标题 **title**

作者 *author*

2023 年 7 月 25 日

前言

目录

第一部分

科学的逻辑

第二部分

信息与数据

第三部分

决策与优化

第六章　凸分析

本章将会建立关于决策与优化的基本理论，这些方法论都是数据驱动的机器学习的基础，他们涉及从数据到建立模型的步骤（即训练）. 优化与分析有着密不可分的联系，所以本章我们会立足优化问题的一些基本事实，建立凸分析理论。凸分析是优化理论的基础.

§6.1　决策与优化的基本原理

6.1.1　统计决策理论

[lhy: 这部分要细化]

我们在前一部分讨论过，数据（或者说信息）的意义总是体现在集合的对象中，我们把我们所关心的集合对象称为（随机）*总体* P。从概率论角度看，总体就是一个概率分布。现在我们从总体 P 中抽取一个*样本* X。这件事情在概率论上意味着我们得到了一个随机变量 X 服从分布 P。拿到样本之后，我们的任务是做出*好的决策*，因此，决策 T 是一个依赖 X 的函数. 比如说，P 是所有大学生的身高，X 是随机抽选一个人测量的身高，我们的决策 T 是估计大学生的平均身高。

“好的决策”指的是函数 T 能够具备某些量化指标. 其中非常常用的一个方法是通过*损失函数*来衡量，它是总体 P 和决策 $T(X)$ 的函数，即 $L(P,T(X))$. 损失函数在不同语境下有不同称呼。例如，在经济学和金融学的风险理论中，损失函数被称为*风险函数*，它意味着个体在面对不确定的环境下所需要面对的风险。而在优化理论中，损失函数往往被称为*目标函数*，表明所要优化的对象。

决策 T 的一种量化指标是最小化期望意义下的损失函数：

$$\min_T \mathbb{E}_{X\sim P}(L(P,T(X))).$$

在经济学中，这一量化指标实际上是 von Neumann 和 Morgenstern *期望效用理论*的具

体体现。这一理论认为，个体在面对不确定的环境时，会选择最大化期望效用的决策；在风险理论的语境下，则是最小化期望风险的的决策。

现在我们考虑一个非常一般的决策任务。假设我们的任务是估计函数 f，但是我们只知道观测到的自变量 X（来自总体 P）以及它的函数值 $Y = f(X)$，我们的决策是函数的估计值 $\hat{f}$. 在机器学习中，f 通常是需要训练的模型. 我们可以写出若干种损失函数：

- 平方（L^2）损失函数：$L(P, T(X)) = (Y - \hat{f}(X))^2$. 使用此损失函数的时候，我们要假定 f 在实数范围取值。
- L^1 损失函数：$L(P, T(X)) = |Y - \hat{f}(X)|$. 使用此损失函数的时候，我们要假定 f 在实数范围取值。
- SVM 损失函数（hinge 损失函数）：$L(P, T(X)) = \max\{0, 1 - Y \cdot \hat{f}(X)\}$. 使用此损失函数的时候，我们一般要假定 $f(X) \in [-1, 1]$。
- 交叉熵损失函数：$L(P, T(X)) = CH(\hat{f}(X), Y)$.

这些损失函数会用在不同的场景之中。通常来说，机器学习中有两类问题：*回归问题*和*分类问题*和。他们两个的区别主要在于回归问题中 f 取值为实数，而且通常随自变量连续变化；而分类问题中 f 只取有限多个值，他们通常被作为标签（比如这张图片是人还是青蛙）使用。在回归问题中，我们通常使用平方损失函数或者 L^1 损失函数；在分类问题中，我们通常使用 SVM 损失函数或者交叉熵损失函数。

6.1.2 优化问题

现在我们从决策过渡到优化。在最简单的决策问题中，我们的目标就是找到某个 x 使得（期望）损失函数 f 最小. 此时，问题的一般形式为：

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & f_i(x) = 0, \quad i = 1, \dots, m, \\ & g_j(x) \le 0, \quad j = 1, \dots, n, \\ & x \in \Omega. \end{aligned}$$

这里，s.t（subject to）之后的内容表明了 x 取值的限制，因此被称为*约束*. 其中 $f_i(x) = 0$ 和 $g_j(x) \le 0$ 被称为*函数约束*，而 $x \in \Omega$ 被称为*集合约束*.

优化的基本任务就是找到 x 最小化损失函数.

根据损失函数 f、约束条件 f_i 和 g_j 的不同性质，我们可以对优化问题进行分类：

- 无约束优化：约束条件 f_i 和 g_j 实际上不存在，即 $m = n = 0$，并且 Ω 是全空间，比如 $\mathbb{R}^n$.
- 有约束优化：至少存在一个约束条件，即 $\min\{m, n\} \geq 1$，或者 Ω 不是全空间。
- 光滑优化：损失函数和约束条件都是可微函数.[1]
- 线性优化：损失函数和约束条件都是线性函数（形如 $a^\top x + b$）.

注. *[lhy: 介绍一下控制理论与优化理论的异同，特别是连续控制和随机优化。]*

下面我们看几个经典的优化例子。

例 6.1 (最小二乘法) 给定矩阵 $A \in \mathbb{R}^{m\times n}$ 和向量 $b \in \mathbb{R}^m$，考虑如下优化问题：

$$\begin{aligned} \min_x \quad & \|Ax - b\|_2^2 \\ s.t. \quad & x \in \mathbb{R}^n. \end{aligned}$$

这个问题被称为**最小二乘法**. 目标函数可以被写为 $(Ax-b)^\top(Ax-b)$，因此最小二乘法是一种典型的无约束光滑优化问题。

最小二乘法的解 x^* 实际上是**投影解**：b 的行向量投影到 A 的列向量形成的线性空间，正好是 Ax^*. *[lhy: 加个图，补全细节]* 因此，求投影也可以被写作一个优化问题.

例 6.2 (线性规划) 给定矩阵 $A \in \mathbb{R}^{m\times n}$ 和向量 $b \in \mathbb{R}^m$，考虑如下优化问题：

$$\begin{aligned} \min_x \quad & c^\top x \\ s.t. \quad & Ax \leq b, \\ & x \geq 0. \end{aligned}$$

这个问题被称为**线性规划**. 目标函数和约束条件都是线性的，因此线性规划是一种典型的线性优化问题. *[lhy: 加个图，补全细节]*

上面两个例子远远不能覆盖所有的优化问题，实际上，相当多的运筹学、机器学习和计算机科学中的问题都可以被视作（非线性）优化问题。

- 运筹学：线性规划、二次规划、整数规划、网络流问题、组合优化问题等.

[1] 光滑这一词的含义在不同的文献中大相径庭，它可以指（连续）可微、连续可微、二次（连续）可微或者无穷次可微。

- 金融学：投资组合优化、风险控制等.
- 机器学习：模型的训练.
- 计算机科学：图论中的极值问题，例如最短路径问题、最小生成树问题等.

因此，如果有一个能够解决通用优化问题的灵丹妙药，那么将会有极其重大的意义. 然而我们后面将会看到，一般的优化是一个难解的问题，更严谨一点说，不存在通用高效算法。

我们先需要明确解优化问题的算法到底是什么. 我们通过给出算法的一些特征来最终明确这一点。大部分优化算法都用了**迭代法**的思想：算法 A 接受一个自变量 x，输出一个自变量 $A(x)$，并把它作为下一轮的输入. 此外，一个算法还应该具有**通用性**，即它必须要能解决一类优化问题 F. 然后，算法具备通用性就意味着它在进行**黑箱优化**：F 必须要给算法提供必要的信息来完成求解，我们将这样的提供机制抽象为**先知**，记为 $\mathcal{O}$. 具体来说算法输入 x 给 $\mathcal{O}$，$\mathcal{O}$ 返回一些信息给算法（例如 x 处的函数值、导数值、Hessian 矩阵）.

接下来的问题是衡量优化算法的性能好坏。我们关注的是最坏情况，也就是说假如我们关注的是问题类 $P \subseteq F$，那么，我们要看的是优化算法在 P 中最差的表现如何。衡量优化算法性能的指标有以下几个：

- 近似程度：我们需要求在允许误差 ϵ 的情况下的近似解. 例如，函数值不大于最优值的 ϵ，或者离最优点距离不超过 ϵ. 考虑近似解是优化问题非常重要的一个想法，因为计算机的表示精度是有限的，我们不可能在所有情况下都求出精确解，所以求近似解是合理的要求。
- 运行时间（收敛速度，复杂度）：找到目标近似解需要调用先知的次数. 通常来说，运行时间会随近似度要求变高而变长，因此运行时间是一个关于近似程度的函数.

> **注.** 通常来说，优化算法的执行过程中还会进行除了调用先知之外的操作，例如进行加减乘除。然而，如果我们把所有这些操作都算入复杂度之中，算法的分析会变得非常困难，因此我们通常只考虑调用先知的次数。这样做的合理性在于，每一次的加减乘除等额外操作，几乎都是因为调用一次先知所以才进行的，因此我们可以把这些额外操作的时间都算入先知调用的时间之中。

有了上面这些准备，我们就可以将“没有万能算法”这一陈述写成定理了。

定理 6.1 (没有免费午餐定理) *[lhy: 给一个证明，以及更加严格的表述]* 设 F 是有限个优化问题的集合，F 上有一个任意的概率分布。考虑一个 F 上的优化算法，记号 d_t 表示 t 轮迭代之后算法产生的点列

$$(x_t(1), y_t(1)), \ldots, (x_t(t), y_t(t)).$$

给定迭代轮数 t，优化问题 f，算法 A，优化过程所产生的点列概率分布为 $P(d_t|f,t,A)$。那么，对任意优化算法 A_1, A_2，

$$\sum_{f\in F} P(d_t|f,t,A_1) = \sum_{f\in F} P(d_t|f,t,A_2).$$

这一定理意味着，对特定的点列，任何算法在所有实例上产生它的概率总和是一样的。

那么，点列和“没有万能算法”有什么样的关系呢？实际上，衡量算法性能的指标和点列有非常密切的联系。比如说，算法花了 k 步找到一个 ϵ-近似解，用点列的语言来说就是算法迭代产生的点列，长度至多是 k 并且最后一个点距离最优解距离不大于 ϵ. 粗略地说，任意点列成立的性质意味着任意指标成立的性质。因此，对于任何一类优化问题来说，不论以何种指标来衡量性能，优化算法在某些问题上表现出来的突出性能一定会在另一些问题上被抵消. 没有一个万能的算法可以高效解决所有优化问题！

6.1.3 例子：网格搜索算法

前面对于概念的讨论依然非常抽象，所以下面我们看一个具体的例子，这个例子将会展示从算法分析的角度，优化所关注的主要问题。考虑如下优化问题：

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & x \in [0,1]^n. \end{aligned} \tag{6.1}$$

其中 $f(x)$ 是 Lipschitz 连续函数，即它满足

$$|f(x) - f(y)| \le L \|x - y\|_\infty, \quad \forall x, y \in [0,1]^n.$$

关于优化算法的假设如下。首先，我们可以访问**零阶先知**，即 $\mathcal{O}(x) = f(x)$。其次，优化算法需要去找到 **ϵ-近似解**，即函数值至多比最小值大 ϵ 的解.

> **注.** 在优化中，我们会经常使用词语“零阶”“一阶”等等，所谓的“阶”指的是函数导数阶数，零阶先知指的是我们可以访问函数值，一阶先知指的是我们可以访问一阶导数，以此类推. 后面还会有零阶条件、一阶条件等等，他们的含义类似。

我们考虑一个非常简单的算法，他被称为网格搜索：

- 将 $[0,1]$ 等分成 p 份，$[0,1] = [0,1/p] \cup \ldots [(p-1)/p, 1]$.
- 遍历 $(p+1)^n$ 个格点：

$$x_{(i_1,\ldots,i_n)} = \left(\frac{i_1}{p}, \ldots, \frac{i_n}{p}\right)^{\mathsf{T}},$$

  $i_k \in \{0,1,\ldots,p\}$.
- 对每个格点询问先知得到其函数值，输出函数值最小的一个（记为 $(\bar{x}, f(\bar{x}))$）.

我们对于网格搜索算法问的问题是，它的复杂度如何。也就是说，它需要调用先知多少次才能找到一个 ϵ-近似解？我们从一个引理开始。

引理 6.1 *设 (6.1) 的最优值为 f^*，那么*

$$f(\bar{x}) - f^* \leq \frac{L}{2p}.$$

证明 设 x^* 是最优点，存在一个方格包含 x^*：

$$x_{(i_1,\ldots,i_n)} \leq x^* \leq x_{(i_1+1,\ldots,i_n+1)}.$$

这个方格的长为 $1/p$，所以我们可以选取方格的某个顶点 $\hat{x}$，使得它的每一个轴离 x^* 的距离都不超过 $1/(2p)$.[lhy: 画个图]

于是根据 Lipschitz 条件，

$$f(\bar{x}) - f^* \leq f(\hat{x}) - f(x^*) \leq L \|\hat{x} - x^*\|_\infty \leq \frac{L}{2p}.$$ □

利用这个引理，我们可以证明网格搜索算法的复杂度.

定理 6.2 *网格搜索算法可以找到找到一个 ϵ-近似解，其调用 $\mathcal{O}$ 的次数至多为*

$$\left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 2\right)^n.$$

- 证明：取 $p = \lfloor L/(2\epsilon) \rfloor + 1$，代入引理 6.1 即可.

网格搜索法的运行时间给了优化问题 (6.1) 一个求解时间的**上界**. 然而这个上界维数呈指数关系，通常来说都是不可接受的复杂度. (6.1) 会有更好的算法呢？这就是**下界问题**. 令人惊讶的是，对于这一个问题，我们可以证明网格搜索法是渐近意义下最优的！

定理 6.3 设 $\epsilon < L/2$，任何访问 $\mathcal{O}$ 的算法（零阶算法）找到 (6.1) 的 ϵ-近似解至少需要调用 $\mathcal{O}$

$$\left\lfloor \frac{L}{2\epsilon} \right\rfloor^n$$

次.*[lhy: 改一下表述，看不懂]*

证明 设 $p = \lfloor L/(2\epsilon) \rfloor$，对任意算法 A，我们尝试构造一个函数，使得 A 调用 $\mathcal{O}$ p^n 次时最多找到一个 ϵ-近似解.

构造思路：对任何测试点，使得 $\mathcal{O}$ 总是返回 0，于是，算法 A 只能找到 $f = 0$ 的解 $\bar{x}$. 注意到算法只能根据先知的返回来进行操作，因此我们先假定这样的函数存在。[lhy: 改一下，读不懂]。

根据鸽巢原理，网格中至少有一个长为 $1/p$ 的小方格 B 内部没有包含任何测试点. 假设这个小方格的中心是 x^*，构造 $\bar{f}(x) = \min\{0, L\|x - x^*\|_\infty - \epsilon\}$。容易看出，$\bar{f}$ 是 L-Lipschitz 函数，并且最小值为 $-\epsilon$.

函数 $\bar{f}$ 非零的点只在方格 $B' = \{x \in [0,1]^n : \|x - x^*\|_\infty \le \epsilon/L\}$ 内部. 因为 $1/(2p) \ge \epsilon/L$，所以 $B' \subseteq B$. 所以所有测试点上 $\mathcal{O}$ 都会返回 0，这是一个 ϵ-近似解. 因此 A 通过小于 p^n 次对 $\mathcal{O}$ 的调用最多只能找到 ϵ-近似解. □

以上两个结论分别给出了 (6.1) 问题的上下界，对比他们：

问题的上界：	问题的下界：
$\left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 2\right)^n$	$\left\lfloor \frac{L}{2\epsilon} \right\rfloor^n$

尽管网格搜索是一个很慢的算法，但是我们证明了，在渐近意义下，优化问题 (6.1) 的最优算法就是网格搜索！因此，我们可以说，一般的优化问题是难解的。

当我们聚焦在特定的问题类上，优化问题并不一定是难解的。比如，线性规划可以在关于约束个数和变量个数的多项式时间内解出精确解。然而，现实中大部分重要的问题并不是线性的，因此，我们接下来的关键问题是识别出一类可以快速求解的非线性优化问题，这就是凸函数的意义。

§6.2 凸函数

我们首先看无约束优化，看看什么样的损失函数可以快速求最小值. **梯度下降方法**是最古老也最常用的方法. 梯度下降每步计算函数的导数（梯度），然后朝着负梯度方向移动到下一个点. 与梯度下降算法相关的最小值必要条件是一阶条件。

定理 6.4 (一阶条件) *如果 x^* 是可微函数 f 的局部最小值，那么*

$$f'(x^*) = 0.$$

证明 根据局部最小值的定义，存在 $r > 0$，对于任意 $\|y - x^*\| < r$，$f(y) \geq f(x^*)$. 因此 $f(y) = f(x^*) + \langle f'(x^*), y - x^* \rangle + o(\|y - x^*\|) \geq f(x^*)$. 因此，对任意 $s \in \mathbb{R}^n$，$\langle f'(x^*), s \rangle \geq 0$. 考虑方向 s 和 $-s$ 可得 $\langle f'(x^*), s \rangle = 0$. 由 s 的任意性，$f'(x^*) = 0$. □

现在，从一阶条件出发，我们考虑如下优化函数类 $\mathcal{F}$，满足如下三个假设：

- 假设 1：对任意 $f \in \mathcal{F}$，如果 x 满足一阶条件，那么 x 是 f 的全局最小值点.
- 假设 2：对任意 $f, g \in \mathcal{F}$，$\alpha, \beta \geq 0$，$\alpha f + \beta g \in \mathcal{F}$.
- 假设 3：线性函数 $f(x) = \langle \alpha, x \rangle + b \in \mathcal{F}$.

假设 1 使得利用一阶条件的算法可以找到全局最优解. 假设 2 描述了对 $\mathcal{F}$ 封闭的操作，这样的操作实际上就是要求函数对线性组合封闭。要求系数 α 和 β 非负是为了保证一阶条件得到的确实是最小值而不是最大值。一个例子是，如果 $x^2 \in \mathcal{F}$，并且线性组合不限制非负系数，那么 $-x^2 \in \mathcal{F}$，但是后者一阶条件对应的是最大值而非最小值，这就会与假设 1 矛盾。假设 3 提供了 $\mathcal{F}$ 的基本函数，即线性函数。我们之前说过，线性规划是易解的，所以 $\mathcal{F}$ 至少要包含线性函数。

从这三个假设出发，我们可以给出函数类 $\mathcal{F}$ 的刻画。

固定一个函数 $f \in \mathcal{F}$，一个点 $x \in \mathbb{R}^n$，定义 $\phi(y) = f(y) - \langle f'(x), y \rangle$。根据假设 2 和假设 3，$\phi(y) \in \mathcal{F}$. $\phi'(y)|_{y=x} = f'(x) - f'(x) = 0$，根据假设 1，$x$ 是 ϕ 的全局最小值. 因此，$\phi(y) \geq \phi(x)$，即

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle . \tag{6.2}$$

这一不等式给出了**可微凸函数**的定义：任意 x, y 都满足 (6.2) 的函数. 这一不等式有很强的几何直观，从 x 处做函数 f 的切线，那么切线上的点都在函数下方. 从这个角度来看，凸函数的定义是向下凸的函数. [lhy: 画个图]

非常有趣的是，$\mathcal{F}$ 完全由可微凸函数组成，这一点可以通过下面的定理得到证明。

定理 6.5 *函数 $f \in \mathcal{F}$ 当且仅当 f 是可微凸函数.*

证明 只需验证满足 (6.2) 的函数属于 $\mathcal{F}$.

- 假设 1 令 $f'(x) = 0$ 即得任意 y 都有 $f(y) \geq f(x)$.

- 假设 2 利用内积的双线性性和导数加法公式.
- 假设 3 是平凡的. □

[lhy: 习题：如果 f 是二次可微的，那么他的二阶导数（Hessian 矩阵）$f''(x)$ 和凸函数有何关系？]

[lhy: 给一些凸函数的例子]

从数学的角度来说，给了凸性的定义，下一步任务就是给出保持凸性不变的操作，这样我们可以用基本函数构造出更多的函数。

假设 2 实际上已经给出了一种凸性不变的操作，我们将它写成以下命题：

命题 6.1 *对任意 $f,g \in \mathcal{F}$，$\alpha,\beta \geq 0$，$\alpha f + \beta g \in \mathcal{F}$.*

另一个可以保持凸性的操作是仿射变换可以保持凸性。所谓仿射变换，指的是向量空间 $\mathbb{R}^n$ 到 $\mathbb{R}^m$ 的映射 $x \mapsto Ax + b$，其中 A 是 $m \times n$ 矩阵，$b \in \mathbb{R}^m$. 仿射变换实际上就是线性函数，只是我们用变换的方式来表示它。

命题 6.2 *假设函数 $f : \mathbb{R}^n \to \mathbb{R}$ 属于 $\mathcal{F}$，那么对任意仿射变换 $x \mapsto Ax + b$，$g(x) = f(Ax + b) \in \mathcal{F}$.*

证明 $g'(x) = A^\mathsf{T} f'(Ax + b)$，因此

$$\begin{aligned} g(y) = f(Ay + b) &\geq f(Ax + b) + \langle f'(Ax + b), (Ay + b) - (Ax + b) \rangle \\ &= f(Ax + b) + \langle f'(Ax + b), A(y - x) \rangle \\ &= g(x) + \left\langle A^\mathsf{T} f'(Ax + b), y - x \right\rangle \\ &= g(x) + \langle g'(x), y - x \rangle . \end{aligned}$$

□

更多保持凸性不变的操作，见习题。[lhy: 习题：给出更多保持凸性不变的操作]

凸函数的一个重要性质是 Jensen 不等式：

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y). \tag{6.3}$$

Jensen 不等式具有很强的几何解释：画一条 f 的割线，那么 f 的函数图像位于割线上方。实际上，Jensen 不等式给了凸函数一种等价的定义：

定理 6.6 *设 f 是连续可微的函数，那么 f 满足 (6.2) 当且仅当 f 满足 (6.3).*

证明 $\Longrightarrow$：在 (6.2) 中，取 x 为 $\alpha x+(1-\alpha)y$，y 分别取为 x 和 y，如此得到两个不等式，加权求和即得 (6.3).

$\Longleftarrow$：$f(y) \geq (1-\alpha)^{-1}(f(\alpha x+(1-\alpha)y)-\alpha f(x))$

$$= f(x)+(1-\alpha)^{-1}(f(x+(1-\alpha)(y-x))-f(x)).$$

令 $\alpha \to 1$ 即得 (6.2). □

如果函数 f 不是可微的，那么定理 6.6 给了一个凸函数更加本质的定义：

定义 6.1 (凸函数) *函数 f 满足对任意 x,y 成立(6.3)，那么称 f 是*凸函数.

扩展定义之后的凸函数包括了我们之前讲的 L^p（$p=1,2$）损失和 SVM 损失，以及机器学习中用到的大部分损失函数. 在实际情况中，凸函数是一类存在快速收敛算法的函数，例如梯度下降和 Netwon 迭代法. 因此，我们可以说，凸函数类划定了非线性优化中可以快速求解的函数类. 自此，凸性成为了优化中的核心概念，正如 R.T.Rockafellar [?] 所说：

> In fact the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity.

§6.3 凸集

接下来我们考虑约束优化问题：

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

一个自然的问题是，什么样 Ω 会存在快速收敛的算法？我们将看到，凸集将会是这个问题的答案.

6.3.1 基本定义和性质

回忆凸函数的一般定义：任意 $\alpha \in [0,1]$ 和 $x,y \in \mathbb{R}^n$，

$$f(\alpha x+(1-\alpha)y) \leq \alpha f(x)+(1-\alpha)f(y).$$

这里，我们隐含的要求是线段 xy 上的每一点都可以求函数值. 因此，如果我们希望凸函数能够包含在带约束的优化中，一个自然的要求就是对任意 $x,y \in \Omega$，线段 $xy \subseteq \Omega$. 这就是凸集的定义：

定义 6.2 (凸集) 集合 C 被称为**凸集**当且仅当对任意 $x,y \in C$，线段 $\{\alpha x + (1-\alpha)y : \alpha \in [0,1]\} \subseteq C$.

我们来看一些凸集的例子：

例 6.3

- 超平面：$\{x \in \mathbb{R}^n : a^\top x = b\}$，$a \in \mathbb{R}^n$，$b \in \mathbb{R}$.
- 半空间：$\{x \in \mathbb{R}^n : a^\top x \geq b\}$，$a \in \mathbb{R}^n$，$b \in \mathbb{R}$.
- 球：$\{x \in \mathbb{R}^n : \|x - x_0\| \leq r\}$，其中 $\|\cdot\|$ 是任意一种范数.
- 锥：C 是一个锥指的是任意 $x,y \in C$ 和任意 $\alpha,\beta \geq 0$，$\alpha x + \beta y \in C$.

另外一些重要的例子是凸函数诱导的凸集。首先是上图。

定义 6.3 (上图) 函数 f 的**上图**是指集合 $\mathrm{epi}(f) = \{(x,y) \in \mathbb{R}^n \times \mathbb{R} : y \geq f(x)\}$. 直观上说，$\mathrm{epi}(f)$ 是位于函数 f 的图像上方的区域.

[lhy: 画图]

上图揭示了凸集与凸函数的关系：

定理 6.7 上图 $\mathrm{epi}(f)$ 是凸集当且仅当 f 是凸函数.

证明 $\Longrightarrow$：$(x,f(x)),(y,f(y)) \in \mathrm{epi}(f)$，因此 $(\alpha x + (1-\alpha)y, \alpha f(x) + (1-\alpha)f(y)) \in \mathrm{epi}(f)$，所以 $\alpha f(x) + (1-\alpha)f(y) \geq f(\alpha x + (1-\alpha)y)$.

$\Longleftarrow$：取 $(x_1,y_1),(x_2,y_2) \in \mathrm{epi}(f)$，得到 $f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2) \leq \alpha y_1 + (1-\alpha)y_2$，所以 $(\alpha x_1 + (1-\alpha)x_2, \alpha y_1 + (1-\alpha)y_2) \in \mathrm{epi}(f)$. □

然后是下水平集。

定义 6.4 (下水平集) 给定 $t \in \mathbb{R}$，函数 f 的**下水平集**是指集合 $C_t(f) = \{x \in \mathbb{R}^n : f(x) \leq t\}$. 直观上说，下水平集是函数值小于 t 的区域.

命题 6.3 如果函数 f 是凸函数，那么对任意 $t \in \mathbb{R}$，下水平集 $C_t(f)$ 是凸集.

这个命题的证明是直接的，我们留做习题。值得注意的是，这一命题的逆命题是不成立的，我们也在习题中讨论。[lhy: 习题：证明命题 6.3]

接下来，我们研究凸集的性质。根据定义，直接有：

命题 6.4 凸集的任意交依然是凸集.

我们可以利用这个性质来构造新的凸集.

例 6.4
- 仿射空间：有限个超平面的交，等价地写作 $\{x \in \mathbb{R}^n : Ax = b\}$，$A \in \mathbb{R}^{m\times n}$，$b \in \mathbb{R}^m$.
- 多面体：有限个半空间的交，等价地写作 $\{x \in \mathbb{R}^n : Ax \le b\}$，$A \in \mathbb{R}^{m\times n}$，$b \in \mathbb{R}^m$.
- 单纯形：$\Delta_n = \{x \in \mathbb{R}^n : x_1 + \cdots + x_n = 1, x_i \ge 0, \forall i\}$，是一种特殊的多面体.
- 凸包：给定任意集合 S，可以定义包含它的最小凸集：

$$\bigcap_{S\subseteq C \text{ 是凸的}} C.$$

从优化的角度来看，凸集本身具有最优近似性质. 我们之前在例 6.1 讨论过，求点到线性空间的投影是一个优化问题. 任何一个点都可以唯一地投影到线性空间的某个点上，因此整个空间通过投影就被近似到了一个线性子空间中.

现在我们来推广这一考虑。给定任意非空集合 $C \subseteq \mathbb{R}^n$，我们尝试将整个空间近似到集合 C 中. 定义点 x 到 C 的距离为：$d(x, C) = \inf_{p\in C} \|x - p\|_2$. 如果存在 $p \in C$ 达到了距离 $d(x, C)$，我们就说 p 是 x 在 C 上的一个投影. 到当 C 就是线性空间的时候，这个定义恰好也是原来投影的定义.

如果 $\mathbb{R}^n$ 中的每个点都在 C 中有唯一的投影，那么就称 C 是 **Chebyshev 集**. C 是 Chebyshev 集意味着 C 是整个空间的一个好的近似. 我们有如下定理：

定理 6.8 *在 $\mathbb{R}^n$ 中，C 是 Chebyshev 集当且仅当 C 是闭凸集.*

这一定理的证明非常复杂，我们留做习题。[lhy: 习题：证明上述定理]

因此，闭凸集是唯一具有良好近似性质的集合类，这又一次从优化角度说明了凸性的重要性.

6.3.2 分离超平面定理

[lhy: 扩展这部分内容，把 Banach-Hahn 定理还有画图的事情处理好。]

凸集还有一个不平凡且重要的性质：

定理 6.9 (分离超平面定理) *设 C, D 是两个非空不交凸集，也就是 $C \cap D = \varnothing$. 那么，存在 $a \neq 0$ 和 $b \in \mathbb{R}$ 使得*

- *任意 $x \in C$，$a^\top x \le b$.*

- 任意 $x \in D$，$a^{\mathsf{T}} x \geq b$.

如果两个凸集只有一个公共点，并且其中一个凸集有内点，分离超平面定理依然成立，证明留做习题。[lhy: 习题：证明分离超平面定理]

下面我们来证明定理 6.9.

证明 定义两个集合间的距离为：

$$d(C,D) = \inf_{x \in C, y \in D} \|x - y\|_2 .$$

我们只证明 C 和 D 都是有界闭集的情况. 此时，存在 $c \in C, d \in D$ 使得 $\|c-d\|_2 = d(C,D)$. 令 $a = d - c$，$b = (\|d\|_2^2 - \|c\|_2^2)/2$. 只需证明 $f(x) = a^{\mathsf{T}} x - b$ 在 C 上非正在 D 上非负. 对称地，只证明在 D 上非负.

注意到 $f(x) = a^{\mathsf{T}} x - b = (d-c)^{\mathsf{T}}(x - (d+c)/2)$. 假设对某个 $u \in D$，$f(u) < 0$，于是

$$f(u) = (d-c)^{\mathsf{T}}(u-d) + \frac{1}{2}\|d-c\|_2^2 < 0 \implies (d-c)^{\mathsf{T}}(u-d) < 0.$$

因此，对充分小的 $t > 0$，$\|d + t(u-d) - c\|_2 < \|d + 0 \cdot (u-d) - c\|_2 = \|d-c\|_2$. 同时，因为 D 是凸集，$d + t(u-d) \in D$. 这与 d 和 c 的假设矛盾！ □

第四部分

逻辑与博弈

第五部分

认知逻辑

参考文献

[Bre57] Leo Breiman. The Individual Ergodic Theorem of Information Theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

[CT12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[Huf52] David A. Huffman. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*, 40(9):1098–1101, September 1952.

[Inf] Information | Etymology, origin and meaning of information by etymonline. https://www.etymonline.com/word/information.

[Jay02] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2002.

[KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[LLG+19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019.

[McM53] Brockway McMillan. The Basic Theorems of Information Theory. *The Annals of Mathematical Statistics*, 24(2):196–219, June 1953.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, USA, January 1986.

[Rob49] Robert M. Fano. *The Transmission of Information*. March 1949.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1996.

[Tin62] Hu Kuo Ting. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4):439–447, January 1962.

[Uff22] Jos Uffink. Boltzmann's Work in Statistical Physics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2022 edition, 2022.

[李 10] 李贤平. 概率论基础. 高等教育出版社, 2010.

索引