h_*) \leq \bar{f}_x(0) = f(x),$$

除非 $h_* = 0$ 是 $\bar{f}_x(\cdot)$ 的全局极小值点, 上述不等式是严格成立的. 后一种情况发生当且仅当 x 满足无约束光滑优化问题的二阶必要条件:

$$\nabla f(x) = 0, \nabla^2 f(x) \succeq 0.$$

假设任给的初始点 x_0 使得集合 $X_0 = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ 是紧的. 除此之外, 存在闭凸集 X 使得 $X_0 \subset \text{int}X$, 并且 f 在 X 上是三次连续可微的. 三次正则化的通用牛顿法可归为:

在步 t , 给定前一个迭代 x_{t-1} , 记

$$\bar{f}(h) = f(x_{t-1}) + h^T \nabla f(x_{t-1}) + \frac{1}{2} h^T \nabla^2 f(x_{t-1}) h + \frac{L_t}{6} \|h\|^3.$$

我们选择 $L_t > 0$ 是良好的——使得位移

$$h_t \in \operatorname{argmin}_h \bar{f}(h)$$

满足 $f(x_{t-1} + h_t) \leq \bar{f}(h_t)$, 并且置 $x_t = x_{t-1} + h_t$.

一个重要的**事实**是只要 $x_{t-1} \in X_0$, 所有足够大的 L_t , 特别地, 那些

$$L_t \geq M_X(f) = \max_{x \in X, h \in \mathbb{R}^n: \|h\| \leq 1} \left. \frac{d^3}{dt^3} \right|_{t=0} f(x + th)$$

是良好的. 算法是适定的且确保 $f(x_0) \geq f(x_1) \geq \dots$, 除过算法在到达满足二阶必要条件 $\nabla f(x) = 0, \nabla^2 f(x) \succeq 0$ 的点 x 外, 所有的不等式严格成立. 算法在满足二阶必要条件的点处陷入困境.

通过线搜索可以轻松保证 L_t 的有界性和良好性: 给定 x_{t-1} 和 L_{t-1} (例如 $L_{t-1} = 1$), 逐一检查 L_t 的候选值 $L^k = 2^k L_{t-1}$ 是否是良好的. 从 $k = 0$ 开始:

如果 L^0 是好的, 尝试 L^{-1}, L^{-2}, \dots 直到良好性消失, 或者达到较小的阈值(比如 10^{-6}), 并使用 L_t 的最后一个良好的候选值 L^k 作为 L_t 的实际值.

如果 L^0 是不好的, 尝试 L^1, L^2, \dots 直到复现良好性, 由 L_t 的第一个良好的候选值 L^k 作为 L_t 的实际值. 这一举措确保 $L_t \leq 2 \max[M_X(f), L_{-1}]$. 通过保持 L_t 有界的举措, 该算法确保

- (i) 轨迹的所有极限点(确实存在——轨迹属于有界集合 X_0)满足无约束最小化的二阶必要最优性条件;
- (ii) 只要轨迹的极限点是 f 的非退化局部极小值点, 轨迹就会二次收敛于该极小值点.

算法的实现步需要求解无约束极小化问题

$$\min_h [p^T h + h^T P h + c \|h\|^3], \quad [P = P^T, c > 0] \quad (\text{CRS})$$

计算特征值分解 $P = U \operatorname{Diag}\{\beta\} U^T$, 并且从变量 h 变化到 $g = U^T h$, 问题变成

$$\min_g \left\{ q^T g + \sum_i \beta_i g_i^2 + c \left(\sum_i g_i^2 \right)^{\frac{3}{2}} \right\} \quad [q = U^T p]$$

在最优的情况下, $\operatorname{sign}(g_i) = -\operatorname{sign}(q_i)$, 从而问题化简成

$$\min_g \left\{ -\sum_i |q_i| |g_i| + \sum_i \beta_i g_i^2 + c \left(\sum_i g_i^2 \right)^{\frac{3}{2}} \right\}$$

进行变量替换, 令 $s_i = g_i^2$, 问题变成凸的

$$\min_{s \geq 0} \left\{ - \sum_i |q_i| \sqrt{s_i} + \sum_i \beta_i s_i + c \left(\sum_i s_i \right)^{\frac{3}{2}} \right\}. \quad (9.1)$$

由(9.1)的最优解 s^* 导出(CRS)的最优解 h^* :

$$h^* = U g^*, g_t^* = -\text{sign}(q_i) \sqrt{s_i^*}.$$

最简单的求解(9.1)方法是把(9.1)重新写成

$$\min_{s, r} \left\{ \sum_i [\beta_i s_i - |q_i| \sqrt{s_i}] + c r^{\frac{3}{2}} : s \geq 0, \sum_i s_i \leq r \right\}$$

并转到它的拉格朗日对偶

$$\max_{\lambda \geq 0} \left\{ L(\lambda) := \min_{s \geq 0, r \geq 0} \left[c r^{\frac{3}{2}} - \lambda r + \sum_i [(\beta_i + \lambda) s_i - |q_i| \sqrt{s_i}] \right] \right\} \quad (9.2)$$

$L(\cdot)$ 易于计算, 所以可以利用二分法求解(9.2). 假设 $|q_i| > 0$ (通过对 q_i 的微扰来实现), 可以由对偶问题的最优解 λ_* 得到(9.1)的最优解

$$(s_*, r_*) \in \operatorname{argmin}_{s \geq 0, r \geq 0} \left[c r^{\frac{3}{2}} - \lambda_* r + \sum_i [(\beta_i + \lambda_*) s_i - |q_i| \sqrt{s_i}] \right].$$

9.1.1 传统修正: 变度量方案(scheme)

牛顿法的所有传统修正都运用了自然的变度量思想. 在谈论GD 时, 已经提到使用非奇异矩阵 B 的方法

$$x_{t+1} = x_t - \gamma_{t+1} \underbrace{BB^T}_{A^{-1} \succ 0} f'(x_t) \quad (9.3)$$

和梯度下降

$$x_{t+1} = x_t - \gamma_{t+1} f'(x_t)$$

具有相同的“存在权”. 前一种方法只不过是应用于

$$g(y) = f(By).$$

等价地, 设 A 是正定对称矩阵. 我们有完全相同的理由用

$$\frac{d^T f'(x)}{\sqrt{d^T d}} \quad (9.4)$$

来度量 f 的“局部方向下降速度”, 就像用

$$\frac{d^T f'(x)}{\sqrt{d^T A d}} \quad (9.5)$$

一样.

当根据(9.4)选择最速下降方向作为当前搜索方向时, 我们得到负梯度方向 $-f'(x)$, 并且获得GD. 当根据(9.5)选择最速下降方向作为当前搜索方向时, 我们得到“比例负梯度方向” $-A^{-1}f'(x)$, 并且获得“比例”GD:

$$x_{t+1} = x_t - \gamma_{t+1} A^{-1} f'(x_t) \quad (9.6)$$

我们已经激发出缩放GD(9.6). 为什么不在进行迭代步之前考虑采用随步变化的“比例矩阵” $A_{t+1} \succ 0$ 的一般变度量算法

$$x_{t+1} = x_t - \gamma_{t+1} A_{t+1}^{-1} f'(x_t) \quad (\text{VM})$$

来领先一步呢?

请注意 当 $A_{t+1} \equiv I$, (VM)就变成普通梯度下降法; 当 f 是强凸的并且 $A_{t+1} = f''(x_t)$ 时, (VM)变成一般牛顿法. 并且, 当 x_t 不是 f 的临界点时, 搜索方向 $d_{t+1} = -A_{t+1}^{-1}f'(x_t)$ 是 f 的下降方向:

$$d_{t+1}^T f'(x_t) = -[f'(x_t)]^T A_{t+1}^{-1} f'(x_t) < 0.$$

这样, 我们对单调线搜索版本(VM)的理解没有概念上的困难. 事实表明, 变度量方法具有良好的全局收敛性质.

定理9.1. 设水平集 $G = \{x : f(x) \leq f(x_0)\}$ 是有界闭的, 且设 f 在 G 的邻域是二次连续可微的. 进一步假设更新矩阵 A_t 的策略确保其的一致正定性和有界性:

$$\exists 0 < \ell \leq L < \infty : \ell I \preceq A_t \preceq LI \quad \forall t.$$

那么对于初始点是 x_0 的精确步长和Armijo版本的(VM), 轨迹是适定的, 属于 G (因此是有界的), 并且除过到达 f 的临界点外, f 沿着轨迹严格减少. 进一步, 轨迹的所有极限点都是 f 的临界点.

通过谱分解的实现: 给定 x_t , 计算 $H_t = f''(x_t)$, 然后求 H_t 的谱分解:

$$H_t = V_t \text{Diag}\{\lambda_1, \dots, \lambda_n\} V_t^T.$$

给定一次选取永远适用的容差 $\delta > 0$, 置

$$\hat{\lambda}_i = \max[\lambda_i, \delta]$$

并且

$$A_{t+1} = V_t \text{Diag}\{\hat{\lambda}_1, \dots, \hat{\lambda}_n\} V_t^T.$$

请注意, 假如水平集 $G = \{x : f(x) \leq f(x_0)\}$ 是紧的, 并且 f 在 G 的邻域是二次连续可微的, 上述构造确保 $\{A_t\}_t$ 满足一致正定性和有界性.

Levenberg-Marquard 实现：选择适当的 $\delta > 0$. 令

$$A_{t+1} = \epsilon_t I + H_t$$

其中选择 $\epsilon_t \geq 0$ 确保 $A_{t+1} \succeq \delta I$. 用二分法求解

$$\min\{\epsilon : \epsilon \geq 0, H_t + \epsilon I \succeq \delta I\}$$

得到 ϵ_t . 对于给定的 ϵ , 二分法需要检查条件

$$H_t + \epsilon I \succ \delta I \Leftrightarrow H_t + (\epsilon - \delta)I \succ 0$$

是否成立. 基础测试源自Choleski分解.

Choleski 分解. 由线性代数可知, 对称矩阵 P 是正定的当且仅当

$$P = DD^T, \quad (\text{CF})$$

其中 D 是下三角矩阵. 当Choleski 分解(CF)存在时, 可以由如下简单算法找到:

表达式(CF)意味着

$$p_{ij} = d_i d_j^T$$

其中

$$d_i = (d_{i1}, d_{i2}, \dots, d_{ii}, 0, 0, 0, 0, \dots, 0)$$

$$d_j = (d_{j1}, d_{j2}, \dots, d_{ji}, \dots, d_{jj}, 0, \dots, 0)$$

是 D 的行. 特别地, $p_{i1} = d_{i1}d_{i1}$, 并且我们可以置 $d_{11} = \sqrt{p_{11}}$, $d_{i1} = p_{i1}/d_{11}$, 由此确定了 D 的第 1 列. 进一步, $p_{22} = d_{21}^2 + d_{22}^2$, 随之 $d_{22} = \sqrt{p_{22} - d_{21}^2}$. 在知道 d_{22} 之后, 可以利用关系

$$p_{i2} = d_{i1}d_{21} + d_{i2}d_{22} \Rightarrow d_{i2} = \frac{p_{i2} - d_{i1}d_{21}}{d_{22}}, i > 2$$

找到 D 的第 2 列中的所有剩余元素. 我们接着以这种方式继续做: 在找到 D 的前 $(k-1)$ 列后, 根据

$$d_{kk} = \sqrt{p_{kk} - d_{k1}^2 - d_{k2}^2 - \dots - d_{k,k-1}^2}$$

$$d_{ik} = \frac{p_{ik} - d_{i1}d_{k1} - \dots - d_{i,k-1}d_{k,k-1}}{d_{kk}}, i > k$$

填充第 k 列. 概括的方法要么得到满足要求的 D , 要么在无法执行当前转轴时, 即当

$$p_{kk} - d_{k1}^2 - d_{k2}^2 - \dots - d_{k,k-1}^2 \leq 0$$

时终止. 此“不合标准的终止”表明 P 不是正定的.

如果Choleski分解存在, 概括的Choleski算法允许在大约 $\frac{n^3}{6}$ 步找到Choleski分解. 通常用Choleski分解求解 P 正定的线性方程组

$$Px = p. \quad (\text{LS})$$

为了求解系统, 首先计算Choleski分解

$$P = DD^T$$

然后通过两个回代求解(LS)

$$b \mapsto y : Dy = b, y \mapsto x : D^T x = y,$$

即通过求解两个三角方程组(仅需 $O(n^2)$ 步).

算法的另一种应用(比如在Levenberg-Marquardt方法中)是检查对称矩阵的正定性.

请注意, 假如集合 $G = \{x : f(x) \leq f(x_0)\}$ 是紧的, 并且 f 在 G 的邻域是二次连续可微的, Levenberg-Marquardt方法产生一致正定有界的矩阵序列 $\{A_t\}$.

修正牛顿法”最实用的”实现是基于对 $H_t = f''(x_t)$ 的Choleski分解. 当在处理当前转轴的过程中(得到 d_{kk})变得不可能或者导致 $d_{kk} < \delta$, 增加 H_t 相应的对角线元素, 直到条件 $d_{kk} = \delta$ 成立.

假如集合 $G = \{x : f(x) \leq f(x_0)\}$ 是紧的, 并且 f 在 G 的邻域是二次连续可微的, 用这个方法得到 H_t 的一个使得矩阵是”良好正定的”对角校正, 且确所得到的序列 $\{A_t\}$ 是一致正定和有界的.

9.1.2 共轭梯度法

最小化二次函数. 考虑正定二次型

$$f(x) = \frac{1}{2}x^T Hx - b^T x + c$$

的最小化问题. 这里是用于最小化 f , 或者与求解方程组

$$Hx = b$$

相同的“概念算法”. 任给初始点 x_0 , 设 $g_0 = f'(x_0) = Hx_0 - b$, 并且设

$$E_k = \text{Lin}\{g_0, Hg_0, H^2g_0, \dots, H^{k-1}g_0\}$$

和

$$x_k = \operatorname{argmin}_{x \in x_0 + E_k} f(x)$$

事实I: 设 k_* 是使得 $E_{k+1} = E_k$ 的最小整数 k . 那么 $k_* \leq n$, 并且 x_{k_*} 是 f 在 \mathbb{R}^n 上的唯一极小值点.

事实 II: 有

$$f(x_k) - \min_x f(x) \leq 4 \left[\frac{\sqrt{Q_f} - 1}{\sqrt{Q_f} + 1} \right]^{2k} [f(x_0) - \min_x f(x)].$$

事实 III: 轨迹 $\{x_k\}$ 由下面的显式循环给出:

初始化: 设

$$d_0 = -g_0 \equiv -f'(x_0) = b - Hx_0;$$

Step t: 如果 $g_{t-1} \equiv \nabla f(x_{t-1}) = 0$, 终止, x_{t-1} 即为所求. 否则

$$\gamma_t = \frac{g_{t-1}^T d_{t-1}}{d_{t-1}^T H d_{t-1}}$$

$$x_t = x_{t-1} + \gamma_t d_{t-1}$$

$$g_t = f'(x_t) \equiv Hx_t - b$$

$$\beta_t = \frac{g_t^T H d_{t-1}}{d_{t-1}^T H d_{t-1}}$$

$$d_t = -g_t + \beta_t d_{t-1}$$

并循环至步 $t+1$.

请注意在上述过程中, 梯度 $g_0, \dots, g_{k_*-1}, g_{k_*} = 0$ 是相互正交的. 方向 $d_0, d_1, \dots, d_{k_*-1}$ 是 H 正交的, 即

$$i \neq j \Rightarrow d_i^T H d_j = 0,$$

并且有

$$\gamma_t = \underset{\gamma}{\operatorname{argmin}} f(x_{t-1} + \gamma d_{t-1})$$

$$\beta_t = \frac{g_t^T g_t}{g_{t-1}^T g_{t-1}}.$$

可将求解强凸二次形 f 的共轭梯度法看作求解线性方程组

$$Hx = b$$

的迭代算法. 与类似像 Choleski 分解或者高斯消元法这种“直接求解器”相比, CG 的优点是:

- (i) 在精确算术运算的情况下, 最多可以在 n 步内找到解, 且每一步一个矩阵向量乘法和 $O(n)$ 个加法运算. 因此求解成本至多是 $O(n)L$, 其中 L 是矩阵向量乘法的算术运算价格.

请注意, 当 H 是稀疏的时, $L \ll n^2$, 并且求解成本变得比直接的线性代数方法的成本 $O(n^3)$ 更少.

- (ii) 原则上, 不必形成 H , 我们所需要的只是能够与 H 相乘.

(iii) 非渐进误差界

$$f(x_k) - \min_x f(x) \leq 4 \left[\frac{\sqrt{Q_f} - 1}{\sqrt{Q_f} + 1} \right]^{2k} [f(x_0) - \min_x f(x)]$$

表明收敛速度仅取决于 H 的条件数，与维数完全无关.

示例：

表 9.2: 1000×1000 的方程组, $Q_f = 1.e2$

Iter	$f - f_*$	$\ x - x_*\ $
1	$2.297e + 003$	$2.353e + 001$
11	$1.707e + 001$	$4.265e + 000$
21	$3.624e - 001$	$6.167e - 001$
31	$6.319e - 003$	$8.028e - 002$
41	$1.150e - 004$	$1.076e - 002$
51	$2.016e - 006$	$1.434e - 003$
61	$3.178e - 008$	$1.776e - 004$
71	$5.946e - 010$	$2.468e - 005$
81	$9.668e - 012$	$3.096e - 006$
91	$1.692e - 013$	$4.028e - 007$
94	$4.507e - 014$	$2.062e - 007$

表 9.3: 1000×1000 的方程组, $Q_f = 1.e4$

Itr	$f - f_*$	$\ x - x_*\ $
1	$1.471e + 005$	$2.850e + 001$
51	$1.542e + 002$	$1.048e + 001$
101	$1.924e + 001$	$4.344e + 000$
151	$2.267e + 000$	$1.477e + 000$
201	$2.248e - 001$	$4.658e - 001$
251	$2.874e - 002$	$1.779e - 001$
301	$3.480e - 003$	$6.103e - 002$
351	$4.154e - 004$	$2.054e - 002$
401	$4.785e - 005$	$6.846e - 003$
451	$4.863e - 006$	$2.136e - 003$
501	$4.537e - 007$	$6.413e - 004$
551	$4.776e - 008$	$2.109e - 004$
601	$4.954e - 009$	$7.105e - 005$
651	$5.666e - 010$	$2.420e - 005$
701	$6.208e - 011$	$8.144e - 006$
751	$7.162e - 012$	$2.707e - 006$
801	$7.850e - 013$	$8.901e - 007$
851	$8.076e - 014$	$2.745e - 007$
901	$7.436e - 015$	$8.559e - 008$
902	$7.152e - 015$	$8.412e - 008$

表 9.4: 1000×1000 的方程组, $Q_f = 1.e6$

Itr	$f - f_*$	$\ x - x_*\ $
1	$9.916e + 006$	$2.849e + 001$
1000	$7.190e + 000$	$2.683e + 000$
2000	$4.839e - 002$	$2.207e - 001$
3000	$4.091e - 004$	$1.999e - 002$
4000	$2.593e - 006$	$1.602e - 003$
5000	$1.526e - 008$	$1.160e - 004$
6000	$1.159e - 010$	$1.102e - 005$
7000	$6.022e - 013$	$7.883e - 007$
8000	$3.386e - 015$	$5.595e - 008$
8103	$1.923e - 015$	$4.236e - 008$

表 9.5: 1000×1000 的方程组, $Q_f = 1.e12$

Itr	$f - f_*$	$\ x - x_*\ $
1	$5.117e + 012$	$3.078e + 001$
1000	$1.114e + 007$	$2.223e + 001$
2000	$2.658e + 006$	$2.056e + 001$
3000	$1.043e + 006$	$1.964e + 001$
4000	$5.497e + 005$	$1.899e + 001$
5000	$3.444e + 005$	$1.851e + 001$
6000	$2.343e + 005$	$1.808e + 001$
7000	$1.760e + 005$	$1.775e + 001$
8000	$1.346e + 005$	$1.741e + 001$
9000	$1.045e + 005$	$1.709e + 001$
10000	$8.226e + 004$	$1.679e + 001$

9.1.3 非二次扩展.

能够被应用于不必非是二次的任何形式函数 f 的CG(Fletcher-Reeds CG)如

下:

$$\begin{aligned}
d_0 &= -g_0 = -f'(x_0) \\
\gamma_t &= \operatorname{argmin}_{\gamma} f(x_{t-1} + \gamma d_{t-1}) \\
x_t &= x_{t-1} + \gamma_t d_{t-1} \\
g_t &= f'(x_t) \\
\beta_t &= \frac{g_t^T g_t}{g_{t-1}^T g_{t-1}} \\
d_t &= -g_t + \beta_t d_{t-1}.
\end{aligned}$$

类似地, 另一种在二次形时是等价的Polak-Ribiere CG:

$$\begin{aligned}
d_0 &= -g_0 = -f'(x_0) \\
\gamma_t &= \operatorname{argmin}_{\gamma} f(x_{t-1} + \gamma d_{t-1}) \\
x_t &= x_{t-1} + \gamma_t d_{t-1} \\
g_t &= f'(x_t) \\
\beta_t &= \frac{(g_t - g_{t-1})^T g_t}{g_{t-1}^T g_{t-1}} \\
d_t &= -g_t + \beta_t d_{t-1}.
\end{aligned}$$

两种方法在二次函数时是等价的, 在非二次情况下变得不同!

CG的非二次推广可以在有重新启动和没有重新启动的情况下使用. 在二次CG情况中, 忽略舍入误差, 在最多 n 步终止并找到精确解. 在非二次情况下不是这样. 在具有重新启动功能的非二次CG中, 执行分为 n 步一个周期, 并且周期 $t+1$ 从上一个周期的最后一个迭代 x_t 开始, 就像从初始点开始. 与此相反, 没有重新启动的循环

$$\begin{aligned}
d_0 &= -g_0 = -f'(x_0) \\
\gamma_t &= \operatorname{argmin}_{\gamma} f(x_{t-1} + \gamma d_{t-1}) \\
x_t &= x_{t-1} + \gamma_t d_{t-1} \\
g_t &= f'(x_t) \\
\beta_t &= \frac{(g_t - g_{t-1})^T g_t}{g_{t-1}^T g_{t-1}} \\
d_t &= -g_t + \beta_t d_{t-1}
\end{aligned}$$

重不“刷新”.

定理9.2. 设 f 的水平集 $G = \{x : f(x) \leq f(x_0)\}$ 是紧的, 并且 f 在 G 的邻域上是二次连续可微的. 当使用结合精确线搜索和重新启动功能的Fletcher-Reeves或者Polak-Ribiere 共轭梯度法最小化 f 时,

- (i) 轨迹是适定并且有界的,
- (ii) f 不增,
- (iii) 包括子循环的迭代的序列 x^t 的所有极限点均是 f 的临界点.
- (iv) 另外, 如果 x^t 收敛到 f 的非退化局部极小值点 x_* , 并且 f 在 x_* 邻域上是三次连续可微的, 那么 x^t 二次收敛到 x_* .

9.2 拟牛顿法

拟牛顿法属于一般形式为

$$\begin{aligned} x_{t+1} &= x_t - \gamma_{t+1} \underbrace{S_{t+1}}_{\text{null}} f'(x_t) \\ &= A_{t+1}^{-1} \end{aligned}$$

的变度量方法, 其中 $S_{t+1} \succ 0$ 并且 γ_{t+1} 由线搜索给出. 与修正牛顿方法相反, 在拟牛顿算法中, 人们直接对矩阵 S_{t+1} 进行运算, 其最终目标是确保在有利的情况下

$$S_{t+1} - [f''(x_t)]^{-1} \rightarrow 0, t \rightarrow \infty \quad (9.7)$$

为了达到(9.7), 在拟牛顿法中, 将 S_t 更新至 S_{t+1} 的方式需要确保 $S_{t+1} \succ 0$ 和

$$S_{t+1}(g_t - g_{t-1}) = x_t - x_{t-1},$$

其中 $g_\tau = f'(x_\tau)$.

通用拟牛顿法

初始化: 选择某个初始点 x_0 和矩阵 $S_1 \succ 0$, 计算 $g_0 = f'(x_0)$.

Step t: 给定 $x_{t-1}, g_{t-1} = f'(x_{t-1})$ 和 $S_t \succ 0$, 当 $g_{t-1} = 0$ 时终止, 否则, 置 $d_t = -S_t g_{t-1}$, 并沿方向 d_t 执行精确线搜索得到 γ_t . 因此获得新迭代

$$x_t = x_{t-1} + \gamma_t d_t$$

计算 $g_t = f'(x_t)$, 并且置

$$p_t = x_t - x_{t-1}, q_t = g_t - g_{t-1};$$

将 S_t 更新成正定对称矩阵 S_{t+1} 使得

$$S_{t+1} q_t = p_t.$$

然后循环.

请注意 $g_{t-1}^T d_t < 0$ (由于 $g_{t-1} \neq 0$ 和 $S_t \succ 0$) 和 $g_t^T p_t = 0$ (因为 x_t 是 f 在射线 $\{x_{t-1} + \gamma d_t : \gamma > 0\}$ 上的最小点) 推出 $p_t^T q_t > 0$. 该事实在证明标准拟牛顿法中 S_t 的正

定性时非常有用. 对更新规则 $S_t \mapsto S_{t+1}$ 的要求: (1) 为确保 d_{t+1} 是 f 的下降方向, 规则应该保证 $S_{t+1} \succ 0$; (2) 在强凸二次 f 的情况下, 规则应该保证当 $t \rightarrow \infty$ 时, $S_t - [f''(x_t)]^{-1} \rightarrow 0$.

Davidon-Fletcher-Powell法:

$$S_{t+1} = S_t + \frac{1}{p_t^T q_t} p_t p_t^T - \frac{1}{q_t^T S_t q_t} S_t q_t q_t^T S_t.$$

当应用于强凸二次函数时, DFP方法在 n 步之内得到精确解. 以 $S_1 = I$ 作为初始矩阵的方法生成的轨迹恰好是共轭梯度法的轨迹, 因此初始矩阵为单位矩阵的DFP法是共轭梯度法. **Broyden族**. Broyden-Fletcher-Goldfarb-Shanno更新公式:

$$S_{t+1}^{\text{BFGS}} = S_t + \frac{1 + q_t^T S_t q_t}{(p_t^T q_t)^2} p_t p_t^T - \frac{1}{p_t^T q_t} [p_t q_t^T S_t + S_t q_t p_t^T]$$

将BFGS公式结合DFP公式

$$S_{t+1}^{\text{DFP}} = S_t + \frac{1}{q_t^T p_t} p_t p_t^T - \frac{1}{q_t^T S_t q_t} S_t q_t q_t^T S_t$$

产生单参数的更新Broyden公式族

$$S_{t+1}^\phi = (1 - \phi) S_{t+1}^{\text{DFP}} + \phi S_{t+1}^{\text{BFGS}},$$

其中 $\phi \in [0, 1]$ 是参数.

事实是当将Broyden方法应用于强凸 n 维二次形 f 时, 在 n 步之内精确地最小化 f . 如果 S_0 与单位矩阵成比例, 那么该方法关于 f 的轨迹恰好是一个共轭梯度法. 从同一对 (x_0, S_1) 开始的精确线搜索Broyden方法应用于相同的问题时, 它们生成的迭代序列(尽管矩阵 S_t 的顺序可能不相同!)是相同的, 与参数 ϕ 的选择无关. 在实践中, 共轭梯度法和拟牛顿法中纯粹的BFGS方法($\phi = 1$)似乎是最好的.

拟牛顿法的收敛. 仅对于某些版本的方法并且仅在 f 是强凸的假设下, 证明了不重新启动的拟牛顿法是**全局收敛**的. 对于通过每隔 m 步置 $S = S_0$ 来“刷新”更新公式的重新启动的方法, 易于证明在我们的标准假设: 水平集 $G = \{x : f(x) \leq f(x_0)\}$ 是紧的, 并且 f 在 G 的邻域上是连续可微的, 循环初始点的轨迹是有界的, 并且轨迹的所有极限点都是 f 的临界点.

局部收敛. 对于具有重新启动的方案, 可以证明如果 $m = n$ 和 $S_0 = I$, 如果循环初始点 x_t 的轨迹收敛到三次连续可微函数 f 的非退化局部极小值点 x_* , 那么轨迹二次收敛到 x_* .

定理9.3 (Powell, 1976). 考虑不重新启动的BFGS方法, 且假设方法收敛到三次连续可微函数 f 的一个非退化局部极小值点 x^* . 那么该方法超线性收敛到 x^* .

11 约束最小化方法：惩罚/障碍法

可将求解一般约束问题

$$\min_x \left\{ f(x) : \begin{array}{l} g_j(x) \leq 0, j = 1, \dots, m \\ h_i(x) = 0, i = 1, \dots, k \end{array} \right\} \quad (\text{P})$$

的传统方法分为**原始方法**，模仿无约束规划方法，沿着可行集行进，并确保每一步目标函数有所提高；**惩罚/障碍法**将约束最小化问题化归成为求解一系列本质上的无约束问题；**拉格朗日乘子法**着眼于与问题(P)关联的对偶问题，与惩罚/障碍法类似，但以不同于惩罚/障碍策略的“智能”方式，后验拉格朗日乘子法将问题(P)化归成一系列无约束问题；**逐步二次规划法**是直接求解与问题(P)关联的KKT系统的牛顿法。

11.1 惩罚/障碍策略

考虑等式约束问题

$$\min_x \{ f(x) : h_i(x) = 0, i = 1, \dots, k \} \quad (\text{EP})$$

并且利用无约束问题

$$\min_x f_\rho(x) = f(x) + \underbrace{\frac{\rho}{2} \sum_{i=1}^k h_i^2(x)}_{\text{penalty term}} \quad (\text{EP}[\rho])$$

“近似” (EP)，其中 $\rho > 0$ 是惩罚参数。请注意在可行集内，惩罚项消失，因此 $f_\rho \equiv f$ ；当 ρ 很大并且 x 是不可行的时， $f_\rho(x)$ 会非常大：

$$\lim_{\rho \rightarrow \infty} f_\rho(x) = \begin{cases} f(x), & x \text{ 可行} \\ \text{null} & \\ +\infty, & \text{否则} \end{cases} \quad (11.1)$$

从而自然期待当 $\rho \rightarrow \infty$ 时 (EP[ρ]) 的解靠近 (EP) 的最优解集。

对于一般约束问题(P)，采用同样的惩罚约束违反的想法，得到无约束问题

$$\min_x f_\rho(x) = f(x) + \underbrace{\frac{\rho}{2} \left[\sum_{i=1}^k h_i^2(x) + \sum_{j=1}^m [g_j(x)^+]^2 \right]}_{\text{penalty term}} \quad (\text{P}[\rho])$$

以近似(P)，其中

$$g_j^+ = \max[g_j(x), 0]$$

且 $\rho > 0$ 是惩罚参数。这里又一次有(11.1)成立。我们再次期待当 $\rho \rightarrow \infty$ 时 (P[ρ]) 的解接近(P)的最优解集合。

障碍策略通常用于满足“Slater 条件”的不等式约束问题

$$\min_x \{f(x) : g_j(x) \leq 0, j = 1, \dots, m\}. \quad (\text{IEP})$$

满足“Slater 条件”指问题(IEP)的可行集

$$G = \{x : g_j(x) \leq 0, j \leq m\} \quad (11.2)$$

具有非空内部 $\text{int}G$ ，这里 $\text{int}G$ 在 G 中是稠密的，并且对 $x \in \text{int}G$ ， $g_j(x) < 0$ 。

给定问题(IEP)，可以为 G 构造一个障碍(等价于内部惩罚)——函数 F ，其在 $\text{int}G$ 中是适定的且光滑的，沿着收敛到 G 的边界点的每个点 $x_i \in \text{int}G$ ，增大至 ∞ ：

$$x_i \in \text{int}G, \lim_{i \rightarrow \infty} x_i = x \notin \text{int}G \Rightarrow F(x_i) \rightarrow \infty, i \rightarrow \infty$$

常见的例子有对数障碍 $F(x) = -\sum_j \ln(-g_j(x))$ 和Carroll障碍 $F(x) = -\sum_j \frac{1}{g_j(x)}$

在为(IEP)可行域的内部选取了惩罚 F 后，由“本质上无约束”问题

$$\min_{x \in \text{int}G} F_\rho(x) = f(x) + \frac{1}{\rho} F(x) \quad (\text{B}[\rho])$$

近似(IEP). 当惩罚参数 ρ 很大时，除了边界周围的细条之外，函数 F_ρ 在 G 中的所有位置都接近 f 。从而，自然期待当 $\rho \rightarrow \infty$ 时(B[ρ])的解接近(IEP)的最优解的集合。

11.2 探讨惩罚策略

让我们重点讨论等式约束问题(EP)和关联的惩罚问题(EP[ρ])。对于一般情况(P)，结果是类似的。我们感兴趣的问题：是否的确当 $\rho \rightarrow \infty$ 时，惩罚目标 f_ρ 的无约束极小点收敛到(P)的最优解集合？有哪些方法可用来最小化惩罚目标？

定理11.1 (简单事实). 设(EP)是可行的，(EP)的目标和约束是连续的，并且设 f 的水平集 $\{x : f(x) \leq a\}$ 是有界的。进一步设 X_* 是(EP)的全局解集合。那么 X_* 是非空的，近似问题(EP[ρ])是可解的，并且当 $\rho \rightarrow \infty$ 时，它们的全局解靠近 X_* ：

$$\begin{aligned} & \forall \epsilon > 0 \exists \rho(\epsilon) : \rho \geq \rho(\epsilon), x_*(\rho) \text{ 求解 (EP}[\rho]) \\ & \Rightarrow \text{dist}(x_*(\rho), X_*) \equiv \min_{x_* \in X_*} \|x_*(\rho) - x_*\|_2 \leq \epsilon. \end{aligned}$$

证明. 1⁰. 由假设可知，(EP)的可行集是非空闭的， f 是连续的并且当 $\|x\|_2 \rightarrow \infty$ 时， $f(x) \rightarrow \infty$ 。由此可见， f 在可行集上达到最小值，并且 f 在可行集上的全局最小值点集合 X_* 是有界闭集。

2⁰. (EP[ρ])的目标函数是连续的，并且当 $\|x\|_2 \rightarrow \infty$ 时，其趋于 ∞ ；所以，(EP[ρ])是可解的。

3⁰. 有待证明的: 对于每一个 $\epsilon > 0$, 具有足够大的 ρ 值使得 (EP[ρ]) 的解都属于 X_* 的邻域. 相反, 假设对于某个 $\epsilon > 0$, 存在序列 $\rho_i \rightarrow \infty$ 使得 (EP[ρ]) 的最优解 x_i 距 X_* 的距离大于 ϵ , 并且让我们由此假设导出矛盾.

设 f_* 是 (EP) 的最优值. 我们清楚地有

$$f(x_i) \leq f_{\rho_i}(x_i) \leq f_*. \quad (11.3)$$

因此 $\{x_i\}$ 是有界的. 传递给子序列, 我们可以假设当 $i \rightarrow \infty$ 时 $x_i \rightarrow \bar{x}$. 这样,

$$x_i \in \operatorname{argmin}_x f_{\rho_i}(x), x_i \rightarrow \bar{x} \notin X_*.$$

我们断言 $\bar{x} \in X_*$, 这给出了想要的矛盾. 确实, \bar{x} 是 (EP) 的可行解, 由于否则

$$\lim_{i \rightarrow \infty} \underbrace{\left[f(x_i) + \frac{\rho_i}{2} \|h(x_i)\|_2^2 \right]}_{f_{\rho_i}(x_i)} = f(\bar{x}) + \lim_{i \rightarrow \infty} \frac{\rho_i}{2} \underbrace{\|h(x_i)\|_2^2}_{\rightarrow \|h(\bar{x})\|_2^2 > 0} = +\infty,$$

这与 (11.3) 矛盾. 由 (11.3) 可知 $f(\bar{x}) = \lim_{x \rightarrow \infty} f(x_i) \leq f_*$; 由于 \bar{x} 是 (EP) 的可行解, 所以 $f(\bar{x}) = f_*$. 得出结论 $\bar{x} \in X_*$. \square

简单事实的缺点是在非凸情况下, 我们无法找到/近似受惩罚目标的全局最小值点, 因此简单事实是“虚幻的”.

定理11.2. 设 x^* 是问题 (EP) 的一个非退化的局部最优解, 即一个可行解使得 f, h_i 在 x_* 的附近二次连续可微. 约束函数在 x_* 处的梯度是线性无关的, 且在 x_* 处, 二阶充分最优性条件满足. 那么

- 存在 x^* 的一个邻域 V 和 $\bar{\rho} > 0$ 使得对于每一个 $\rho \geq \bar{\rho}$, f_ρ 在 V 中恰好具有一个临界点 $x_*(\rho)$; $x_*(\rho)$ 是 f_ρ 的一个非退化局部最小值点和 f_ρ 在 V 的最小值点, 并且当 $\rho \rightarrow \infty$ 时 $x_*(\rho) \rightarrow x_*$.
- 局部“惩罚最优值”

$$f_\rho(x_*(\rho)) = \min_{x \in V} f_\rho(x)$$

关于 ρ 是不减的; 当 $\rho \rightarrow \infty$ 时, 约束违反量 $\|h(x_*(\rho))\|_2$ 单调趋于 0; $x_*(\rho)$ 处的目标函数真值 $f(x_*(\rho))$ 关于 ρ 是不减的, 并且量 $\rho h_i(x_*(\rho))$ 收敛到 (EP) 在 x_* 处的最优拉格朗日乘子.

注记. 的确, $f_\rho(\cdot) = f(\cdot) + \frac{\rho}{2} \|h(\cdot)\|_2^2$ 随着 ρ 递增. 由此易得 $f_\rho(x_*(\rho))$ 关于 ρ 是不减的. 此外, 设 $\rho'' > \rho'$, 并且设 $x' = x_*(\rho')$, $x'' = x_*(\rho'')$. 那么由

$$f(x') + \frac{\rho''}{2} \|h(x')\|_2^2 \geq f(x'') + \frac{\rho''}{2} \|h(x'')\|_2^2$$

和

$$f(x'') + \frac{\rho'}{2} \|h(x'')\|_2^2 \geq f(x') + \frac{\rho'}{2} \|h(x')\|_2^2$$

有

$$f(x') + f(x'') + \frac{\rho''}{2} \|h(x')\|_2^2 + \frac{\rho'}{2} \|h(x'')\|_2^2 \geq f(x') + f(x'') + \frac{\rho''}{2} \|h(x'')\|_2^2 + \frac{\rho'}{2} \|h(x')\|_2^2.$$

进而有

$$\frac{\rho'' - \rho'}{2} \|h(x')\|_2^2 \geq \frac{\rho'' - \rho'}{2} \|h(x'')\|_2^2.$$

这样, 约束违反量 $\|h(x_*(\rho))\|_2$ 关于 ρ 是单调递减的. 最后, 由局部一阶必要最优性条件, $x_*(\rho)$ 满足

$$0 = f'_\rho(x_*(\rho)) = f'(x_*(\rho)) + \sum_i (\rho h_i(x_*(\rho))) h'_i(x_*(\rho)).$$

由此可得 $\rho h_i(x_*(\rho))$ 收敛到 (EP) 在 x_* 处的最优拉格朗日乘子.

求解惩罚问题. 原则上, 可以用任何用于无约束最小化的方法求解 (EP[ρ]). 然而 f_ρ 的条件数随着 $\rho \rightarrow \infty$ 而变差. 事实上, 当 $\rho \rightarrow \infty$ 时, 我们有

$$\underbrace{d^T f''_\rho(x_*(\rho)) d}_{\rightarrow x_*} = \underbrace{d^T [f''(x) + \sum_i \rho h_i(x) h''_i(x)] d}_{\rightarrow \nabla_x^2 L(x_*, \mu^*)} + \underbrace{\rho \sum_i (d^T h'_i(x))^2}_{\rightarrow \infty, \rho \rightarrow \infty \text{ except for } d^T h'(x_*)=0}$$

由此, 当惩罚参数很大时, 收敛放慢和/或严重的数值困难.

11.3 用障碍法求解凸规划

考虑等式约束优化问题 (IEP), 其可行域 G 的定义见 (11.2). 设 F 是 $G = \text{cl}(\text{int}G)$ 的内部惩罚函数, 即 F 在 $\text{int}G$ 上是光滑的, 且 F 沿着每个收敛到 G 的边界点的序列 $x_i \in \text{int}G$ 趋于 ∞ .

定理11.3. 假设 $G = \text{cl}(\text{int}G)$ 是有界的并且 f, g_j 在 G 上是连续的. 那么 (B[ρ]) 的最优解集合 X_* 和 $(P[\rho])$ 的最优解集合 $X_*(\rho)$ 是非空的, 并且当 $\rho \rightarrow \infty$ 时, 第二个集合收敛为第一个集合: 对于每一个 $\epsilon > 0$, 那么存在 $\rho = \rho(\epsilon)$ 使得

$$\rho \geq \rho(\epsilon), x_*(\rho) \in X_*(\rho) \Rightarrow \text{dist}(x_*(\rho), X_*) \leq \epsilon.$$

在凸规划的情况下, 即

$$\min_{x \in G} f(x), \quad (\text{CP})$$

其中 G 是有界闭凸集并且目标 f 是凸的, 有多种方式为定义域 G 设计二次连续可微的强凸惩罚 $F(x)$. 假设 f 在 $\text{int}G$ 上是二次连续可微的, 聚合函数

$$F_\rho(x) = \rho f(x) + F(x)$$

在 $\text{int}G$ 上是强凸的, 因此在单一点

$$x_*(\rho) = \underset{x \in \text{int}G}{\text{argmin}} F_\rho(x)$$

取到它的最小值. 易于看出, 路径 $x_*(\rho)$ 是连续可微的, 并且当 $\rho \rightarrow \infty$ 时收敛到(CP)的最优集合, 即

$$x_*(\rho) = \operatorname{argmin}_{x \in \operatorname{int} G} F_\rho(x) \xrightarrow{\rho \rightarrow \infty} \operatorname{argmin}_G f.$$

在经典路径跟踪策略(Fiacco和McCormic, 1967)中, 根据以下的通用策略, 随着 $\rho \rightarrow \infty$ 跟踪路径 $x_*(\rho)$:

给定 $x_i \in \operatorname{int} G, \rho_i > 0$, 并且 x_i 接近 $x_*(\rho_i)$. 将 ρ_i 更新为更大的惩罚值 ρ_{i+1} . 以 x_i 作为初始点最小化 $F_{\rho_{i+1}}(\cdot)$, 直到构造出足够接近

$$x_*(\rho_{i+1}) = \operatorname{argmin}_{x \in \operatorname{int} G} F_{\rho_{i+1}}(x)$$

的新迭代 x_{i+1} , 并且循环.

为了将 $\operatorname{argmin} F_{\rho_i}(x)$ 的紧近似解 x_i 更新为 $\operatorname{argmin} F_{\rho_{i+1}}(x)$ 的紧近似解 x_{i+1} , 可以对 $F_{\rho_{i+1}}(\cdot)$ 应用针对“本质上无约束”最小化的方法, 首选是牛顿法. 当使用牛顿法时, 可以尝试以“安全”的速度增加惩罚, 使 x_i 保持在应用于 $F_{\rho_{i+1}}(\cdot)$ 的牛顿法的二次收敛域范围内, 因此利用方法的快速局部收敛.

问题是如何选择 F ? 如何测量与路径的贴近程度? 如何在不减慢方法速度的情况下确保“安全”的更新惩罚?

请注意, 随着 $\rho \rightarrow \infty$, $F'_\rho(x_*(\rho))$ 的条件数可能放大至 ∞ , 根据传统牛顿法的理论, 这使得将 x_i 更新为 x_{i+1} 的问题越来越困难. 因此, 减速似乎是不可避免的……

在80年代后期, 人们发现, 与适当选择的障碍关联的经典路径跟踪策略允许“安全”实现而不会减慢速度. 这一发现导致了用于凸规划的多项式时间内点方法的发明.

大多数多项式时间内点法都大量利用经典路径跟踪策略; 新颖之处在于它的障碍——那些特别适合于牛顿法最小化的特定**self-concordant**函数.

设 G 是内部非空的闭凸域并且不含直线. 称三次连续可微凸函数

$$F(x) : \operatorname{int} G \rightarrow \mathbb{R}$$

是**self-concordant**的, 如果 F 是 G 的内部惩罚:

$$x_i \in \operatorname{int} G, x_i \rightarrow x \in \partial G \Rightarrow F(x_i) \rightarrow \infty$$

并且 F 满足关系

$$\left| \frac{d^3}{dt^3} \Big|_{t=0} F(x + th) \right| \leq 2 \left(\frac{d^2}{dt^2} \Big|_{t=0} F(x + th) \right)^{3/2}$$

设 $\vartheta \geq 1$. 称 F 是 G 的 ϑ -**self-concordant**障碍, 如果除了在 G 上是**self-concordant**外, F 还满足关系

$$\left| \frac{d}{dt} \Big|_{t=0} F(x + th) \right| \leq \sqrt{\vartheta} \left(\frac{d^2}{dt^2} \Big|_{t=0} F(x + th) \right)^{1/2},$$

称 ϑ 是s.-c.b.F的参数.

每个凸规划(CP)都可以转换成具有线性目标的凸规划, 即

$$\min_{t,x} \{t : x \in G, f(x) \leq t\}$$

假设这种转换在一开始就已经完成, 不失一般性我们可以专注于线性目标的凸规划

$$\min_{x \in G} c^T x \quad (\text{CP}')$$

假设 G 是内部非空的有界闭凸集, 设 F 是 G 的 ϑ -s.c.b.障碍.

事实I: F 在 $\text{int}G$ 上是强凸的: 对于所有的 $x \in \text{int}G$, $F''(x) \succ 0$. 因此

$$F_\rho(x) \equiv \rho c^T x + F(x)$$

在 $\text{int}G$ 上也是强凸的. 特别地, 称数量

$$\lambda(x, F_\rho) = ([F'_\rho(x)]^T \underbrace{[F''_\rho(x)]}_{=F''(x)}^{-1} F'_\rho(x))^{1/2}$$

是 F_ρ 在 x 处的牛顿减量(decrement), 它对于所有 $x \in \text{int}G$ 和所有 $\rho > 0$ 是适定的.

注意到

$$\frac{1}{2}\lambda^2(x, F_\rho) = F_\rho(x) - \min_y [F_\rho(x) + (y-x)^T F'_\rho(x) + \frac{1}{2}(y-x)^T F''_\rho(x)(y-x)]$$

$\lambda(x, F_\rho) \geq 0$ 且 $\lambda(x, F_\rho) = 0$ 当且仅当 $x = x_*(\rho)$. 因此, 可将牛顿减量视为“贴进度”, 即从 x 到 $x_*(\rho)$ 的距离.

事实II: 设通过以下经典惩罚策略实施待求解的问题(CP'): 该策略下 G 的障碍是 ϑ -s.-c.b.F;

由关系 $\lambda(x, F_\rho) \leq 0.1$ 指定 x 和 $x_*(\rho)$ 的“贴进度”;

惩罚更新是 $\rho_{i+1} = (1 + \frac{\gamma}{\sqrt{\vartheta}})\rho_i$, 其中 $\gamma > 0$ 是参数.

为了将 x_i 更新为 x_{i+1} , 对 $F_{\rho_{i+1}}$ 应用从 x_i 开始的阻尼牛顿法:

$$x \mapsto x - \frac{1}{1 + \lambda(x, F_{\rho_{i+1}})} [F''_{\rho_{i+1}}(x)]^{-1} F'_{\rho_{i+1}}(x)$$

方法是适定的, 并且在更新 $x_i \mapsto x_{i+1}$ 中所产生的阻尼牛顿步数只依赖 γ (并且对于 $\gamma = 0.1$, 它和1一样小). 有 $c^T x_i - c_* \leq \frac{2\vartheta}{\rho_i}$. 因此使用上述方法需要 $O(\sqrt{\vartheta})$ 个牛顿步将不准确度 $c^T x - c_*$ 减小一个绝对常数倍数!

事实III: 每个凸域 $G \subset \mathbb{R}^n$ 都拥有 $O(n)$ -s.-c.b.. 对于凸规划中出现的典型可行域, 可以指出明确的“可计算”s.-c.b.. 例如,

设 G 由 m 个凸二次约束给出:

$$G = \{x : \underbrace{x^T A_j^T A_j x + 2b_j^T x + c_j}_{g_j(x)} \leq 0, 1 \leq j \leq m\}$$

并且满足Slater条件：存在某 \bar{x} 使得 $g_j(\bar{x}) < 0, j = 1, \dots, m$. 那么对数障碍

$$F(x) = - \sum_{j=1}^m \ln(-g_j(x))$$

是 G 的 m -s.-c.b.. 设 G 由线性矩阵不等式给出

$$G = \{x : \underbrace{A_0 + x_1 A_1 + \dots + x_n A_n}_{A(x):m \times m} \succeq 0\}$$

并且满足Slater条件：对于某个 \bar{x} , $A(\bar{x}) \succ 0$. 那么log-det障碍

$$F(x) = - \ln \text{Det}(A(x))$$

是 G 的 m -s.-c.b..

11.4 求解LP的原始对偶内点法

考虑线性规划问题

$$\min_z \{c^T z : Az - b \geq 0\} \quad (\text{LP})$$

和对偶问题

$$\max_y \{b^T y : A^T y = c, y \geq 0\},$$

并且假设两个问题都是严格可行的：

$$\exists \bar{z} : A\bar{z} - b > 0 \ \& \ \exists y > 0 : A^T y = c.$$

注意从 z 转变到“原始松弛” $x = Az - b$, 我们可以将(LP)重写作

$$\min_x \{e^T x : x \geq 0, x + b \in \text{Im} A\} \quad (11.4)$$

其中 e 是满足 $A^T e = c$ 的向量, 因此

$$e^T x = e^T (Az - b) = (A^T e)^T z - \text{const} = c^T z - \text{const}.$$

从而对偶问题即可表述为

$$\max_y \{b^T y : \underbrace{A^T y = c \equiv A^T e}_{\Leftrightarrow y - e \in (\text{Im} A)^\perp}\}. \quad (\text{DP})$$

设 $\Phi(x) = - \sum_{i=1}^m \ln x_i$. 为(LP)的定义域设计 m -s.c.b.的 $F(z) = \Phi(Az - b)$, 考虑

$$\begin{aligned} z_*(\rho) &= \underset{z}{\operatorname{argmin}} [\rho c^T z + F(z)] \\ \text{null} \quad &= \underset{z}{\operatorname{argmin}} [\rho e^T (Az - b) + \Phi(Az - b)] \end{aligned}$$

观察到点 $x_*(\rho) = Az_*(\rho) - b$ 在(11.4)的可行集上最小化 $\rho e^T x + \Phi(x)$, 即点 $x = x_*(\rho)$ 满足:

$$x > 0, x + b \in \text{Im}A, \rho e + \Phi'(x) \in (\text{Im}A)^\perp.$$

由此推出 $y = y_*(\rho) = -\rho^{-1}\Phi'(x_*(\rho))$ 满足

$$y > 0, y - e \in (\text{Im}A)^\perp, \underbrace{-\rho b + \Phi'(y)}_{=-\rho(x+b)} \in \text{Im}A$$

即点 $y_*(\rho)$ 在(DP)的可行集上最小化 $-\rho b^T y + \Phi(y)$.

我们得出一个很好的对称图: 原始中心路径 $x_* = x_*(\rho)$ 在由下式给出的原始可行集

$$x_* > 0, x_* + b \in \text{Im}A, \rho e + \Phi'(x_*) \in (\text{Im}A)^\perp$$

上最小化原始聚合函数

$$\rho e^T x + \Phi(x) \quad [\Phi(x) = -\sum_i \ln x_i].$$

对偶中心路径 $y_* = y_*(\rho)$ 在由下式给出的对偶可行集

$$y_* > 0, y_* - e \in (\text{Im}A)^\perp, -\rho b + \Phi'(y_*) \in \text{Im}A$$

最小化对偶聚合函数

$$-\rho b^T y + \Phi(y) \quad [\Phi(y) = -\sum_i \ln y_i].$$

路径之间由

$$y_* = -\rho^{-1}\Phi'(x_*) \Leftrightarrow x_* = -\rho^{-1}\Phi'(y_*) \Leftrightarrow [x_*]_i [y_*]_i = \frac{1}{\rho}, \forall i$$

关联. 由此推出在路径上 $x = x_*(\rho)$, $y = y_*(\rho)$ 处的

$$\text{DualityGap}(x, y) = x^T y = [c^T x - \text{Opt}(P)] + [\text{Opt}(D) - b^T y]$$

是 $m\rho^{-1}$.

用于线性规划的通用原始-对偶内点法:

给定当前迭代—原始对偶严格可行对 x^i, y^i 和惩罚值 ρ_i , 按如下方式将其更新为新迭代 $x^{i+1}, y^{i+1}, \rho_{i+1}$

◇ 更新 $\rho_i \mapsto \rho_{i+1} \geq \rho_i$

◇ 由针对方程组

$$x > 0, x + b \in \text{Im}A; y > 0, y - e \in (\text{Im}A)^\perp$$

$$\text{null Diag}\{x\}y = \frac{1}{\rho_{i+1}} \underbrace{(1, \dots, 1)^T}_e \quad [\Leftrightarrow x_s y_s = \frac{1}{\rho_{i+1}}, 1 \leq s \leq m]$$

的牛顿步定义原始-对偶中心路径:

$$x^{i+1} = x^i + \Delta x, y^{i+1} = y^i + \Delta y,$$

其中 $\Delta x, \Delta y$ 求解线性方程组

$$\Delta x \in \text{Im}A, \Delta y \in (\text{Im}A)^\perp$$

$$\begin{aligned} & \text{Diag}\{x^i\}y^i + \text{Diag}\{x^i\} \Delta y + \text{Diag}\{y^i\} \Delta x = \frac{e}{\rho_{i+1}} \\ & \left[\Leftrightarrow x_s^i y_s^i + x_s^i \cdot [\Delta y]_s + y_s^i \cdot [\Delta x]_s = \frac{1}{\rho_{i+1}}, 1 \leq s \leq m \right] \end{aligned}$$

应用于问题(LP)和 m -s.c.b $F(z) = \Phi(Az - b)$ 的经典路径跟随策略允许追踪路径 $z_*(\rho)$ (并且因此 $x_*(\rho) = Az_*(\rho) - b$). 更高级的原始-对偶路径跟踪方法同时跟踪原始和对偶中心路径, 由此得到的算法策略的实际性能优于“纯粹原始”策略得到算法中的.

通过恰当的实现, 两种方法都可以为LP带来迄今为止最有名的理论复杂度界. 根据这些界, 对具有 $m \times n$ 矩阵 A 的严格可行LP, 产生一个原始-对偶可行 ϵ 解的“算术运算开销”是

$$O(1)mn^2 \ln \left(\frac{mn\Theta}{\epsilon} \right),$$

其中 $O(1)$ 是一个绝对常数, Θ 是一个与数据相关的常数. 在实践中, 恰当实现的原始-对偶方法远胜于纯粹的原始方法, 并用数十次牛顿迭代中求解了具有数万和数十万个变量和约束的现实世界中的线性规划问题.

12 约束最小化方法：增广拉格朗日方法

考虑等式约束问题

$$\min_x \{f(x) : h_i(x) = 0, i = 1, \dots, k\} \quad (\text{EP})$$

惩罚方法利用无约束问题

$$\min_x f_\rho(x) = f(x) + \underbrace{\frac{\rho}{2} \sum_{i=1}^k h_i^2(x)}_{\text{penalty term}} \quad (\text{EP}[\rho])$$

“近似”它，其中 $\rho > 0$ 是惩罚参数。

惩罚方法的缺点是为了高精度地求解问题(EP)，应该使用较大的惩罚值，这导致很难最小化(EP[ρ])中被惩罚的目标。

增广拉格朗日方法以“明智的方式”使用惩罚机制，从而避免了使用非常大的 ρ 值进行工作的必要性。

12.1 局部拉格朗日对偶

问题(EP)的Lagrange函数是

$$L(x, \lambda) = f(x) + \sum_i \lambda_i h_i(x).$$

设 x_* 是问题(EP)的非退化局部解，因此存在 λ^* 使得

$$\nabla_x L(x_*, \lambda^*) = 0 \quad (12.1)$$

$$d^T \nabla_x^2 L(x_*, \lambda^*) d > 0 \quad \forall 0 \neq d \in T_{x_*} := \{d : d^T h'_i(x^*) = 0, i = 1, \dots, k\}. \quad (12.2)$$

暂且将(12.2)替换为假设更强的条件：

$$\text{矩阵 } \nabla_x^2 L(x_*, \lambda^*) \text{ 在整个空间上是正定的} \quad (12.3)$$

成立. 在假设(12.3)的基础上， x_* 是光滑函数

$$L(\cdot, \lambda^*)$$

的非退化无约束局部极小点，因此可以用无约束最小化方法找到该点。

这样，如果足够聪明以至于猜测出拉格朗日乘子向量 λ^* ，并且很幸运的有 $\nabla_x^2 L(x_*, \lambda^*) \succ 0$ ，那么可以通过无约束优化技术找到 x_* 。

在幸运的情况下如何变得智慧？采用的技术是局部拉格朗日对偶。现在面对的问题： x_* 是问题(EP)的非退化局部解，并且我们很幸运：

$$\exists \lambda^* : \nabla_x L(x_*, \lambda^*) = 0, \nabla_x^2 L(x_*, \lambda^*) \succ 0. \quad (12.4)$$

事实： 在假设(12.4)的前提下，存在 x_* 的凸邻域 V 和 λ^* 的凸邻域 Λ 使得

- (i) 对于每个 $\lambda \in \Lambda$ ，函数 $L(x, \lambda)$ 在 $x \in V$ 上是强凸的且在 V 中拥有唯一定义的临界点 $x_*(\lambda)$ ，其在 $\lambda \in \Lambda$ 是连续可微的. $x_*(\lambda)$ 是 $L(\cdot, \lambda)$ 的非退化局部极小点；
- (ii) 函数

$$\underline{L}(\lambda) = L(x_*(\lambda), \lambda) = \min_{x \in V} L(x, \lambda)$$

在 Λ 上是 C^2 -光滑和凹的，

$$\underline{L}'(\lambda) = h(x_*(\lambda)),$$

并且 λ_* 是 $\underline{L}(\lambda)$ 在 Λ 上的非退化极大点.

这样，我们有

$$\begin{aligned} \lambda^* &= \operatorname{argmax}_{\text{null}} \underline{L}(\lambda) := \min_{x \in V} L(x, \lambda) \\ x_* &= \operatorname{argmin}_{x \in V} L(x, \lambda^*). \end{aligned}$$

因此，我们可以利用一阶无约束最小化方法求出 $\underline{L}(\lambda)$ 在 $\lambda \in \Lambda$ 上的最大点 λ^* ，进而求解问题(EP). 在关于 $\underline{L}(\lambda)$ 的最大化方法中，通过求解辅助无约束问题

$$x_*(\lambda) = \operatorname{argmin}_{x \in V} L(x, \lambda)$$

可通过 $\underline{L}(\lambda) = L(x_*(\lambda), \lambda)$ 和 $\underline{L}'(\lambda) = h(x_*(\lambda))$ 获得 $\underline{L}(\lambda)$ 的一阶信息.

请注意在此方案中，没有“巨大的参数”！然而如何确保是幸运的？通过惩罚来凸化可以确保是幸运的. 对于每一个 $\rho \geq 0$ ，观察感兴趣的问题(EP)，它恰好等价于

$$\min_x \left\{ f_\rho(x) = f(x) + \frac{\rho}{2} \|h(x)\|_2^2 : h_i(x) = 0, i = 1, \dots, k \right\}. \quad (\mathbf{P}_\rho)$$

结果发现：如果 x_* 是问题(EP)的非退化局部最优解且 ρ 足够大，那么 x_* 是问题(\mathbf{P}_ρ)的局部最优并且“幸运”的解. 从而，只要 ρ 适当地大，我们即可将提到的“原始-对偶”方案应用于(\mathbf{P}_ρ)以便求解问题(EP)！

请注意，尽管在新的方案中，我们需要惩罚参数“足够大”，但与直接惩罚方案相比，我们仍然具有优势：在后者中，随着要求求解(EP)的非精确度 ϵ 趋于0，对应所需的 ρ 像 $O(1/\epsilon)$ 那样趋于 ∞ ，而在我们的新方案中，单个“足够大”的 ρ 值就可以了！

解释断言： 记(\mathbf{P}_ρ)的拉格朗日函数

$$L_\rho(x, \lambda) = f(x) + \frac{\rho}{2} \|h(x)\|_2^2 + \sum_i \lambda_i h_i(x),$$

因此(EP)的拉格朗日函数是 $L(x, \lambda)$. 给定(EP) 的非退化局部最优解 x_* ，设 λ^* 是对

应的拉格朗日乘子. 我们有

$$\begin{aligned}\nabla_x L_\rho(x_*, \lambda^*) &= \nabla_x L(x_*, \lambda^*) + \rho \sum_i h_i(x_*) h'_i(x_*) = \nabla_x L(x_*, \lambda^*) = 0 \\ \text{null } \nabla_x^2 L_\rho(x_*, \lambda^*) &= \nabla_x^2 L(x_*, \lambda^*) + \rho \sum_i h_i(x_*) h''_i(x_*) + \rho \sum_i h'_i(x_*) [h'_i(x_*)]^T \\ &= \nabla_x^2 L(x_*, \rho^*) + \rho H^T H\end{aligned}$$

其中

$$H = \begin{bmatrix} [h'_1(x_*)]^T \\ \vdots \\ [h'_k(x_*)]^T \end{bmatrix} \quad (12.5)$$

与 $h'_i(x_*)$, $i = 1, \dots, k$, 正交的方向 d 恰好是使得 $Hd = 0$ 的方向. 因此, 对于所有 $\rho \geq 0$, 问题 (P_ρ) 在 x_* 的二阶充分最优性条件:

$$Hd = 0, d \neq 0 \Rightarrow d^T \nabla_x^2 L_\rho(x_*, \lambda^*) d > 0$$

成立. 进一步, 为了证明对于大的 ρ , x^* 是“幸运”的解, 我们需要如下的线性代数事实.

命题12.1. 设 Q 是 $n \times n$ 对称矩阵, H 是 $k \times n$ 矩阵. 假设 Q 在 H 的零空间上是正定矩阵:

$$d \neq 0, Hd = 0 \Rightarrow d^T Q d > 0.$$

那么对于所有足够大的 ρ , 矩阵 $Q + \rho H^T H$ 是正定的.

证明. 用反证法. 假设存在序列 $\rho_i \rightarrow \infty$ 和 d_i , $\|d_i\|_2 = 1$ 满足:

$$d_i^T [Q + \rho_i H^T H] d_i \leq 0, \forall i.$$

因为 $\{d_i\}$ 是有界序列, 必有收敛子列. 不妨设 $d_i \rightarrow d$, $i \rightarrow \infty$. 将 d_i 分解为它分别在 $\text{Null}(H)$ 和 $[\text{Null}(H)]^\perp$ 中的投影之和, 即 $d_i = h_i + h_i^\perp$. 类似地有 $d = h + h^\perp$. 那么

$$d_i^T H^T H d_i = \|H d_i\|_2^2 = \|H h_i^\perp\|_2^2 \rightarrow \|H h^\perp\|_2^2$$

从而得

$$0 \geq d_i^T [Q + \rho_i H^T H] d_i = \underbrace{d_i^T Q d_i}_{\rightarrow d^T Q d} + \rho_i \underbrace{\|H h_i^\perp\|_2^2}_{\rightarrow \|H h^\perp\|_2^2}. \quad (12.6)$$

如果 $h^\perp \neq 0$, 那么 $\|H h^\perp\|_2 > 0$ (因为 $h^\perp \notin \text{Null}(H)$). 从而当 $i \rightarrow \infty$ 时 (12.6) 的右边趋于 $+\infty$. 这是不可能的. 因此, $h^\perp = 0$. 但是这样导致 $0 \neq d \in \text{Null}(H)$, 因而 $d^T Q d > 0$. 因此 (12.6) 右边对于大的 i 是正的, 这也是不可能的. \square

12.2 整合起来：增广拉格朗日方案

现在，考虑求解 (P_ρ) 的Lagrange对偶问题，即

$$\max_{\lambda} \underline{L}_\rho(\lambda) \quad (D_\rho)$$

其中

$$\underline{L}_\rho(\lambda) = \min_x L_\rho(x, \lambda).$$

一般的增广拉格朗日策略. 对于给定的 ρ 值，用无约束极小化问题的一阶方法求解对偶问题 (D_ρ) ，其中通过求解辅助问题

$$x_*(\lambda) = \operatorname{argmin}_x L_\rho(x, \lambda), \quad (P_\lambda)$$

由关系

$$\underline{L}_\rho(\lambda) = L_\rho(x_*(\lambda), \lambda) \text{ 和 } \underline{L}'_\rho(\lambda) = h(x_*(\lambda))$$

得到 (D_ρ) 的一阶信息. **请注意**如果 ρ 足够大，并且将 (P_λ) 和 (D_ρ) 中的优化问题分别限制在 (P_ρ) 的非退化局部解 x_* 的适当凸邻域和对应格朗日乘子向量 λ^* 的适当邻域内，则 (D_ρ) 的目标函数是凹的和 C^2 的，并且 λ^* 是 (D_ρ) 的非退化解； (P_λ) 的目标函数是凸的和 C^2 的，且 $x_*(\lambda) = \operatorname{argmin}_x L_\rho(x, \lambda)$ 是 (P_λ) 的非退化局部解；随着针对 (D_ρ) 的“主方法”收敛到 λ^* 时，对应的原始迭代 $x_*(\lambda)$ 收敛到 x_* .

实现问题：关于求解辅助问题 (P_λ) ，如果能得到二阶信息，最好的选择是线搜索牛顿法或线搜索修正牛顿法；否则，可以使用拟牛顿法、共轭梯度法等. 至于求解主问题 (D_ρ) ，令人惊讶的是，这里选择最简单的恒定步长梯度上升方法，即

$$\lambda^t = \lambda^{t-1} + \rho \underline{L}'_\rho(\lambda^{t-1}) = \lambda^{t-1} + \rho h(x^{t-1}) \quad (12.7)$$

其中 x^{t-1} 是 $L_\rho(x, \lambda^{t-1})$ 关于 x 的(近似)极小点，即

$$x^{t-1} = \operatorname{argmin}_x L_\rho(x, \lambda^{t-1}). \quad (12.8)$$

这里给出方法(12.4)-(12.8)的动机：可以得到

$$0 \approx \nabla_x L_\rho(x^{t-1}, \lambda^{t-1}) = f'(x^{t-1}) + \sum_i [\lambda_i^{t-1} + \rho h_i(x^{t-1})] h'_i(x^{t-1})$$

这看起来与KKT条件

$$0 = f'(x_*) + \sum_i \lambda_i^* h'_i(x_*)$$

是类似的. 下面给出该方法的**依据**. 直接计算表明

$$\Psi_\rho \equiv \nabla_\lambda^2 \underline{L}_\rho(\lambda^*) = -H[\nabla_x^2 L(x_*, \lambda^*) + \rho H^T H]^{-1} H^T$$

其中 H 的定义见(12.5). 从这里, 当 $\rho \rightarrow \infty$ 时, $-\rho\Psi_\rho \rightarrow I$. 所以, 当 ρ 足够大且(12.4)-(12.8)中的起始点 λ_0 足够接近 λ^* , (12.4)-(12.8)确保 λ^t 线性收敛到 λ^* , 且随着 $\rho \rightarrow +\infty$, 对应的收敛因子趋于0. 的确, (12.4)-(12.8)的渐近行为就好像 $\underline{L}_\rho(\lambda)$ 是二次函数

$$\Phi(\lambda) = \text{const} + \frac{1}{2}(\lambda - \lambda^*)^T \Psi_\rho(\lambda - \lambda^*),$$

并且我们通过梯度上升 $\lambda \mapsto \lambda + \rho\Phi'(\lambda)$ 来极大化这个函数. 这个循环是

$$\lambda^t - \lambda^* = \underbrace{(I + \rho\Psi_\rho)}_{\rightarrow 0, \rho \rightarrow \infty}(\lambda^{t-1} - \lambda^*).$$

最后讨论如何调整惩罚参数. 当 ρ “足够大”, 以便(12.4)-(12.8)以合适的收敛比线性收敛时, $\|\underline{L}'_\rho(\lambda^t)\|_2 = \|h(x^t)\|_2$ 应该以本质上相同的速度线性收敛到0. 因此, 可以使用 $\|h(\cdot)\|_2$ 的进展控制 ρ , 比如当

$$\|h(x^t)\|_2 \leq 0.25\|h(x^{t-1})\|_2$$

时, 保持 ρ 的当前值不变, 否则将惩罚增加10倍. 并用新的 ρ 值重新计算 x^t .

12.3 纳入不等式约束

给定一般的约束问题

$$\min_x \{f(x) : h_i = 0, i \leq m, g_j(x) \leq 0, j \leq k\}, \quad (\text{P})$$

可以把它等价地变换成等式约束问题

$$\min_{x,s} \{f(x) : h_i = 0, i \leq m, g_j(x) + s_j^2 = 0, j \leq k\}, \quad (\text{P}')$$

并将增广拉格朗日策略应用于这个重新表述的问题, 从而得到增广拉格朗日函数

$$L_\rho(x, s; \lambda, \mu) = f(x) + \sum_i \lambda_i h_i(x) + \sum_j \mu_j [g_j(x) + s_j^2] + \frac{\rho}{2} \left[\sum_i h_i^2(x) + \sum_j [g_j(x) + s_j^2]^2 \right]$$

相应的对偶问题是

$$\max_{\lambda, \mu} \{ \underline{L}_\rho(\lambda, \mu) = \min_{x, s} L_\rho(x, s; \lambda, \mu) \}.$$

对偶目标是关于 x, s 极小化得到的最优值, 我们可以关于 s 极小化, 将得到的解析解代入目标函数, 得到

$$\underline{L}_\rho(\lambda, \mu) = \min_x \left\{ f(x) + \frac{\rho}{2} \sum_{j=1}^k (g_j(x) + \frac{\mu_j}{\rho})_+^2 + \sum_{i=1}^m \lambda_i h_i(x) + \frac{\rho}{2} \sum_{i=1}^m h_i(x)^2 \right\} - \sum_{j=1}^k \frac{\mu_j^2}{2\rho},$$

其中 $a_+ = \max\{0, a\}$. 从而增广拉格朗日方案中需要求解的辅助问题是关于原始设计变量的!

等式约束问题的增广拉格朗日策略的理论分析是基于“我们正在尝试逼近非退化的局部解”的假定. 当我们将不等式约束问题化归成等式约束问题时, 我们保持了局部解的非退化性质吗? 答案是肯定的!

定理12.2. 设 x_* 是问题(P)的非退化局部解. 那么点

$$(x_*, s^*) : s_j^* = \sqrt{-g_j(x_*)}, j = 1, \dots, m$$

是问题(P')的非退化局部解.

12.4 凸情况：增广拉格朗日法

考虑凸优化问题

$$\min_x \{f(x) : g_j(x) \leq 0, j = 1, \dots, m\} \quad (\text{CP})$$

其中 f, g_j 在 \mathbb{R}^n 上是凸的且是 C^2 的. 假设问题(CP)是可解的且满足Slater条件

$$\exists \bar{x} : g_j(\bar{x}) < 0, j = 1, \dots, m.$$

在凸的情形下, 由于拉格朗日对偶定理, 前面的局部考虑可被全局化.

定理12.3. 设(CP)是凸的、可解的且满足Slater条件. 那么对偶问题

$$\max_{\lambda \geq 0} \underline{L}(\lambda) \equiv \min_x \underbrace{\left[f(x) + \sum_j \lambda_j g_j(x) \right]}_{L(x, \lambda)} \quad (\text{DP})$$

有如下性质:

- (i) 对偶目标 \underline{L} 是凹的.
- (ii) 问题(DP)是可解的.
- (iii) 对于问题(DP)的每一个最优解 λ^* , 问题(CP)所有的最优解均属于集合 $\operatorname{argmin}_x L(x, \lambda^*)$.

该定理有如下含义. 有时能够显式构造(DP)(比如, 在线性规划, 线性约束二次规划和几何规划中). 在这些情况下, 能够通过求解(DP), 然后由(DP)的解恢复(CP)的解.

在一般情况, 可以用一阶方法数值求解(DP), 从而将具有一般凸约束的问题化归为具有简单线性约束的问题. 为了用数值方法求解(DP), 我们应该能够计算 \underline{L} 的一阶信息. 这可以通过求解辅助问题

$$x_* = x_*(\lambda) = \operatorname{argmin}_x L(x, \lambda) \equiv f(x) + \sum_j \lambda_j g_j(x), \quad (12.9)$$

并基于

$$\underline{L}(\lambda) = L(x_*(\lambda), \lambda), \quad \underline{L}'(\lambda) = g(x_*(\lambda))$$

来完成. 请注意(12.9)是一个目标函数光滑的凸的无约束规划.

实施该想法时会碰到两个潜在困难: (1) 在一些点处 $\underline{L}(\cdot)$ 可能是 $-\infty$; 如何求解(DP)? (2) 找到 λ^* 之后, 如何恢复(CP)的最优解? 可能发生的是, 集合 $\operatorname{argmin}_x L(x, \lambda^*)$ 要比(CP)的最优解集合更大!

例12.1 (直接求解对偶问题的困难, LP). 原始凸优化:

$$\min_x \{c^T x : Ax - b \leq 0\}.$$

这里

$$\underline{L}(\lambda) = \min_x [c^T x + (A^T \lambda)^T x - b^T \lambda] = \begin{cases} -b^T \lambda, & A^T \lambda + c = 0 \\ \text{null} & \\ -\infty, & \text{否则} \end{cases}$$

这时如何求解对应的对偶问题(DP)? 同时, 对于每个 λ , 函数 $L(x, \lambda)$ 关于 x 是线性的; 因此, $\operatorname{argmin}_x L(x, \lambda)$ 是 \emptyset 或者 \mathbb{R}^n . 这时, 如何恢复 x_* ?

我们观察到上面提到的两个困难均来自辅助问题(12.9)的解可能不存在/不唯一. 事实上, 如果在某特定集合 Λ 上(12.9)的解 $x_*(\lambda)$ 存在、唯一且关于 λ 是连续的, 那么由于

$$\underline{L}(\lambda) = L(x_*(\lambda), \lambda), \quad \underline{L}'(\lambda) = g(x_*(\lambda)),$$

从而 $\underline{L}(\lambda)$ 在 Λ 上有限且连续可微. 除此之外, 如果 $\lambda^* = \operatorname{argmax}_{\lambda \geq 0} \underline{L}(\lambda)$ 属于 Λ , 那么由 λ^* 恢复(CP)的最优解就没有问题了.

例12.2. 假设函数

$$r(x) = f(x) + \sum_{j=1}^k g_j(x)$$

是局部强凸的($r''(x) \succ 0, \forall x$)并且使得

$$r(x)/\|x\|_2 \rightarrow \infty, \|x\|_2 \rightarrow \infty.$$

那么对于集合 $\Lambda = \{\lambda > 0\}$ 中的 λ , $x_*(\lambda)$ 存在, 并且是唯一的和连续的. 当 f 自身是局部强凸的并且当 $\|x\|_2 \rightarrow \infty$ 时有 $f(x)/\|x\|_2 \rightarrow \infty$, 那么结论在 $\Lambda = \{\lambda \geq 0\}$ 上成立.

在增广拉格朗日策略中, 为了确保

$$r(\cdot) = f(x) + \text{sum of constraints}$$

的局部强凸性，我们通过将原始问题(CP) 重新表述为等价问题

$$\min_x \{f(x) : \theta_j(g_j(x)) \leq 0, j = 1, \dots, m\}, \quad (\text{CP}')$$

其中 $\theta_j(\cdot)$ 是增的强凸光滑函数，且满足规范化条件

$$\theta_j(0) = 0, \theta'_j(0) = 1.$$

关于原始问题(CP)和等价表述(CP')，我们有以下事实：(i) 问题(CP)是凸的并且等价于问题(CP'). (ii) 因为

$$\nabla_x [f(x) + \sum_j \lambda_j^* g_j(x)] = 0 \quad \& \quad \lambda_j^* g_j(x) = 0, \forall j$$

当且仅当

$$\nabla_x [f(x) + \sum_j \lambda_j^* \theta_j(g_j(x))] = 0 \quad \& \quad \lambda_j^* \theta_j(g_j(x)) = 0, \forall j.$$

所以问题(CP)的最优拉格朗日乘子和问题(CP')的是相同的. (iii) 在温和的正则性假设下，

$$r(x) = f(x) + \sum_j \theta_j(g_j(x))$$

是局部强凸的，并且当 $\|x\|_2 \rightarrow \infty$ 时有 $r(x)/\|x\|_2 \rightarrow \infty$.

用粗略介绍的策略，我们从(CP)的经典拉格朗日函数

$$L(x, \lambda) = f(x) + \sum_j \lambda_j g_j(x)$$

转到等价问题(CP')的增广拉格朗日函数

$$\tilde{L}(x, \lambda) = f(x) + \sum_j \lambda_j \theta_j(g_j(x)),$$

由其得到的对偶问题

$$\max_{\lambda \geq 0} \tilde{L}(\lambda) \equiv \max_{\lambda \geq 0} \min_x \tilde{L}(x, \lambda) \quad (\text{DP}')$$

比(CP)的惯常拉格朗日对偶更适合于数值求解和恢复(CP) 的解.

惩罚机制进一步增加了灵活性. 带有惩罚参数的增广拉格朗日函数³

$$\tilde{L}(x, \lambda) = f(x) + \sum_j \lambda_j \rho^{-1} \theta_j(\rho g_j(x))$$

³对于给定的正数 μ_j ，这里

$$\theta_j(s) = \frac{\mu_j}{2} \left[\left(\frac{s}{\mu_j} + 1 \right)_+^2 - 1 \right],$$

其中 $a_+ = \max\{0, a\}$.

相等于“重新缩放”，这里将 $\theta_j(s)$ 取成与惩罚参数 ρ 相关的，即

$$\theta_j^{(\rho)}(s) = \rho^{-1}\theta_j(\rho s).$$

因此， ρ 越大，用来求解(DP')的一阶方法的收敛速度越快，同时辅助问题

$$\min_x [f(x) + \sum_j \lambda_j \rho^{-1} \theta_j(\rho g_j(x))]$$

变得更难于求解.

13 约束最小化方法：逐步二次规划

逐步二次规划法(Sequential Quadratic Programming, SQP)被认为是求解具有光滑目标和约束的普通型优化问题的最有效技术, 它通过牛顿型迭代过程直接解决问题的KKT系统. 具体地, 考虑等式约束问题(EP), 它的Lagrange函数

$$L(x, \lambda) = f(x) + h^T(x)\lambda,$$

对应的KKT系统是

$$\begin{aligned} \nabla_x L(x, \lambda) &\equiv f'(x) + [h'(x)]^T \lambda = 0 \\ \nabla_\lambda L(x, \lambda) &\equiv h(x) = 0 \end{aligned} \quad (13.1)$$

对于(EP)的每个正则(即梯度 $\{h'_i(x_*)\}_{i=1}^k$ 是线性无关的)局部最优解 x_* , 都可以适当选择 $\lambda = \lambda^*$ 以扩展到成(13.1)的解. 而(13.1)是具有 $n + k$ 个方程和 $n + k$ 个未知数的非线性方程组. 我们可以尝试用牛顿法求解该系统.

13.1 牛顿法求解非线性方程组

求解具有 N 个未知数的 N 个非线性方程的系统

$$P(u) \equiv (p_1(u), \dots, p_N(u))^T = 0,$$

其中 $p_i : \mathbb{R}^N \mapsto \mathbb{R}$ 是 C^1 实值函数. 按如下方式迭代:

给定当前迭代 \bar{u} , 我们在迭代 \bar{u} 处将系统线性化, 从而得线性方程组

$$P(\bar{u}) + P'(\bar{u})(u - \bar{u}) \equiv \begin{bmatrix} p_1(\bar{u}) + [p'_1(\bar{u})]^T(u - \bar{u}) \\ \vdots \\ p_N(\bar{u}) + [p'_N(\bar{u})]^T(u - \bar{u}) \end{bmatrix} = 0$$

假设 $N \times N$ 矩阵 $P'(\bar{u})$ 是非奇异的, 求解线性化得到的方程组, 由此得新迭代

$$\bar{u}^+ = \bar{u} - \underbrace{[P'(\bar{u})]^{-1}P(\bar{u})}_{\text{Newton位移}}$$

$$\bar{u} \mapsto \bar{u}^+ = \bar{u} - [P'(\bar{u})]^{-1}P(\bar{u}) \quad (\text{N})$$

请注意无约束最小化的基本牛顿法不是别的, 就是将上述过程应用于费马方程

$$P(x) \equiv \nabla f(x) = 0.$$

像在优化情况中那样, 牛顿法具有快速局部收敛性:

定理13.1. 设 $u_* \in \mathbb{R}^N$ 是正方形非线性方程组

$$P(u) = 0$$

的解, 其中 P 的分量在 u_* 的邻域上是 C^1 的. 假设 u_* 是非退化的, 即 $\text{Det}(P'(u_*)) \neq 0$, 那么从足够接近 u_* 开始的牛顿法(N)超线性收敛到 u_* . 此外, 如果 P 在 u_* 的邻域是 C^2 的, 那么上述收敛是二次的.

考虑对KKT系统(13.1)应用如上所述策略. 在具体的行动之前, 首先应该回答如下重要问题: 何时KKT点 (x_*, λ^*) 是KKT系统(13.1)的非退化解. 设

$$P(x, \lambda) = \nabla_{x, \lambda} L(x, \lambda) = \begin{bmatrix} \nabla_x L(x, \lambda) \equiv f'(x) + [h'(x)]^T \lambda \\ \nabla_\lambda L(x, \lambda) \equiv h(x) \end{bmatrix}$$

那么(13.1)即

$$P(x, \lambda) = 0.$$

请注意

$$P'(x, \lambda) = \begin{bmatrix} \nabla_x^2 L(x, \lambda) & [h'(x)]^T \\ h'(x) & 0 \end{bmatrix}.$$

定理13.2. 设 x_* 是(EP)的非退化局部解并且 λ^* 是对应的拉格朗日乘子. 那么 (x_*, λ^*) 是KKT系统(13.1)的非退化解, 即矩阵 $P' \equiv P'(x_*, \lambda^*)$ 是非奇异的.

证明. 置 $Q = \nabla_x^2 L(x_*, \lambda^*)$, $H = h'(x_*)$, 得到

$$P' = \begin{bmatrix} Q & H^T \\ H & 0 \end{bmatrix}.$$

首先由 x^* 满足正则性知 H 的行是线性无关的; 其次, 由 x^* 是局部非退化解的事实(这里指满足二阶必要条件), 知道当 $d \neq 0, Hd = 0$ 时有 $d^T Q d > 0$.

如果

$$0 = P' \begin{bmatrix} d \\ g \end{bmatrix} \equiv \begin{bmatrix} Qd + H^T g \\ Hd \end{bmatrix}$$

得到

$$Hd = 0 \tag{13.2}$$

和

$$0 = Qd + H^T g \tag{13.3}$$

由(13.2)和(13.3)得到

$$d^T Qd + (Hd)^T g = d^T Qd = 0.$$

正如我们知道的那样, 这种情况是可能的当且仅当 $d = 0$. 将 $d = 0$ 代入(13.3), 得 $H^T g = 0$; 再由 H 的行是线性无关的得 $g = 0$. \square

13.2 牛顿位移的结构和解释

在讨论的情况下, 牛顿系统

$$P'(u)\Delta = -P(u) \quad [\Delta = u^+ - u]$$

变成

$$\begin{aligned} [\nabla_x^2 L(\bar{x}, \bar{\lambda})]\Delta x + [h'(\bar{x})]^T \Delta \lambda &= -f'(\bar{x}) - [h'(\bar{x})]\bar{\lambda} \\ [h'(\bar{x})]\Delta x &= -h(\bar{x}) \end{aligned}$$

其中 $(\bar{x}, \bar{\lambda})$ 是当前迭代. 表述成关于变量 Δx 和 $\lambda^+ = \bar{\lambda} + \Delta \lambda$ 的系统, 其变成

$$\begin{aligned} [\nabla_x^2 L(\bar{x}, \bar{\lambda})]\Delta x + [h'(\bar{x})]^T \lambda^+ &= -f'(\bar{x}) \\ h'(\bar{x})\Delta x &= -h(\bar{x}). \end{aligned} \tag{N}$$

解释. 假设我们知道最优拉格朗日乘子 λ^* 和 x_* 处可行曲面的切平面 T :

$$T = \{y = x_* + \Delta x : h'(x_*)\Delta x + h(x_*) = 0\}.$$

由于 $\nabla_x^2 L(x_*, \lambda^*)$ 在 T 上是正定的, 并且 $\nabla_x L(x_*, \lambda^*)$ 与 T 正交, x_* 是 $L(x, \lambda^*)$ 在 $x \in T$ 上的局部极小点, 且能够通过将牛顿极小化方法应用于函数 $L(x, \lambda^*)$ 在 T 上的限制来找到 x_* :

$$\bar{x} \mapsto \bar{x} + \operatorname{argmin}_{\Delta x: \bar{x} + \Delta x \in T} \left[L(\bar{x}, \lambda^*) + \Delta x^T \nabla_x L(\bar{x}, \lambda^*) + \frac{1}{2} \Delta x^T \nabla_x^2 L(\bar{x}, \lambda^*) \Delta x \right].$$

实际上, 我们既不知道 λ^* , 也不知道 T , 仅知道 x_* 和 λ^* 的当前近似 $\bar{x}, \bar{\lambda}$. 可以使用这些近似来逼近上述策略: 给定 \bar{x} , 用平面

$$\bar{T} = \{y = \bar{x} + \Delta x : h'(\bar{x})\Delta x + h(\bar{x}) = 0\}$$

逼近 T . 我们应用上面提到的步, 将其中的 λ^* 替换为 $\bar{\lambda}$, T 替换为 \bar{T} :

$$\bar{x} \mapsto \bar{x} + \operatorname{argmin}_{\Delta x: \bar{x} + \Delta x \in \bar{T}} \left[L(\bar{x}, \bar{\lambda}) + \Delta x^T \nabla_x L(\bar{x}, \bar{\lambda}) + \frac{1}{2} \Delta x^T \nabla_x^2 L(\bar{x}, \bar{\lambda}) \Delta x \right]. \tag{A}$$

请注意, 根据事实: 对于 $\bar{x} + \Delta x \in \bar{T}$ 有

$$\Delta x^T \nabla_x L(\bar{x}, \bar{\lambda}) = \Delta x^T f'(\bar{x}) + \bar{\lambda}^T h'(\bar{x}) \Delta x = \Delta x^T f'(\bar{x}) - \bar{\lambda}^T h(\bar{x}).$$

从而可将步简化成

$$\bar{x} \mapsto \bar{x} + \underset{\Delta x: \bar{x} + \Delta x \in \bar{T}}{\operatorname{argmin}} \left[f(\bar{x}) + \Delta x^T f'(\bar{x}) + \frac{1}{2} \Delta x^T \nabla_x^2 L(\bar{x}, \lambda^*) \Delta x \right]. \quad (\text{B})$$

由此可推出：当 $\bar{x} + \Delta x \in \bar{T}$ 时，我们在(A)和(B)中最小化的关于 Δx 的函数仅相差一个常数。综上，我们得到以下策略：

已知等式约束问题(EP)的非退化KKT点 (x_*, λ^*) 的近似 $(\bar{x}, \bar{\lambda})$ 。求解辅助二次规划

$$\Delta x_* = \underset{\Delta x}{\operatorname{argmin}} \left\{ f(\bar{x}) + \Delta x^T f'(\bar{x}) + \frac{1}{2} \Delta x^T \nabla_x^2 L(\bar{x}, \bar{\lambda}) \Delta x : h(\bar{x}) + h'(\bar{x}) \Delta x = 0 \right\} \quad (\text{QP})$$

用 $\bar{x} + \Delta x_*$ 代替 \bar{x} 。

请注意倘若在可行平面 $\bar{T} = \{\Delta x : h(\bar{x}) + h'(\bar{x}) \Delta x = 0\}$ 上， $\nabla^2 L(\bar{x}, \bar{\lambda})$ 是正定的(当 $(\bar{x}, \bar{\lambda})$ 足够接近 (x_*, λ^*) 时确实如此)，那么(QP)是良好的线性代数问题。这样，应用于(EP)KKT系统的牛顿迭代是

$$(\bar{x}, \bar{\lambda}) \mapsto (\bar{x}^+ = \bar{x} + \Delta x, \lambda^+),$$

其中 $(\Delta x, \lambda^+)$ 满足牛顿系统(N)。下面考虑相关二次规划(QP)与牛顿系统(N)的关系。

定理13.3 (重要观察). 假设正在考虑的牛顿系统(N)中的系数矩阵是非奇异的。那么由(N)确定的牛顿位移 Δx 是二次规划(QP)的唯一KKT点，而 λ^+ 是对应的拉格朗日乘子向量。

证明. 设 z 是(QP)的KKT点，并且 μ 是对应的拉格朗日乘子向量。那么(QP)的KKT系统是

$$\begin{aligned} \underset{\text{null}}{f'(\bar{x}) + \nabla_x^2 L(\bar{x}, \bar{\lambda}) z + [h'(\bar{x})]^T \mu} &= 0 \\ h'(\bar{x}) z &= -h(\bar{x}). \end{aligned}$$

这正是(N)中的方程。由于系统(N)中的矩阵是非奇异的，因此有 $z = \Delta x$ 和 $\mu = \lambda^+$ 。

□

综上，应用于(EP)的KKT系统的牛顿法按如下方式工作：

已知当前迭代 $(\bar{x}, \bar{\lambda})$ ，我们将约束线性化，由此得到“近似可行集”

$$\bar{T} = \{\bar{x} + \Delta x : h'(\bar{x}) \Delta x = -h(\bar{x})\},$$

并且在这个可行集上最小化二次函数

$$f(\bar{x}) + (x - \bar{x})^T f'(\bar{x}) + \frac{1}{2} (x - \bar{x})^T \nabla_x^2 L(\bar{x}, \bar{\lambda}) (x - \bar{x}).$$

所得到的线性等式约束二次问题的解是新的 x -迭代，与此解对应的拉格朗日乘子向量是新的 λ -迭代。需要注意的是辅助二次目标中的二次部分来自于(EP)的拉格朗日函数，而不是来自(EP)的目标！

13.3 一般约束的情况

13.3.1 基本SQP法

可将应用于等式约束问题KKT系统的牛顿方法“基于最优化”的解释推广到一般约束问题

$$\min_x \left\{ f(x) : \begin{array}{l} \text{null} \\ h(x) = (h_1(x), \dots, h_k(x)) = 0 \\ g(x) = (g_1(x), \dots, g_m(x)) \leq 0 \end{array} \right\}, \quad (\text{P})$$

得到如下基本SQP策略:

置 $L(x; \lambda, \mu) = f(x) + h^T(x)\lambda + g^T(x)\mu$. 已知(P)的非退化局部解 x_* 以及对应的最优拉格朗日乘子 λ^*, μ^* 的当前近似 $x_t, \mu_t \geq 0, \lambda_t$, 求解辅助线性约束二次问题

$$\min_{\Delta x} \left\{ f(x_t) + \Delta x^T f'(x_t) + \frac{1}{2} \Delta x^T \nabla_x^2 L(x_t; \lambda_t, \mu_t) \Delta x : \begin{array}{l} \text{null} \\ h'(x_t) \Delta x = -h(x_t) \\ g'(x_t) \Delta x \leq -g(x_t) \end{array} \right\} \quad (\text{QP}_t)$$

得到最优解 Δx_* , 置 $x_{t+1} = x_t + \Delta x_*$, 并将 λ_{t+1}, μ_{t+1} 定义为 (QP_t) 的最优拉格朗日乘子.

定理13.4. 设 $(x_*; \lambda^*, \mu^*)$ 是(P)的非退化局部最优解及对应的最优拉格朗日乘子. 倘若除始点足够接近 $(x_*; \lambda^*, \mu^*)$, 并且限制只在恰当小的 Δx 的情况下工作, 那么基本SQP方法是适定的, 并且二次收敛于 $(x_*; \lambda^*, \mu^*)$.

存在的困难也不少. 从“全局”角度来看, 要求解的辅助二次问题可能很糟糕(例如, 不可行或无下界). 对等式约束的情况, 当我们靠近非退化的局部解时, 就永远不会发生这种情况. 对一般情况, 即使靠近非退化局部解, 也可能发生这些糟糕的事情. 一种矫正办法是当矩阵 $\nabla_x^2 L(x_t; \lambda_t, \mu_t)$ 在整个空间上不是正定的时, 用正定矩阵 B_t 取代它, 从而得到辅助二次问题

$$\text{null} \min_{\Delta x} \left\{ f(x_t) + \Delta x^T f'(x_t) + \frac{1}{2} \Delta x^T B_t \Delta x : \begin{array}{l} \text{null} \\ h'(x_t) \Delta x = -h(x_t) \\ g'(x_t) \Delta x \leq -g(x_t) \end{array} \right\}. \quad (\text{QP}_t)$$

通过这种修改, 辅助问题是凸的并且可求解的(倘若它们是可行的, 那么当 x_t 靠近(P)的非退化解时的确如此), 并且它有唯一最优解.

13.3.2 确保全局收敛

“矫正的”基本SQP策略具有良好的局部收敛性质; 然而, 它通常不是全局收敛的. 实际上, 在最简单的无约束情况下, SQP变成了基本/修正牛顿法, 除非结合了线性搜索, 否则它不一定是全局收敛的.

为了确保SQP的全局收敛，我们引入线搜索. 在线性搜索方案中，将辅助二次问题(QP_t)的最优解 Δx_* 和相应的拉格朗日乘子 λ^+ ， μ^+ 作为搜索方向，而不是新迭代. 新迭代是

$$\begin{aligned}x_{t+1} &= x_t + \gamma_{t+1} \Delta x_* \\ \text{null } \lambda_{t+1} &= \lambda_t + \gamma_{t+1} (\lambda^+ - \lambda_t) \\ \mu_{t+1} &= \mu_t + \gamma_{t+1} (\mu^+ - \mu_t)\end{aligned}$$

其中 $\gamma_{t+1} > 0$ 是由线搜索确定的步长. 请注意，在(QP_t)中看不到 λ_t 和 μ_t . 然而，作为构建 B_t 时使用的数据，它们隐含地出现在此问题中.

现在的问题是：通过线搜索应该最小化什么？在受约束的情况下，选择原始问题(P)作为线搜索最小化的辅助目标是行不通的. 对于SQP，一个好的辅助目标(“价值函数, merit function”)是

$$M(x) = f(x) + \theta \left[\sum_{i=1}^m |h_i(x)| + \sum_{j=1}^k g_j^+(x) \right],$$

其中 $g_j^+(x) = \max\{0, g_j(x)\}$, $\theta > 0$ 是参数.

事实. 设 x_t 是当前迭代，辅助二次问题(QP_t)中使用的矩阵 B_t 是正定的， Δx 是该问题的解， $\lambda \equiv \lambda_{t+1}$ ， $\mu \equiv \mu_{t+1}$ 是相应的拉格朗日乘子. 假设 θ 足够大：

$$\theta \geq \max\{|\lambda_1|, \dots, |\lambda_k|, \mu_1, \mu_2, \dots, \mu_m\}.$$

那么或者 $\Delta x = 0$ ，那么 x_t 是原始问题(P)的KKT点，或者 $\Delta x \neq 0$ ，那么 Δx 是 $M(\cdot)$ 的下降方向，即对于充分小的 $\gamma > 0$ 有

$$M(x + \gamma \Delta x) < M(x).$$

带价值函数的通用SQP算法如下：

初始化： 选择 $\theta_1 > 0$ 和初始点 x_1 .

步 t： 已知当前迭代 x_t ，选取矩阵 $B_t \succ 0$ ，构造并求解辅助问题(QP_t)以获得最佳的 Δx 和相应的拉格朗日乘子 λ, μ .

如果 $\Delta x = 0$ ，终止： x_t 是原始问题(P)的KKT点，否则进行如下操作：

——检查

$$\theta_t \geq \bar{\theta}_t \equiv \max\{|\lambda_1|, \dots, |\lambda_k|, \mu_1, \dots, \mu_m\}$$

是否成立. 如果是这种情况，置 $\theta_{t+1} = \theta_t$ ，否则置

$$\theta_{t+1} = \max[\bar{\theta}_t, 2\theta_t].$$

——通过旨在搜索射线 $\{x_t + \gamma \Delta x | \gamma \geq 0\}$ 以极小化价值函数

$$M_{t+1}(x) = f(x) + \theta_{t+1} \left[\sum_{i=1}^m |h_i(x)| + \sum_{j=1}^k g_j^+(x) \right]$$

的线搜索找到新迭代:

$$x_{t+1} = x_t + \gamma_{t+1} \Delta x.$$

用 $t+1$ 代替 t , 并且循环.

定理13.5. 设使用价值函数的SQP 算法可以求解一般约束问题(P). 假使(i)存在紧集 $\Omega \subset \mathbb{R}^n$ 使得对于 $x \in \Omega$, 以 Δx 为未知数的线性不等式约束系统

$$S(x) : h'(x)\Delta x = -h(x), g'(x)\Delta x \leq -g(x)$$

的解集 $D(x)$ 非空, 并且每个向量 $\Delta x \in D(x)$ 是系统 $S(x)$ 的正则解; (ii) 算法的轨迹 $\{x_t\}$ 属于 Ω , 并且是无限的(即该方法没有终止于某个精确的KKT点); (iii) 该方法中使用的矩阵 B_t 是一致有界并且一致正定的: 对于所有 t 和某 $0 < c \leq C < \infty$ 有

$$cI \preceq B_t \preceq CI.$$

那么该方法轨迹的所有聚点都是问题(P)的KKT点.

参考文献

- [1] D. Bertsimas, M. Frankovich, and A. Odoni. Optimal Selection of Airport Runway Configurations. *Operations Research*, 59(6):1047-1419, 2011.
- [2] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer-Verlag, Berlin/Heidelberg, 2011.
- [3] J. B. Lasserre. *Moments, Positive polynomials and Their Applications, volume 1 of Imperial College Press Optimization Series*. Imperial College Press, London, United Kingdom, 2009.
- [4] A. S. Lewis and M. L. Overton. Eigenvalue Optimization. *Acta Numerica*, 5:149-190, 1996.
- [5] I. Bárány, *A generalization of Carathéodory's theorem*. *Discrete Mathematics*, 40(2-3):141-152, 1982.
- [6] I. Bárány and S. Onn. *Carathéodory's theorem, colourful and applicable*. In I. Bárány and K. Böröczky, editors, *Intuitive Geometry*, volume 6 of Bolyai Society Mathematical Studies, pages 11-21. Bolyai János Matematikai Társulat (The János Bolyai Mathematical Society), Budapest, Hungary, 1997.
- [7] A. Barvinok. *A Course in Convexity*, volume 54 of *Graduate Studies in Mathematics*, American Mathematical Society, Providence, Rhode Island, 2002.
- [8] S. Boyd and L. Vandenberghe. *Convex Optimization*, Cambridge University Press, Cambridge, 2004. Available online at <http://www.stanford.edu/~boyd/cvxbook/>.
- [9] A. Brøndsted. *An Introduction to Convex Polytopes*, volume 90 of *Graduate Texts in Mathematics*, Springer-Verlag, New York, 1983.
- [10] C. Ding, D. Sun, and K.-C. Toh. An Introduction to a Class of Matrix Cone Programming. *Mathematical Programming, Series A*, 144(1-2): 141-179, 2014.
- [11] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Grundlehren Text Editions. Springer-Verlag, Berlin/Heidelberg, 2001.
- [12] R. A. Horn and C. R. Johnson. *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

- [13] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Chichester, England, revised edition, 1999.
- [14] E. A. Papa Quiroz, L. Mallma Ramirez, and P. R. Oliveira. An Inexact Proximal Method for Quasiconvex Minimization. *European Journal of Operational Research*, 246(3):721-729, 2015.
- [15] R. T. Rockafellar. *Convex Analysis*, Princeton Landmarks in Mathematics and Physics, Princeton University Press, Princeton, New Jersey, 1997.
- [16] H. L. Royden. *Real Analysis*, Macmillan Publishing Company, New York, third edition, 1988.
- [17] A. Ruszczyński. *Nonlinear Optimization*, Princeton University Press, Princeton, New Jersey, 2006.
- [18] G. W. Stewart and J. Sun. *Matrix perturbation Theory*. Academic Press, Boston, 1990.
- [19] G. M. Ziegler. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*, Springer-Verlag, New York, revised first edition, 1995.