标题 title

作者 *author*

2023 年 7 月 26 日

前言

目录

第一部分

科学的逻辑

第一章　合情推理

第二章　Markov 链与决策

第二部分

信息与数据

第三章　信息论基础

信息是什么？不同于真实的物理世界，信息仿佛看不见，摸不着．然而，任何人都可以体会到信息的存在，信息是我们认识世界的基础．信息的存在正如同物理世界中的能量、动量一般，抽象而具有一般性．信息论已经在计算机、AI、认知理论等诸多领域中得到了广泛的应用．本章探讨信息论的基础，并给出他们在 AI 中的一些应用．

在第 3.1 节，我们讨论熵的概念与性质．在第 3.2 节，我们讨论 Kullback-Leibler 散度的概念与性质．在第 3.3 节，我们给出 Shannon 定理证明．

§3.1　熵

3.1.1　概念的导出

我们常说“恐惧来源于未知”，信息似乎代表着某种确定的东西，某种知识，因而和不确定性有相反的关系．更精确地说，*消除不确定性的东西被称为信息*．当然，这句话本身似乎是一种循环论证，它并没有真正回答信息或者不确定性到底是什么．所以我们进一步的问题是，给定一个“对象”，如何定量衡量它不确定性（或信息量）？

然而，单个对象的信息是一个非常难以划定的概念．同样的内容，对于不同的人来说，信息量是完全不同的．比如说，已经学过信息论的读者再看这一部分内容，他获得的信息一定比没有学过的读者要少得多．因而实际上，一种更加容易的办法是我们将世界视为不确定的，因而有多种可能的对象，然后考虑这一堆对象的信息量．比如说，这本书的读者的背景是不确定的，可能学过信息论，也可能没学过，但是我们可以综合考虑不同读者的背景，然后给出一个信息的概率分析．

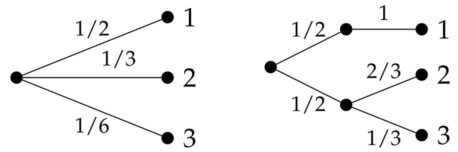
我们可以用数学来表述上面的考虑，假如我们进行一次试验，一共有 n 种可能的结果，第 i 种发生的概率为 p_i．我们预测试验的结果，如果越能正确地预测，那么就说明我们对这个试验中包含的信息知道的越多．假如 $p_1=1$，那么我们完全确定试验一定会产生结果 1．如果 $p_i=1/n$，那么我们完全无法预计试验的结果．我们对试验结果的预期与

试验结果的概率分布有密切联系. 因此概率分布给我们带来了**信息**，使得我们能够产生不同的判断. 另一方面，概率分布带来了**不确定性**，使我们不能总是确信预言会成真.

我们遵循“信息论之父”Shannon 的思路，为信息提供一个严格的数学模型：熵. 假设随机变量 X 表示了所有可能的结果（编号为 1 到 n），$\Pr(X=i)=p_i$，$p=(p_1,\ldots,p_n)$，有时候也把 p_i 写作 $p(i)$. 我们把不确定性度量记为 $H(p)$. Shannon 假设 H 满足以下三个性质：

1. H 是一个连续函数.
2. 事件结局可能数变多则不确定性增大：$p_i=1/n$ 时，$H(p)$ 随 n 单调递增，n 是正整数.
3. 如果一个试验被分成了两个相继的试验，那么原来的 H 应该等于分开之后的 H 的加权和.

注. 第三个假设可以用下图来理解.



假设我们有一个试验，有三种可能的结果，1,2,3，概率分别为 1/2,1/3,1/6. 该试验的不确定性是 $H(1/2,1/3,1/6)$. 我们把试验分成两步相继的试验，第一步试验有两种可能的结果，概率分别都是 1/2. 当第一步试验出现上面的结果时，第二步试验以概率 1 产生结果 1；当第二步试验出现下面的结果时，第二步试验以概率 2/3 产生结果 2，以概率 1/3 产生结果 3. 我们可以看到，分成两步之后，第一步试验的不确定性是 $H(1/2,1/2)$，第二步试验的不确定性有一半概率是 $H(1)$（上面的分支），有一半概率是 $H(2/3,1/3)$（下面的分支），因而加权的不确定性是 $1/2\cdot 0+1/2\cdot H(2/3,1/3)$. 因此第三个假设可以具体表述为

$$H\left(\frac{1}{2},\frac{1}{3},\frac{1}{6}\right)=H\left(\frac{1}{2},\frac{1}{2}\right)+\left[\frac{1}{2}\cdot H(1)+\frac{1}{2}\cdot H\left(\frac{2}{3},\frac{1}{3}\right)\right].$$

这里，我们可以看出 Shannon 的哲学思想：不确定性只来自于概率分布而不是具体对象. 他的考虑具有浓厚的工程意味，正如他自己针对通信的数学理论所说：“消息是具有含义的……然而，通信的语义层面并不是工程问题所关心的.“正是因为抽象掉了具体考虑的对象，信息论的应用才变得如此广泛.

基于上面三个假设，Shannon 证明了如下定理，这一定理直接给出了熵的概念.

定理 3.1 (Shannon 定理) H 满足三个假设当且仅当

$$H(p) = -C \sum_i p_i \log p_i,$$

其中 C 是正常数，$0 \log 0 = 0$.

这一定理的证明较长并且和后面的讨论关联较小，所以我们在第 3.3 节中给出证明.

根据对数的的换底公式，可以将 $C \log p_i$ 写为 $\log_b p_i$，这里 $C = 1/\log b$. 于是，Shannon 定理直接给出了熵的如下定义：

定义 3.1 (熵) 分布列 $p = (p_1, \ldots, p_n)$ 的**熵**定义为

$$H(p) = -\sum_{i=1}^{n} p_i \log_b p_i.$$

其中 $b = \mathrm{e}$（自然对数底数），$0 \log 0 = 0$. 当 $b = 2$ 时，我们记熵为 $H_2(p)$.

通常来说，使用 e 作为底数会使得数学推导简洁，而用 2 为底数则常常是讨论信息量时的习惯. 在后面通信理论中，我们将讨论熵在通信中的含义，以 2 为底的时候熵的实际意义会更清楚些. 如果没有特别强调，我们在讨论时总是假设 $b = \mathrm{e}$.

熵的定义还可以用数学期望的形式写出. 假设 X 的分布列是 p，$p(i) = \Pr(X = i)$，那么我们也可以把熵写成期望的形式：

$$H(p) = -\mathbb{E}[\log p(X)].$$

每一个（离散）随机变量 X 会确定一个分布列 p_X，因此我们也可以定义随机变量的熵：

定义 3.2 (随机变量的熵) 随机变量 X 的**熵**定义为

$$H(X) = -\mathbb{E}[\log p_X(X)].$$

其中 p_X 是 X 的分布列，$0 \log 0 = 0$.

尽管从信息论的角度我们可以唯一确定熵的定义，但是熵的概念在物理学上早就已经存在. 下面我们给出统计力学中熵的推导过程. 在经典力学中，物理系统的状态由粒子的位置和动量（速度）完全确定，将粒子位置和动量可能的值集合称为*相空间*，于是物理系统的演化就是相空间中的粒子状态的变化. 将相空间等分成 m 个单元，编号 1 到 m. 假设相空间中有 N 个可区分的粒子，相互独立，没有相互作用，每个粒子等可能出现在每一个单元中. 如果单元 i 中有 N_i 个粒子，那么按照粒子在单元中的分布来看，系统处于某个特定状态的概率为

$$P = \frac{N!}{N_1! \ldots N_m!} \left(\frac{1}{m}\right)^N.$$

这是一个多项分布. 两边取对数，得

$$\log P = \log(N!) - \sum_i \log(N_i!) - N \log m.$$

考虑充分大的 N_i，由 Stirling 公式，有

$$\log(N_i!) \sim \log\left(\sqrt{2\pi N_i}\left(\frac{N_i}{\mathrm{e}}\right)^{N_i}\right) \sim N_i \log N_i.$$

因此，

$$\log P \sim N \log N - \sum_i N_i \log N_i - N \log m \sim N \log N - \sum_i N_i \log N_i. \tag{3.1}$$

假设 N_i 充分大的时候，N_i/N 呈现固定的比例 p_i，那么

$$\begin{aligned} N \log N - \sum_i N_i \log N_i &\sim N \log N - \sum_i N p_i \log(N p_i) \\ &= -N \sum_i p_i \log p_i. \end{aligned}$$

$\log P \sim -N \sum_i p_i \log p_i$. 于是我们证明了：

$$\frac{1}{N} \log P \to H(p_1, \ldots, p_m), \quad N \to \infty.$$

因此，熵刻画了充分多粒子的物理系统某种特定状态出现概率！熵越大的系统越有可能达到. 更进一步，在统计力学中有 Boltzmann H-定理：孤立的粒子系统会向着熵（H）增加的方向演化，并最终达到熵最大的状态. H-定理是热力学第二定律的微观解释，熵越大的系统出现概率越大、越混乱、越接近均衡.

3.1.2 概念与性质

现在，我们将进一步探讨熵的若干拓展定义，并讨论他们的性质.

首先，我们考虑最简单的情形，即分布列为 (p_1, p_2)，此时，我们不妨设 $p_1 = p$，$p_2 = 1 - p$，那么熵就是

$$H(p_1, p_2) = H(p, 1-p) = -p \log p - (1-p) \log(1-p).$$

H 是关于 p 的函数，作图如图 3.1 所示.

利用导数的方法，很容易证明：

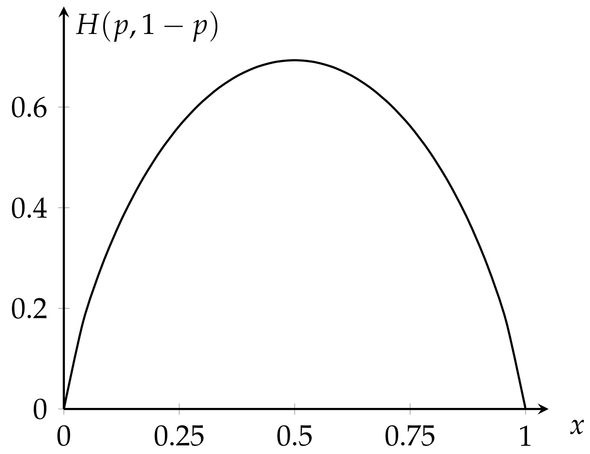


图 3.1: 熵 $H(p)$ 的图像.

命题 3.1 $H(p)$ 在 $p \in (0, 1/2)$ 严格单调递增，在 $p \in (1/2, 1)$ 严格单调递减. 它的最小值是 0，在 $p \in \{0, 1\}$ 取得；它的最大值是 $\log 2$，在 $p = 1/2$ 取得.

这与我们对于"不确定性"的直觉是相一致的：当 p 接近 0 或 1 时，我们对于 X 的取值几乎是确定的，因此熵接近 0；当 p 接近 $1/2$ 时，我们对于 X 的取值几乎是完全不确定的，因此熵接近最大值 $\log 2$. 实际上，这样的性质对于一般的分布也是成立的.

考虑一般分布的熵 $H(p) = H(p_1, \ldots, p_n)$. 我们有如下性质：

命题 3.2 $H(p) \geq 0$，等号成立当且仅当某个 $p_i = 1$.

证明 这是一个典型的证明，主要的技巧是使用熵的期望形式. 考虑随机变量 X，其分布列为 p. 回忆 Jensen 不等式：如果 f 是一个严格凸函数，那么

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

等号成立当且仅当 X 是常数.

因为 $-\log(\cdot)$ 是严格凸函数，所以根据 Jensen 不等式

$$H(X) = \mathbb{E}[-\log p(X)] \geq -\log \mathbb{E}[p(X)] \geq -\log 1 = 0.$$

等号成立当且仅当 X 是常数，即对某个 i，$p(i) = 1$. □

命题 3.3 p_i 朝着相等方向改变的时候 H 增加. 也就是说，假设 $p_i < p_j$，再假设 $p_i' > p_i, p_j' < p_j$ 且 $p_i + p_j = p_i' + p + j'$. 用 p_i' 和 p_j' 代替原来的 p_i 和 p_j，那么 H 变大.

证明 为简化符号，考虑 $i=1$ 和 $j=2$. 利用假设三，第一步试验中，将试验的结果 1 和结果 2 合并，第二步试验再按照 $p_1/(p_1+p_2)$ 和 $p_2/(p_1+p_2)$ 的概率产生结果 1 和结果 2. 于是，

$$
\begin{aligned}
&H(p_1,p_2,\dots)\\
=&H(p_1+p_2,p_3,\dots)+(p_1+p_2)H\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right) \quad \text{（假设三）}\\
\le&H(p_1+p_2,p_3,\dots)+(p_1+p_2)<H\left(\frac{p_1'}{p_1'+p_2'},\frac{p_2'}{p_1'+p_2'}\right) \quad \text{（命题 3.1）}\\
=&H(p_1',p_2',p_3,\dots). \quad \text{（假设三）}
\end{aligned}
$$

□

命题 3.4 当 $p_1=\cdots=p_n=1/n$ 时 H 取得最大值 $\log n$.

证明 若存在 $p_i\neq p_j$，因为 $\sum_i p_i/n=1/n$，根据鸽巢原理，则必有 i,j 满足 $p_i<1/n<p_j$. 根据命题 3.3，我们可以将 p_i 和 p_j 替换为 $1/n$ 和 p_i+p_j-1/n，而 H 增大. 只要还有两个 p_i 不相等，这一过程就可以重复，每一次都会增大 H，直到所有 p_i 都等于 $1/n$. □

至此，命题 3.2 和命题 3.4 证明了一般情形的命题 3.1. 在等可能的时候不确定性最大，熵最大；在确定事件的时候不确定性最小，熵最小. 所以熵是符合直观的定义.

接下来，我们讨论熵的拓展形式.

在一次试验中，我们可以观察多个变量，比如说 X 和 Y. 我们也可以说，我们观察到了一个结果 (X,Y)，服从分布 $p(i,j)$. 因此有对应的熵，这就是**联合分布的熵**：

$$H(X,Y)=-\mathbb{E}[\log p(X,Y)].$$

对应地，我们也可以写成和的形式：

$$H(p)=-\sum_{i,j}p(i,j)\log p(i,j).$$

自然，联合分布也可以引出**边缘分布的熵**：

$$H(X)=-\mathbb{E}[\log p_X(X)]=-\sum_i\sum_j p(i,j)\log\sum_j p(i,j).$$

$$H(Y)=-\mathbb{E}[\log p_Y(Y)]=-\sum_j\sum_i p(i,j)\log\sum_i p(i,j).$$

有了两个随机变量，我们就可以讨论“条件”的概念. 具体来说，我们可以把试验分为两步，第一步观测 X，第二步观测 Y，那么，第二步所产生的熵就是已经知道第一步结果之后的熵，即:

$$H(Y|X=x) = -\mathbb{E}[\log p_{Y|X=x}(Y)|X=x] = -\sum_j p_{Y|X=x}(j)\log p_{Y|X=x}(j),$$

其中 $p_{Y|X=x}(j) = p(x,j)/p_X(x)$. 当我们知道了 $X=x$ 之后，对 Y 的观测就消除了部分的不确定性，因此根据我们对于不确定性和信息关系的讨论，从 $X=x$ 中获得的关于 Y 的信息是

$$I(X=x:Y) = H(Y) - H(Y|X=x).$$

考虑一个特殊情况，$Y=X$，那么刚刚的讨论就变成了自己从自己身上获得的信息，或者说知道 $X=x$ 带来的信息量. 首先有

$$p_{X|X=x}(i) = \begin{cases} 1, & i=x \\ 0, & i\neq x. \end{cases}$$

因此，

$$H(X|X=x) = -\sum_j p_{X|X=x}(j)\log p_{X|X=x}(j) = -1\log 1 = 0.$$

于是，

$$I(X=x:X) = H(X) - H(X|X=x) = H(X).$$

这正是定量版本的“消除不确定性的东西被称之为信息”! 此外，我们之前说过，熵刻画的是一族可能对象的信息，这一点也反映在了这一公式中：只要知道了 X 的值，无论它具体是多少，我们得到的信息量是一样的！

再回到一般情况，还是同样的两步试验，我们定义给定 X 时 Y 的**条件熵**为

$$\begin{aligned} H(Y|X) =& \mathbb{E}[H(Y|X=x)] \\ =& -\mathbb{E}[\log p_{Y|X}(Y)] \\ =& -\sum_x p_X(x)\sum_j p_{Y|X=x}(j)\log p_{Y|X=x}(j) \\ =& -\sum_{x,j} p(x,j)\log p_{Y|X=x}(j). \end{aligned}$$

换言之，我们现在进一步假定 X 也是不知道的，于是 $H(Y|X)$ 就是平均上来说第二步中 Y 的不确定性. 条件熵和熵有着类似的性质:

命题 3.5 $H(Y|X)\geq 0$，等号成立当且仅当 Y 是退化的，即 Y 概率 1 只取一个值.

证明 仿照命题 3.2 的证明即可. $\square$

类似地，我们可以考虑平均上 Y 中包含的关于 X 的信息量：

$$\mathbb{E}[I(X=x:Y)] = H(Y) - H(Y|X).$$

与之相对应地，平均上 X 中包含的关于 Y 的信息量为

$$\mathbb{E}[I(Y=y:X)] = H(X) - H(X|Y).$$

一个自然的问题是，二者相互包含的信息量是什么关系？根据概率的链式法则，$p(x,y) = p_{X|Y}(x|y)p_Y(y)$，带入 $H(X,Y)$ 的定义得熵的链式法则：

命题 3.6 对任意离散随机变量 X, Y，$H(X,Y) = H(Y) + H(X|Y)$.

利用链式法则，我们注意到，$H(X) - H(X|Y) = H(X) - (H(X,Y) - H(Y)) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y|X)$. 所以，$X$ 中包含的 Y 的信息和 Y 中包含的 X 的信息是一样多的！此外，直观上我们还应该觉得，信息量不能是负的，实际上的确如此：

命题 3.7 $H(X) - H(X|Y) \geq 0$，等号成立当且仅当 X 和 Y 相互独立.

我们将在第 3.2 节看到，命题 3.7 就是 K-L 散度信息不等式的一个特例，所以我们就不在这里给出证明了. 命题 3.7 表明知道任何信息都不会增加不确定性，这个原理被称为“Information doesn't hurt.”根据以上讨论，我们可以自然地定义 X 和 Y 的**互信息**为 $I(X;Y) = I(Y;X) = \mathbb{E}[I(X=x:Y)] = \mathbb{E}[I(Y=y:X)]$.

类似联合分布的熵，条件熵和互信息的概念也可以推广到多元情形. 对于三个随机变量 X, Y, Z，我们可以定义条件熵为

$$H(X,Y|Z) = H(X,Y,Z) - H(Z).$$

类似地，我们可以定义互信息为

$$I(X,Y;Z) = H(X,Y) - H(X,Y;Z).$$

他们的含义以及性质和二元情形类似.

同样，我们可以定义**条件互信息**为 $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$，表明 Z 已知时候 Y 给 X 带来的平均信息增益. 类似互信息，我们如下性质：

命题 3.8 条件互信息满足以下性质：

1. 非负性：$I(X;Y|Z) \geq 0$，等号成立当且仅当 X 和 Y 在给定 Z 的条件下相互独立.
2. 对称性：$I(X;Y|Z) = I(Y;X|Z)$.
3. 链式法则：$I(X,Y;Z) = I(X;Z|Y) + I(Y;Z)$.
4. 条件信息量：$I(X:X|Y) = H(X|Y) - H(X|X,Y) = H(X|Y)$.

最后一条性质说的其实是，在平均的意义下，给定 Y 的时候，知道 X 所能够得到的额外信息量就是 $H(X|Y)$. 这一命题的证明和前面都非常相似，我们留做习题.

最后，我们将各种熵以及信息量的关系总结为图 3.2. 在集合论中，这样的图被称为 Venn 图，所以我们可以用集合论来理解信息与熵. 对应关系可以总结为表 3.1.

信息论	集合论
$H(X)$	A
$H(Y)$	B
$H(X\|Y)$	$A \setminus B$
$H(X,Y)$	$A \cup B$
$I(X;Y)$	$A \cap B$

表 3.1: 信息论和集合论的对应关系.

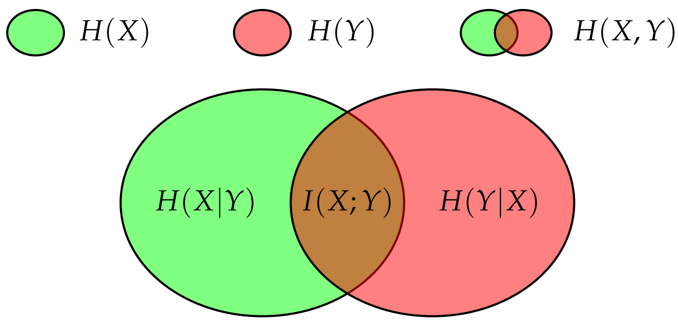


图 3.2: 熵和信息量的关系.

3.1.3 熵与通信理论

最早的时候，Shannon 建立信息论，就是为了给通信理论一个数学基础. 从通信的角度出发，我们可以更本质地理解信息和熵.

通信就是一个发射端和一个接收端，中间有信道传递消息. 将所有可能要传递的消息集合记为 Ω（一个有限集），我们现在考虑 Ω 所蕴含的信息量是多少. 注意到，根据 Shannon 的思想，Ω 里面具体是什么并不重要，重要的是有多少个. 我们可以用自然数 $1, 2, \ldots$ 表示集合 Ω 里的元素. 那么，使用二进制编码，我们至少需要 $\log_2 |\Omega|$ 个比特来表示 Ω 里的元素. 于是，假如说随机变量 X 表示收到的消息，那么 X 的熵就定义为 $H(X) = \log_2 |\Omega|$，它衡量了接收端收到的消息的不确定性. 当我们选定了具体的消息 $m \in \Omega$，X 的不确定性被消除了，于是 $X = a$ 的过程产生了（或者说传递了）$\log_2 |\Omega|$ 比特的信息. 比如说，我们发送一个长为 n 的二进制序列，消息的集合大小就是 2^n，发送任何一条具体的消息，我们就传递了 n 比特的信息.

有时候，我们会把消息看成一个序列. 具体来说，我们可以发送独立的 k 条消息，其中第 i 条 X_i 来自消息集合 Ω_i，$|\Omega_i| = n_i$，那么 $(X_1, \ldots, X_k)$ 的熵就是

$$H(X_1, \ldots, X_k) = \log_2 n_1 + \cdots + \log_2 n_k,$$

它衡量了 k 条消息的不确定性. 在更常见的情况下，每次发送的其实不是一条消息，而是一个字母，所有的字母组成了一个字母表，我们用 $\Sigma = \{x_1, \ldots, x_s\}$ 来表示. 于是，X_i 就是消息的第 i 个字母，于是，一条消息可以写作 $X_1 \ldots X_k$，其中每一个 X_i 都来自 Σ.

我们现在考虑更加简单的情形，即每一个字母 X_i 其实是同一个随机变量 X 的独立采样. 如果我们具体知道某一个 x_i 出现的次数，那么我们其实可以有更高效的传递信息的方式. 譬如说，在极端情况下，如果只有 x_1 和 x_2 会出现，那么我们其实只需要 $\log_2 2 = 1$ 比特就足够传递所有消息了. 在一般情况下，考虑 Ω 中只包含长为 k 的消息，并且 x_i 在消息中出现 k_i 次，那么所有可能的消息数量为

$$N(k) = \frac{k!}{k_1! \cdot \cdots \cdot k_s!}.$$

假定我们需要 $h(\omega)$ 比特来具体确定发的消息是 ω. 首先，无序集合本身需要 $\log_2 |\Omega|$ 比特来编码，其次，我们还需要确定 $(k_1, \ldots, k_s)$，确定它的一种方式是按照顺序给出每一个 k_i. 每个 k_i 最多需要 $\log_2 k$ 比特来表示，所以按顺序表示所有的 k_i 至多需要 $s \log_2 k$ 比特. 于是，我们需要的比特数为

$$\log_2 \frac{k!}{k_1! \cdot \cdots \cdot k_s!} \le h(m) \le s \log_2 k + \log \frac{k!}{k_1! \cdot \cdots \cdot k_s!}.$$

这刚好和我们在统计力学中推导熵的过程是一致的！假设消息足够的长，x_i 出现的频率逐渐接近 p_i，那么同样的推理我们可以知道，

$$h(m) \sim -k \sum_i p_i \log_2 p_i = k H_2(p_1, \ldots, p_s).$$

因此，如果知道字母的出现频率，我们传递单位长度的消息至少需要 $H(p_1, \ldots, p_s)$ 比特，这完全给出了熵的具体含义，而且，我们现在也不难理解熵的形式为何会出现 log 了：熵就是期望上编码一个字母需要的比特数（即 $\log(1/p(X))$）.

那么，是否有一种编码确实达到了这个理论上的编码长度下界呢？答案是肯定的，它被称为 *Huffmann* 编码. 它的核心思想在于把出现频率高的字母用更短的编码表示. 类似的思想被用在了机器学习的决策树中，作为选择节点非常常用的一种依据.

> **注.** 决策树是一种常用的机器学习分类模型. 假设数据有很多属性 $P_1, \ldots, P_k$，这些属性共同决定了某一条数据的类别. 比如，在银行的信用系统中，给定了一个人的性别、是否已婚、是否负债等信息，我们希望给他评估一个信用评级. 决策树的做法是，将决策过程写成一棵树，然后叶节点是决策类别的结果. 比如说，我们会先看这个人是否负债，如果不负债，那么看是否已婚，如果已婚，那么我们信用评级就给 A. 那么，如何选择每个节点需要去判断的属性呢？树本身其实就是一种广义的消息，从根节点沿着树走到叶节点得到的就是一条消息. 直观上，如果先选择带来信息增益比较高的属性，那么我们就可以用更少的比特来表示这条消息，或者说，我们决策树结构更加简单. 这样的选择方式叫做 ID3 策略.

我们进一步的问题是，为什么我们知道了每个字母的频次就可以压缩编码？我们接下来将要说明，其实长为 k 的消息中的“典型的消息”数量是远远少于所有 k 长消息的数目，因此我们实际上相当于只是针对一个子集进行编码. 注意到，当 k 充分大的时候，

$$\log_2 N(k) \sim h(m) \sim kH_2(p_1, \ldots, p_s).$$

因此，

$$N(k) \approx 2^{kH_2(p_1, \ldots, p_s)} = \mathrm{e}^{kH(p_1, \ldots, p_s)}.$$

然而，长为 k 的所有消息数目为

$$s^k = \mathrm{e}^{k \log s}.$$

根据命题 3.4，只有当所有 p_i 相等的时候 $N(k)$ 才会达到这一量级. 从这个意义上说，熵所刻画的信息量定量刻画了数据压缩可能的极限.

以上关于信息编码下界以及数据压缩的讨论，再更一般的情况下也成立，此时这样的性质被称为渐近等分性. 而这一性质成立对应的结果被称为 *Shannon–McMillan–Breiman* 定理，它的陈述以及证明都需要用到更多随机过程的知识，这里就不再给出了.

> **注.** 现代的主流信息论都是从 *Shannon* 发展起来的. 然而，这一信息论也有很多问题. 首先，信息论使用了概率论进行建模. 但我们已经看到，概率要么是作为频率的近似理论（频率学派），要么反映了人们对未知的信念（主观学派）. 无论哪种解释，都将问题简化了. 正如 *Kolmogorov* 所说：“如果事情没有按照我们的预期发展，那么问题一定出在我们对于概率和真实世界的随机之间关系不清晰的认识上.”其次，这一信息论考虑的是一族对象的信息. 我

们是否能够用这样的方式来衡量单个对象的信息量呢？比如，我们要考虑这本书中包含的信息量，是它放在所有可能的书的集合中去考虑呢，还是把它的每一个章节分开考虑成一个随机序列呢？因此，信息论并不能很好地回答“单个对象”的信息量的问题.

现代概率论的奠基人 *Kolmogorov* 也非常严肃地考虑了这一问题. 他提出了被后世称为 ***Kolmogorov* 复杂度**的概念，旨在刻画一个随机字符串的随机程度. 简单来说，一个字符串的 *Kolmogorov* 复杂度就是描述它所需要的最短的代码长度，越随机的字符串就越需要更复杂的程序去描述它的产生方式. 利用这一概念，我们可以将信息的概念变成一个对象自己的属性，而不再需要把对象放在可能的一堆对象中去考虑. 这是信息论的另一种构建思路.

§3.2 Kullback-Leibler 散度

3.2.1 定义

为了引入 K-L 散度，我们从互信息出发. 它的定义是：

$$\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= -\sum_x p_X(x)\log p_X(x) + \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p_Y(y)} \\
&= -\sum_{x,y} p(x,y)\log p_X(x) + \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p_Y(y)} \\
&= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p_X(x)p_Y(y)}.
\end{aligned}$$

根据命题3.7，$I(X;Y)\geq 0$，等号成立当且仅当 X,Y 相互独立，即 $p(x,y)=p_X(x)p_Y(y)$. X,Y 之间的互信息越大，说明他们之间的关联越强，分布越不独立，$p(x,y)$ 越不接近 $p_X(x)p_Y(y)$. 实际上，这样的想法可以被推广到一般分布上.

我们从数理统计的视角出发，考虑两个概率分布的似然函数p_1 和 p_2（也就是他们的分布列）. 抽取一个样本 X，考虑假设检验问题：

$$H_1\text{：样本 } X \text{ 来自 } p_1 \text{ 的分布} \quad \text{vs.} \quad H_2\text{：样本 } X \text{ 来自 } p_2 \text{ 的分布}$$

假设检验中有一种很常用的技巧，称为*似然比检验法*，即考虑两个假设分布的似然比 p_1/p_2. 如果这个比值越大，就越说明 p_1 的值更大，因而更有可能，倾向于接受 H_1，反之则越倾向于接受 H_2. 于是，可以自然定义区分 H_1 和 H_2 的检验量为对数似然比：

$$\log(p_1(x)/p_2(x)).$$

假设 H_1 是真的，那么在 H_1 的世界里，这个检验量的期望为

$$\mathbb{E}_{X\sim p_1}(\log(p_1(X)/p_2(X))) = \sum_i p_1(i)\log\frac{p_1(i)}{p_2(i)}.$$

实际上，上面的期望就是 K-L 散度的定义.

定义 3.3 (Kullback-Leibler 散度) 对于两个概率分布列 p_1, p_2，他们的 ***Kullback-Leibler* 散度**或**相对熵**被定义为

$$D(p_1\|p_2) = \mathbb{E}_{X\sim p_1}(\log(p_1(X)/p_2(X))) = \sum_i p_1(i)\log\frac{p_1(i)}{p_2(i)}.$$

其中规定 $0\log(0/0)=0$，$0\log(0/a)=0$，$a\log(a/0)=+\infty$.

我们马上知道，互信息是 K-L 散度的一种特殊情况：

命题 3.9 对于两个随机变量 X,Y，成立 $I(X;Y)=D(p_{X,Y}\|p_Xp_Y)$，其中 $p_{X,Y}$ 是 X,Y 的联合分布列，p_X, p_Y 分别是 X,Y 的边缘分布列.

K-L 散度可以看成两个分布之间的区分衡量标准，但他不是度量. 一般来说，甚至连对称性都不成立. 例如，设 p_1 和 p_2 都是定义在 $0,1$ 上的 Bernoulli 分布，参数分别为 $1/2$ 和 $1/4$. 于是

$$D(p_1\|p_2) = \frac{1}{2}\log\frac{1/2}{3/4} + \frac{1}{2}\log\frac{1/2}{1/4} = \frac{1}{2}\log\frac{4}{3}.$$

$$D(p_2\|p_1) = \frac{3}{4}\log\frac{3/4}{1/2} + \frac{1}{4}\log\frac{1/4}{1/2} = \frac{1}{2}\log\frac{3\sqrt{3}}{4}.$$

这两个值是不相等的.

我们在定义中还提到了 K-L 散度的另一个名字——相对熵. 实际上，这可以从编码中看出来. 假设事实上消息中字母的分布是 p_1，那么期望上编码单位长度消息需要的比特数是 $H(p_1)=\mathbb{E}_{X\sim p_1}[\log p_1(X)]$. 如果我们错误地认为消息中字母的分布是 p_2 并使用最优编码，那么实际上期望编码单位长度消息需要的比特数是 $\mathbb{E}_{X\sim p_1}[\log p_2(X)]$. 由于错误的认识所产生的额外编码长度是

$$\mathbb{E}_{X\sim p_1}[\log p_1(X) - \log p_2(X)] = D(p_1\|p_2).$$

根据在第 3.1.3 节中的讨论，我们知道，额外的编码长度代表的是额外的不确定性，因而这一概念是某种“熵”的概念. 这正是“相对熵”的由来，$D(p_1\|p_2)$ 表示了当我们错误地把 p_1 当成 p_2 时带来的额外的不确定性，或者说额外的信息损失.

在机器学习中，比起讨论 K-L 散度，更加常用的是直接讨论量 $\mathbb{E}_{X\sim p_1}[\log p_2(X)]$. 从机器学习的观点来说，$p_1$ 是真实的分布，而 p_2 是我们所学习到的分布. 根据刚刚的讨论，这个量越小越说明 p_2 接近真实的 p_1，因此这又是一种衡量两个分布之间关系的量，我们称之为**交叉熵**：

定义 3.4 (交叉熵) 给两个随机变量 X, Y，X 的分布为 p_X，Y 的分布为 p_Y，则 X 的分布 p_X 和 Y 的分布 p_Y 的**交叉熵**[1] 为

$$CH(p_X, p_Y) = \mathbb{E}_{X\sim p_X}[\log p_Y(X)] = -\sum_i p_X(i) \log p_Y(i).$$

在机器学习的分类问题中，我们希望学习到的分布 p_Y 尽可能地接近真实的分布 p_X，所以我们训练的目标经常是最小化交叉熵 $CH(p_X, p_Y)$. 有趣的是，从数理统计的角度来看，最小化交叉熵等价于进行*最大似然估计*，因此这为最大似然估计提供了一种信息论意义下的理解. 相关讨论留作练习.

3.2.2 两个关于信息的不等式

利用 K-L 散度，我们可以给出两个关于信息的不等式，它们分别是*信息不等式*和*数据处理不等式*.

定理 3.2 (信息不等式) *对于两个概率分布列 p, q，成立 $D(p\|q) \geq 0$，当且仅当 $p = q$ 时取等号.*

证明 由于 $\log x$ 是凸函数，所以由 Jensen 不等式，我们有

$$D(p\|q) = -\mathbb{E}_{X\sim p}\left[\log \frac{q(X)}{p(X)}\right] \geq -\log \mathbb{E}_{X\sim p}\left[\frac{q(X)}{p(X)}\right] = -\log \sum_i p(i) \cdot \frac{q(i)}{p(i)} = 0.$$

因此，$D(p\|q) \geq 0$，当且仅当 $p = q$ 时取等号. □

信息不等式表明，K-L 散度虽然不是度量，但却是非负的，因而确实可以被作为熵，用来衡量“额外的不确定性”. 此外，命题 3.7 是信息不等式的直接推论. 利用类似的方法，我们可以证明条件互信息的非负性（即命题 3.8 中的第一条）.

接下来我们叙述并证明数据处理不等式.

定理 3.3 (数据处理不等式) *假设随机变量 X, Y, Z 形成了 Markov 链，那么 $I(X;Y) \geq I(X;Z)$. 特别地，对任意函数 f，成立 $I(X;Y) \geq I(X;f(Y))$.*

证明 根据互信息链式法则，

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y). \end{aligned}$$

---

[1] 文献中，经常会直接写为 $H(p_X, p_Y)$，但是在本书中为了区分熵，我们使用了符号 CH.

根据 Markov 性，条件在 Y 上，X 和 Z 相互独立. 因此，$I(X;Z|Y)=0$，根据条件互信息的非负性，$I(X;Y|Z)\geq 0$，所以 $I(X;Y)\geq I(X;Z)$.

显然，$X,Y,f(Y)$ 也形成了 Markov 链，所以 $I(X;Y)\geq I(X;f(Y))$. □

数据处理不等式表明，无论我们对随机变量 Y 进行了何种处理，甚至是允许带随机的处理，它的信息量都不会增加.

3.2.3 在机器学习中的应用：语言生成模型

现如今，机器学习中最为瞩目的成果之一就是*大语言模型*，它通过学习人类海量的高质量语料库来形成一个生成式的模型，其中最为典型的例子是 ChatGPT. 从思路上来说，大语言模型的核心思想非常简单：给一段话，将其中一些词掩盖掉，让模型填出这些词来. 例如，给出“我在 [mask] 面条，真好吃“，模型应该能够填出“我在吃面条，真好吃“. 这样的思想，对于更一般的数据也是成立的：用（修改改过的）数据本身作为输入，训练一个*编码器*，然后将编码器的输出送入*解码器*，而解码器的输出具有原始数据的格式，我们希望这一输出能够尽量匹配原始的输入. 在自然语言处理中，一个*生成模型*往往同时有编码器和解码器. 比如说，图 3.3 展示的就是 BART [LLG+19] 的结构.

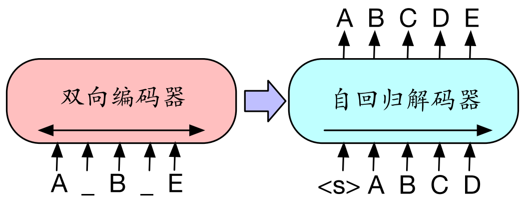


图 3.3: 生成式语言模型 BART 的示意图.

我们已经指出，熵和编码有着密切的联系. 从这个角度出发，我们很容易理解生成模型背后的思想：我们希望通过训练的方式得到一个由神经网络所表示的编码和解码规则，他要尽可能符合真实数据的分布.

我们可以用一种非常简单的模型去理解这一过程. 假设所有的单词的集合为 Σ，单词数为 k 的文本集合为 Ω. 我们希望训练一个生成模型 M，给它输入 $k-1$ 个单词，它可以给出第 k 个单词的概率分布，我们选择出现概率最大的那个词作为预测. 在训练的时候，对于一个句子 ω，我们只保留前 $k-1$ 个词，得到 $\omega[1:k]$，然后将它输入到生成模型 M 中，让它去预测第 k 个词.

对于这一个具体的句子来说，理想的分布应该是一个 *Dirac 分布*[2] $\delta_{\omega[k]}$，即以概率 1

[2] Dirac 分布是一个数学物理中更加常用的名字. 在概率论中，这也被称为*退化分布*；而在机器学习中，分

取到 $\omega[k]$. 假如说生成模型的输出是一个概率分布 $M(\omega[1:k-1]) = p$，那么，我们可以用 K-L 散度去衡量这两个分布的差异，因为 $H(\delta_{\omega[k]})$ 是固定的，所以我们只考虑交叉熵 $CH(\delta_{\omega[k]}, p)$. 因为一次训练会给多个样本，所以我们的目标是同时最小化这些交叉熵的和. 假如训练集是 T，我们的目标就是

$$\min_M \sum_{\omega\in T} CH(\delta_{\omega[k]}, M(\omega[1:k-1])).$$

实际上，这个例子是有普适性的，所有的监督训练的分类问题都可以用这种方式来建模. 而在 **??** 我们也会看到，此时交叉熵实际上被作为了一种*损失函数*.

§3.3 附录：Shannon 定理的证明

我们在这一部分给出 Shannon 定理（定理 3.1）的证明. 整体上的思路是：

1. 证明如果 f 是单调函数，对正整数 m,n 成立 $f(mn) = f(m) + f(n)$，那么 $f(n) = C\log n$.
2. 求出 $H(1/n,\dots,1/n)$ 的表达式.
3. 假设 p_i 是有理数，设 $p_i = n_i/\sum_j n_j$，考虑 $\sum_j n_j$ 个等可能试验结果，利用假设 3 推出 H 的表达式.
4. 利用有理数的稠密性和 H 的连续性推出一般情形.

最后一步是显然的，我们只需要证明前三步即可.

对第一步，我们需要证明的是，如果 f 是单调函数，对正整数 m,n 成立 $f(mn) = f(m) + f(n)$，那么 $f(n) = C\log n$. 首先，利用数学归纳法容易看出，对正整数 n,k，成立

$$f(n^k) = kf(n). \tag{3.2}$$

设 m,n 是任意两个大于 1 的整数，再选任意大的正整数 k，从 m 进制数的性质可以看出，总存在正整数 l 使得

$$m^l \le n^k < m^{l+1}. \tag{3.3}$$

根据 f 的单调性，我们有

$$f(m^l) \le f(n^k) < f(m^{l+1}).$$

利用式 (3.2)，我们有

$$lf(m) \le kf(n) < (l+1)f(m) \iff \frac{l}{k} \le \frac{f(n)}{f(m)} < \frac{l+1}{k}.$$

将式 (3.3) 取对数，得到

$$l\log m \le k\log n < (l+1)\log m \iff \frac{l}{k} \le \frac{\log n}{\log m} < \frac{l+1}{k}.$$

所以

$$\left|\frac{\log n}{\log m} - \frac{f(n)}{f(m)}\right| \le \frac{1}{k}.$$

布经常会表示为一个概率向量，文献中称为*独热向量*.

因为 k 可以是任意大的正整数，取 $k \to \infty$，我们就得到了

$$\frac{\log n}{\log m} = \frac{f(n)}{f(m)}.$$

由 m, n 的任意性，取 $m = 2$，我们就得到了 $f(n) = (f(2)/\log 2) \cdot \log n = C \log n$. 容易检验，$f(1) = 0 = C \log 1$，因此这一等式对所有正整数 n 都成立.

对第二步，我们需要求出 $f(n) = H(1/n, \ldots, 1/n)$ 的表达式. 我们要利用第一步的结果，首先，根据假设二，$f(n)$ 是单调递增的函数. 然后，考虑 mn 个等可能试验，我们可以将它分成两步试验，第一步有 m 中等可能的结果，而在每一种结果之下，第二步有 n 种等可能结果. 根据假设三，

$$f(mn) = f(m) + \frac{1}{n} \cdot nf(n) = f(m) + f(n).$$

所以 $f(n)$ 符合第一步的假设. 第二步就可以直接从第一步推出.

最后，我们证明第三步. 设 $p_1, \ldots, p_n$ 都是有理数，那么，他们可以被写为

$$p_i = \frac{n_i}{\sum_{j=1}^n n_j}.$$

其中 n_i 是非负整数. 我们考虑 $\sum_j n_j$ 个等可能试验，这个试验可以被看成两步的试验，第一步有 n 种可能的结果，第 i 种结果出现的概率是 p_i，而在第 i 种结果之下，第二步有 n_i 种等可能的结果. 根据假设三，和证明的第三步，我们有

$$C \log \sum_{j=1}^n n_j = H(p_1 + \cdots + p_n) + \sum_{i=1}^n p_i \cdot C \log n_i.$$

因此，

$$\begin{aligned}
H(p_1, \ldots, p_n) &= C\left(\log \sum_{j=1}^n n_j - \sum_{i=1}^n p_i \log n_i\right) \\
&= C\left(\log \sum_{j=1}^n n_j - \sum_{i=1}^n p_i \log\left(p_i \sum_{j=1}^n n_j\right)\right) \\
&= C\left(\log \sum_{j=1}^n n_j - \sum_{i=1}^n p_i \log p_i - \sum_{i=1}^n p_i \log \sum_{j=1}^n n_j\right) \\
&= -C \sum_{i=1}^n p_i \log p_i.
\end{aligned}$$

这正是我们要证明的. 于是，我们证明了 Shannon 定理.

§3.4 习题

1. 我们在熵以及 K-L 散度的定义中，都规定了一些无定义的量的值，这些值并不是随便规定的，他们实际上反映了熵或者 K-L 散度定义中的连续性.

(1) 证明：对给定的 $a > 0$，$\lim_{x \to 0+} x \log(x/a) = 0$，因此我们规定了 $0 \log 0 = 0$ 以及 $0 \log(0/a) = 0$.

(2) 证明：对给定的 $a>0$，$\lim_{x\to 0+} x\log(a/x)=+\infty$，因此我们规定了 $0\log(a/0)=+\infty$.

2. 考虑关于 n 的正实数序列 $a_1(n),\ldots,a_k(n)$ 以及 $b_1(n),\ldots,b_k(n)$，假设对所有 i，都成立 $\lim_{n\to\infty} a_i(n)/b_i(n)=1$，证明：

$$\lim_{n\to\infty}\frac{a_1(n)+\cdots+a_k(n)}{b_1(n)+\cdots+b_k(n)}=1.$$

由此证明式 (3.1).

3. 证明命题 3.1.

4. 用 Lagrange 乘子法重新证明命题 3.4.

提示：如果你不知道 Lagrange 乘子法，可以参考 **??**.

5. 证明命题 3.8.

6. [Tin62] 仿照集合论的思路，我们可以定义三个随机变量的互信息为：

$$I(X;Y;Z)=I(X;Y)-I(X;Y|Z).$$

(1) 证明对称性：$I(X;Y;Z)=I(Y;X;Z)=I(X;Z;Y)$.

(2) 举一个例子说明，可能会有 $I(X;Y;Z)<0$，所以这样定义的互信息并不一定真的代表“信息量”.

7. 举一个例子说明，即便 $D(p_1\|p_2)$ 很接近 0，$D(p_2\|p_1)$ 也可能会很大.

8.（单变量数据处理不等式）对任意离散随机变量 X 和函数 f，证明：$H(X)\geq H(f(X))$.

9. 考虑二分类的学习问题，此时对单个样本我们观察到的结果要么是 0 或 1，假设在真实世界中样本总体服从参数为 θ 的 Bernoulli 分布，即 $\Pr(X=1)=1-\Pr(X=0)=\theta$. 假设我们的数据集是 $(x_1,y_1),\ldots,(x_N,y_N)$，他们是从总体中独立采样得到的.

(1) 将问题考虑成一个数理统计问题，估计 θ. 写出似然函数$L(\theta;y_1,\ldots,y_N)$.

(2) 再将问题考虑为一个信息论问题，写出每个样本的真实分布与估计分布之间的交叉熵之和 $CH(\theta;y_1,\ldots,y_N)$.

(3) 证明：$\max_\theta L(\theta;y_1,\ldots,y_N)=\min_\theta CH(\theta;y_1,\ldots,y_N)$，也就是说，最大似然估计等价于最小化交叉熵.

10. 请查找文献回答以下问题:

(1) Fisher 信息量是什么? 它与 K-L 散度有什么样的关系?

(2) 列举其他概率分布之间散度的概念，他们是否是度量?

(3) 列举概率分布之间的度量，他们之间是否有关联?

§3.5 章末注记

信息一词的英文是“information”，从动词“inform”来，意思是告知、通知. 早在 15 世纪中叶, “information”一词的出现了义项“在通信中针对特定主题的知识”.[Inf] 这说明在那个时候人类就已经意识到，通信会产生新的东西，被称为知识或信息. 然而，人类对信息的严谨探索起步晚得多. 关于信息的物理学讨论源自统计力学，Boltzmann 提出了著名的熵，证明了 H 定理，以此给出了热力学第二定律的微观解释. 关于 Boltzmann 的工作，参见 [Uff22].

一般认为，现代信息论的起源是 Shannon 的论文 [Sha48]，他在论文中提出了信息的数学定义，以及信息的基本性质. 但是，Shannon 的工作并不是孤立的，他的工作是在统计力学的基础上发展起来的. 事实上，Shannon 在论文中也提到了 Boltzmann 的熵. 这篇工作也被视为通信理论以及编码理论的奠基性工作. Shannon 在这篇论文中还给出了渐近意义下达到理论下界的最优编码，并且独立地被 Fano [Rob49] 以一种不同的形式发现，因此后世称为 Shannon-Fano 编码. 但是 Shannon-Fano 编码并不是精确地达到下界，实际上，最优编码是 Huffman [Huf52] 给出的. Shannon 在这篇论文中还讨论了渐近等分性，后来 MciMillan 的工作 [McM53] 和 Breiman 的工作 [Bre57] 拓展了这一结果，因此后世称为 Shannon-McMillan-Breiman 定理.

关于信息论与集合论的关系工作，可以参见 Hu Kuo Ting 的工作 [Tin62]. 他的工作还给出了多个随机变量互信息的定义，在这一章习题中有涉及.

相对熵的概念依然是从 Shannon 的奠基性论文 [Sha48] 中提出的，但他只局限于通信的问题. 更加一般的讨论是由 Kullback 和 Leibler 在 [KL51] 给出，他们的是一种数理统计的思路，但是他们也具体地讨论了这一概念与信息的关系. 他们的论文中也讨论了交叉熵这一概念.

机器学习中编码器和解码器的思路，最早是由 Rumelhart，Hinton 和 Williams 在 [RHW86] 中提出，他们将编码器和解码器的整体称作*自编码器*. 这篇工作几乎可以被视为深度学习的开山之作，它还提出了训练神经网络最常用的反向传播算法.

关于信息论的经典教科书，可以参见 [CT12]，此外，概率论的教材中也有很多很好的讨论，比如 [Jay02]，[Shi96] 以及 [李 10].

关于 Kolmogorov 复杂度的讨论，可以参见专著 [?]，这本书对于随机、信息、编码、复杂度，乃至归纳推理等概念都有非常独到的见解，值得一读.

第四章　Johnson-Lindenstrauss 引理

我们已经在上一章看到，使用概率分布建模的信息论在机器学习中起到了举足轻重的作用. 基于概率论的信息论总是考虑一个集合的对象的信息量，因此**数据**成为了这种方法论的核心前提：数据表征了一个集合的对象的某一特征. 在这一章，我们将探讨机器学习中数据的特性，以及一种重要的数据压缩的原理：Johnson-Lindenstrauss 引理. 证明这一引理所用到的概率论技术是**矩法**，这是*机器学习理论*中最为核心的几个技术之一. 因此本章也可以看做机器学习理论的一个引论.

§4.1　机器学习中的数据

从编码的角度来说，数据最简单的表示方法是使用固定长度的字符串. 比如说，人的生理性别有男或者女两种，于是我们可以用字符串 0 表示男，1 表示女. 这样，我们就可以用一个长度为 1 的字符串来表示人的生理性别. 人的属性还有很多，比如说年龄、身高、体重、学历、职业等等，这些属性都可以分别用固定长度的字符串来表示. 于是，一个人就被抽象为了一个固定长度的字符串.

然而，这种表示方式必须要假定数据只取有限个值. 有时候，为了简化建模和计算，我们还会考虑可以取无限个值的数据. 我们看一个具体的例子，人的身高. 从现代物理的角度来说，身高的变化是离散的，它有一个最小变化的单位. 从生物学的角度来说，身高是有上界的，比如说所有人的身高都不会超过十米. 因此，从理论上说，身高也只能取有限个值，所以也可以用字符串来表示. 然而，更加方便的方式是假定身高是一个*非负实数*，因此用一个数而不是一个字符串来表示.

因此，更加常见的情况下，我们会用*实数*或者*整数*来编码数据. 此时，将对象的多种属性按顺序排在一起，我们就得到了一个*向量*. 总而言之，在机器学习的框架，*数据被表示成数值向量*. 例如，要表示一个人的年龄、身高、体重、学历、职业，我们需要

1. 身高使用厘米作为单位；体重用千克作为单位；给学历编一个编号，比如 0 是高中，1 是本科，2 是硕士，3 是博士，-1 是其他；给职业也编号，例如 1 表示提示词工程师.
2. 用一个五维向量来表示年龄、身高、体重、学历、职业. 例如，$(20, 180, 70, 2, 1)$ 表示一个年龄为 20 岁，身高 180 厘米，体重 70 千克，学历为硕士，职业为提示词工程师的人.

注. *[lhy: 介绍一下计算机中对数值的编码].*

机器学习中，如此表示数据具备了独特的性质，一言以蔽之：**维数高，但是稀疏.** [lhy: 介绍一下高维高斯分布的特点，以及图像处理中数据稀疏的特点.]

§4.2 矩法与集中不等式

我们先引入示性函数的概念.

定义 4.1 (示性函数) 对事件 A，定义 A 的示性函数为一个从样本空间 Ω 到 $\mathbb{R}$ 的随机变量：

$$I(A)(\omega) := \begin{cases} 1, & \omega \in A. \\ 0, & \omega \notin A. \end{cases}$$

从定义就可以得到如下基本性质：

命题 4.1 设 A, B 是两个事件，则

1. $I(AB) = I(A)I(B)$.
2. $I(A)^2 = I(A)$.
3. $I(A \cup B) = I(A) + I(B) - I(AB)$.

证明 这里只作为一个示意，证明第三点，其他都类似. 我们需要证明，对任意样本点 $\omega \in \Omega$，我们有

$$I(A \cup B)(\omega) = I(A)(\omega) + I(B)(\omega) - I(AB)(\omega).$$

假设 $\omega \in A \cup B$，那么左边等于 1. 我们分类讨论：

- 如果 $\omega \in A$，那么右边第一项为 1.
  - 如果 $\omega \in B$，那么右边第二项为 1. 此时自然也有 $\omega \in AB$，所以右边第三项为 1，因此右边等于 1，等于左边.
  - 如果 $\omega \notin B$，那么右边第二项为 0. 此时自然也有 $\omega \notin AB$，所以右边第三项为 0，因此右边等于 1，等于左边.
- 如果 $\omega \notin A$，那么右边第一项为 0. 此时必须有 $\omega \in B$，所以右边第二项为 1. 但是此时自然也有 $\omega \notin AB$，所以右边第三项为 0，因此右边等于 1，等于左边.

如果 $\omega \notin A \cup B$，讨论类似，这里不再赘述. □

示性函数之所以重要，是因为它联系了期望与概率. 我们先来看一个显然的命题：

命题 4.2 设 A 是一个事件，则

$$\mathbb{E}[I(A)] = \Pr(A).$$

示性函数可以把对概率的计算变成对期望的计算. 回忆期望的线性性：设 $a, b \in \mathbb{R}$，X, Y 是有期望的随机变量，那么成立

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

利用期望的线性性，示性函数可以导出很多概率恒等式与不等式. 例如：容斥公式

$$\begin{aligned}\Pr(A \cup B) &= \mathbb{E}[I(A \cup B)] = \mathbb{E}[I(A) + I(B) - I(AB)] \\ &= \mathbb{E}[I(A)] + \mathbb{E}[I(B)] - \mathbb{E}[I(AB)] \\ &= \Pr(A) + \Pr(B) - \Pr(AB).\end{aligned}$$

对于概率论以及机器学习理论来说，下面的这个不等式非常重要：

定理 4.1 (Markov 不等式) 如果 X 是非负有期望的随机变量，$a > 0$，那么

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

证明 直接利用示性函数，我们有：

$$\begin{aligned}\mathbb{E}[X] =& \mathbb{E}[XI(X \geq a) + XI(X < a)] \\ =& \mathbb{E}[\underbrace{XI(X \geq a)}_{\geq aI(X \geq a)}] + \mathbb{E}[\underbrace{XI(X < a)}_{\geq 0}] \\ \geq& a\mathbb{E}[I(X \geq a)] = a\Pr(X \geq a).\end{aligned}$$ □

注. 为了使得证明有效，我们必须要假设上面的推导中出现的期望都是存在的，当然这实际上很容易验证. 为了避免不必要的技术细节，在后面的所有证明以及推导中，我们都会默认写出来的期望是存在的，不再赘述.

我们利用 Markov 不等式可以直接得到以下结果.

推论 4.1 (Chebyshev 不等式) 设 X 是任意有方差的随机变量，那么对任意 $a > 0$，成立

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathsf{Var}(X)}{a^2}.$$

证明 设 $Y = (X - \mathbb{E}[X])^2$，$t = a^2$，那么 Y 是非负随机变量，且 $\mathbb{E}[Y] = \mathsf{Var}(X)$，于是由 Markov 不等式，我们有

$$\begin{aligned}\Pr(|X - \mathbb{E}[X]| \geq a) &= \Pr(|X - \mathbb{E}[X]|^2 \geq a^2) \\ &= \Pr(Y \geq t) \\ &\leq \frac{\mathbb{E}[Y]}{t} = \frac{\mathsf{Var}(X)}{a^2}.\end{aligned}$$ □

Chebyshev 不等式告诉我们采样到偏离其期望的概率有一个上界. 像这样利用矩（即 $\mathbb{E}[f(X)]$）来估计概率上界的方法被称为*矩法*.

实际上，很多情况下，偏离期望是非常小概率的事件，远小于上面的估计值. 为了得到更精确的上界，我们需要一些技巧. 考虑任意随机变量 X，对 $\lambda > 0$，

$$X \geq a \iff \lambda X \geq \lambda a \iff e^{\lambda X} \geq e^{\lambda a}.$$

由 Markov 不等式（如何得到?），

$$\Pr(X \geq a) = \Pr\left(e^{\lambda X} \geq e^{\lambda a}\right) \leq e^{-\lambda a} \cdot \mathbb{E}\left[e^{\lambda X}\right].$$

注意到这个不等式应该对任意 $\lambda > 0$ 成立，所以

$$\Pr(X \geq a) \leq \inf_{\lambda > 0} e^{-\lambda a} \cdot \mathbb{E}\left[e^{\lambda X}\right].$$

以上方法可以得到概率更精确的上界. 这样用指数进行推导的方法称为*指数矩*或*Cramér-Chernoff 方法*.

利用指数矩，我们可以更加精确地研究 Chebyshev 不等式中随机变量所表现出来的性质，这种性质被称为概率的*集中性*. 我们可以用*集中不等式*来刻画这样的性质. 这样的不等式描述随机变量 X 有多大概率偏离某个值 μ 多少值（t），它表现为

$$\Pr(|X - \mu| \geq t) \leq \text{小量}.$$

通常来说，μ 是随机变量的期望或者中位数，在这本书中，只会讨论关于期望的集中性. 我们可以看到 Chebyshev 不等式就是一种特殊的集中不等式，但是它的界太松. 利用指数矩，我们将证明界更紧的 Hoeffding 不等式和 Chernoff 不等式.

定理 4.2 (Hoeffding 不等式) 设 $X_1, \ldots, X_n$ 相互独立且服从对称 *Bernoulli* 分布，即 X_i 满足 $\Pr(X_i = 1) = 1 - \Pr(X_i = -1) = 1/2$. 考虑向量 $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$，对任意 $t \geq 0$，我们有

$$\Pr\left(\sum_{i=1}^{n} a_i X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right).$$

证明 由指数矩，我们有

$$\begin{aligned}
\Pr\left(\sum_{i=1}^{n} a_i X_i \geq t\right) &= \Pr\left(\exp\left(\lambda \sum_{i=1}^{n} a_i X_i\right) \geq \exp(\lambda t)\right) \\
&\leq e^{-\lambda t}\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} a_i X_i\right)\right] \\
&= e^{-\lambda t}\prod_{i=1}^{n}\mathbb{E}\left[\exp(\lambda a_i X_i)\right].
\end{aligned}$$

这个不等式对任意 $\lambda > 0$ 都成立. 利用 $X_1, \ldots, X_n$ 服从对称 Bernoulli 分布，得到（习题）[lhy: 习题]

$$e^{-\lambda t}\prod_{i}\mathbb{E}\left[\exp(\lambda a_i X_i)\right] \leq \exp\left(-\lambda t + \frac{\lambda^2}{2}\sum_{i} a_i^2\right). \tag{4.1}$$

由于这一不等式对任意 $\lambda > 0$ 都成立，根据二次函数的性质，取 $\lambda = t/\sum_i a_i^2$，可得

$$\begin{aligned}
\inf_{\lambda>0}\exp\left(-\lambda t + \frac{\lambda^2}{2}\sum_{i} a_i^2\right) &= \exp\left(-\frac{t}{\sum_i a_i^2}t + \frac{1}{2}\left(\frac{t}{\sum_i a_i^2}\right)^2\sum_{i} a_i^2\right) \\
&= \exp\left(-\frac{t^2}{2\|a\|_2^2}\right).
\end{aligned}$$

□

利用相同的证明技巧，我们可以证明一般形式的 Hoeffding 不等式，我们把证明留作习题.[lhy: 习题]

定理 4.3 (Hoeffding 不等式，一般情形) 设 $X_1, \ldots, X_n$ 是相互独立的随机变量，对任意 i 都成立 $X_i \in [m_i, M_i]$. 那么对任意 $t \geq 0$，我们有

$$\Pr\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(M_i - m_i)^2}\right).$$

下面我们介绍 Chernoff 不等式.

定理 4.4 (Chernoff 不等式) 设 $X_1, \dots, X_n$ 是相互独立的随机变量，分别服从于参数为 $p_1, \dots, p_n$ 的 *Bernoulli* 分布. 记 $\sum_{i=1}^n X_i$ 的期望为 $\mu = \sum_{i=1}^n p_i$，对于任意 $t > \mu$，我们有

$$\Pr\left(\sum_{i=1}^n X_i \geq t\right) \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t.$$

这里 e 是自然对数的底数.

证明 和证明 Hoeffding 不等式的第一步相同，我们先利用指数矩，对任意 $\lambda > 0$ 有

$$\Pr\left(\sum_{i=1}^n X_i \geq t\right) \leq e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\left[\exp(\lambda X_i)\right].$$

然后，将 $\prod_{i=1}^n \mathbb{E}[\exp(\lambda X_i)]$ 进一步放缩：

$$\begin{aligned}\prod_{i=1}^n \mathbb{E}\left[\exp(\lambda X_i)\right] &= \prod_{i=1}^n \left(e^\lambda p_i + (1-p_i)\right)\\ &\leq \prod_{i=1}^n \exp\left((e^\lambda - 1)p_i\right).\end{aligned}$$

因此

$$\begin{aligned}\Pr\left(\sum_{i=1}^n X_i \geq t\right) &\leq e^{-\lambda t}\prod_{i=1}^n \exp\left((e^\lambda - 1)p_i\right)\\ &= e^{-\lambda t}\exp\left((e^\lambda - 1)\sum_{i=1}^n p_i\right)\\ &= \exp\left(\mu e^\lambda - t\lambda - \mu\right).\end{aligned}$$

右边的最小值在 $\lambda = \log(t/\mu)$ 取得，代入得到：

$$\Pr\left(\sum_{i=1}^n X_i \geq t\right) \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t.$$

□

§4.3 J-L 引理的陈述与证明

有了上面矩法的准备，我们可以陈述并证明 J-L 引理了.

定理 4.5 (Johnson-Lindenstrauss 引理) 给定 N 个单位向量 $v_1, \ldots, v_N \in \mathbb{R}^m$ 和 $n > 24 \log N/\epsilon^2$，随机矩阵 $A \in \mathbb{R}^{n\times m}$ 每个元素独立重复采样自 $\mathcal{N}(0, 1/n)$，$\epsilon \in (0, 1)$ 是给定常数，那么至少有 $(N-1)/N$ 的概率，使得对所有的 $i \neq j$，都成立

$$(1-\epsilon)\left\|v_i - v_j\right\|_2^2 < \left\|Av_i - Av_j\right\|_2^2 < (1+\epsilon)\left\|v_i - v_j\right\|_2^2.$$

我们可以把 n 理解成降维后的维度，Av_i 是降维后的向量. 这个引理告诉我们只要 $n > 24 \log N/\epsilon^2$，我们就可以用变换 A 把原本 m 维的向量映射到 n 维空间，并且保证它们相对距离的偏离不超过 ϵ，因此我们可以把 A 看成一个损失率很低的压缩变换. 不严格地说，塞下 N 个向量，只需要 $\mathcal{O}(\log N)$ 维空间.

下面我们开始证明 J-L 引理. 为了看出来证明的思路，我们第一个任务是算出压缩后 Av_i 的分布. 我们首先回忆一些正态向量的基本性质. 关于正态向量的讨论，可以参考**??**.

命题 4.3 假设 $u \sim \mathcal{N}(\mu, \Sigma)$ 是一个 n 维正态向量，M 是一个 $m \times n$ 矩阵，那么 Mu 是一个 m 维正态向量，并且 $Au \sim \mathcal{N}(M\mu, M\Sigma M^T)$.

利用这一个命题，很容易可以得到 Av_i 的分布：

引理 4.1 假设 $u \in \mathbb{R}^m$ 是一个单位向量，那么 $Au \sim \mathcal{N}(0, n^{-1}I_n)$.

证明 将 A 视作一个 mn 维的正态向量，注意到，$(Au)_i = \sum_{j=1}^m A_{ij}u_j$，所以 Au 是一个从向量 A 线性变换得到的向量. 根据命题 4.3，Au 是一个正态向量，只需计算它的期望和协方差矩阵.

注意到，对不同的 i，向量 $(A_{ij})_j$ 相互是独立的，所以分量 $(Au)_i$ 相互也是独立的，因此只需要计算正态变量 $(Au)_i$ 的期望与方差. 其期望为 $\sum_{j=1}^m 0 \cdot u_j = 0$，方差为

$$\sum_{j=1}^m \left(\frac{1}{n} \cdot u_j^2\right) = \frac{1}{n}.$$

所以 Au 的期望是 0，协方差矩阵是 $n^{-1}I_n$. □

然而，我们关心的其实不单单是 Av_i 的分布，更重要的其实是 $Av_i - Av_j$ 的分布，即压缩后的向量之间的相对距离，幸运的是，我们并不需要做额外的什么计算，我们直接有如下结果：

引理 4.2 向量 $u = \frac{v_i - v_j}{\left\|v_i - v_j\right\|_2}$ 是一个单位向量，因此 $Au \sim \mathcal{N}(0, n^{-1}I_n)$.

J-L 引理实际上在说，$\|Au\|_2$ 偏离 1 的一定程度的概率是非常小的. 于是，为了证明 J-L 引理，我们最重要的任务是给出 Au 这样向量模长的集中不等式：

引理 4.3 (单位模引理) 设 $u \sim \mathcal{N}(0, n^{-1}I_n)$，$\epsilon \in (0,1)$ 是给定的常数，那么我们有

$$\Pr(|\,\|u\|_2^2 - 1| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2 n}{8}\right).$$

注意到 $\mathbb{E}[\|u\|_2^2] = n \cdot (1/n) = 1$，所以这个引理在说高维空间中，如果正态向量具有单位模长平方期望，那么它的模长就会集中在单位长度附近，因此称为单位模引理.

证明 $|\,\|u\|_2^2 - 1| \geq \epsilon$ 发生有两种可能，$\|u\|_2^2 - 1 \geq \epsilon$ 和 $1 - \|u\|_2^2 \geq \epsilon$. 我们先来计算 $\|u\|_2^2 - 1 \geq \epsilon$ 的概率，根据指数矩，

$$\Pr\left(\|u\|_2^2 - 1 \geq \epsilon\right) \leq \inf_{\lambda > 0}\left\{e^{-\lambda(\epsilon+1)}\mathbb{E}\left[e^{\lambda\|u\|_2^2}\right]\right\}.$$

因为 u 的各个分量是相互独立的，所以我们可以把 $\|u\|_2^2$ 展开

$$\mathbb{E}\left[e^{\lambda\|u\|_2^2}\right] = \mathbb{E}\left[e^{\lambda\sum_i u_i^2}\right] = \mathbb{E}\left[\prod_i e^{\lambda u_i^2}\right] = \prod_i \mathbb{E}\left[e^{\lambda u_i^2}\right].$$

可以算得 $\mathbb{E}\left[e^{\lambda u_i^2}\right] = \sqrt{n/(n-2\lambda)}$[lhy: 习题]，所以

$$\Pr\left(\|u\|_2^2 - 1 \geq \epsilon\right) \leq \inf_{\lambda > 0}\left\{e^{-\lambda(\epsilon+1)}\left(\frac{n}{n-2\lambda}\right)^{n/2}\right\}.$$

可以验证最小值在 $\lambda = n\epsilon/(2(1+\epsilon))$ 处取到，代入可得

$$\Pr\left(\|u\|_2^2 - 1 \geq \epsilon\right) \leq e^{n(\log(1+\epsilon)-\epsilon)/2} \leq e^{-n\epsilon^2/8}.$$

这里最后一个不等号使用了不等式 $\log(1+\epsilon) \leq \epsilon - \epsilon^2/4$.

计算 $1 - \|u\|_2^2 \geq \epsilon$ 的概率的过程和 $\|u\|_2^2 - 1 \geq \epsilon$ 几乎完全相同的，可以得到

$$\Pr\left(1 - \|u\|_2^2 \geq \epsilon\right) \leq e^{n(\log(1-\epsilon)+\epsilon)/2} \leq e^{-n\epsilon^2/8}.$$

$$\begin{aligned}\Pr\left(|\,\|u\|_2^2 - 1| \geq \epsilon\right) &\leq \Pr\left(\|u\|_2^2 - 1 \geq \epsilon\right) + \Pr\left(1 - \|u\|_2^2 \geq \epsilon\right) \\ &\leq 2e^{-n\epsilon^2/8}.\end{aligned}$$

□

有了单位模引理，我们就可以很容易证明 J-L 引理了. 将引理 4.2 中的 u 带入单位模引理，得到

$$\Pr\left(\left|\left\|\frac{A(v_i - v_j)}{\|v_i - v_j\|_2}\right\|_2^2 - 1\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2 n}{8}\right).$$

这个结论对任意 $i \neq j$ 成立，因此遍历所有 i, j 对，可得

$$\begin{aligned}
\Pr\left(\exists (i,j): \left| \left\| \frac{A(v_i - v_j)}{\left\| v_i - v_j \right\|_2} \right\|_2^2 - 1 \right| \geq \epsilon \right) &\leq 2 \sum_{i \neq j} \exp\left(-\frac{\epsilon^2 n}{8}\right) \\
&= 2\binom{N}{2} \exp\left(-\frac{\epsilon^2 n}{8}\right).
\end{aligned}$$

换言之，对任意 i, j，$\left| \left\| \frac{A(v_i - v_j)}{\left\| v_i - v_j \right\|_2} \right\|_2^2 - 1 \right| < \epsilon$ 都成立的概率不小于

$$1 - 2\binom{N}{2} \exp\left(-\frac{\epsilon^2 n}{8}\right) = 1 - N(N-1) \exp\left(-\frac{\epsilon^2 n}{8}\right).$$

代入 $n > \frac{24 \log N}{\epsilon^2}$，可得这一概率

$$1 - N(N-1) \exp\left(-\frac{\epsilon^2 n}{8}\right) \geq 1 - N(N-1) N^{-3} \geq 1 - N^{-1} = \frac{N-1}{N}.$$

很多时候，我们关心的并不是向量间的距离，而是向量的内积（比如使用余弦度量的时候），这时候我们可以使用内积版本的 J-L 的引理：

定理 4.6 (J-L 引理，内积形式) 给定 N 个单位向量 $v_1, \ldots, v_N \in \mathbb{R}^m$ 和 $n > 24 \log N / \epsilon^2$，随机矩阵 $A \in \mathbb{R}^{n \times m}$ 每一个元素都独立重复采样自 $\mathcal{N}(0, 1/n)$，$\epsilon \in (0,1)$ 是给定常数，那么至少有 $(N-1)/N$ 的概率，使得对所有的 $i \neq j$，都成立

$$|\langle Av_i, Av_j \rangle - \langle v_i, v_j \rangle| < \epsilon.$$

证明 由原始 J-L 引理可知，至少有 $\frac{N-1}{N}$ 的概率满足对于任意 $i \neq j$ 有：

$$\begin{aligned}
(1-\epsilon) \left\| v_i - v_j \right\|_2^2 < \left\| Av_i - Av_j \right\|_2^2 < (1+\epsilon) \left\| v_i - v_j \right\|_2^2, \\
(1-\epsilon) \left\| v_i + v_j \right\|_2^2 < \left\| Av_i + Av_j \right\|_2^2 < (1+\epsilon) \left\| v_i + v_j \right\|_2^2.
\end{aligned}$$

我们将第一行乘 -1 加到第二行可以得到

$$4\langle v_i, v_j \rangle - 2\epsilon(\|v_i\|_2^2 + \|v_j\|_2^2) < 4\langle Av_i, Av_j \rangle < 4\langle v_i, v_j \rangle + 2\epsilon(\|v_i\|_2^2 + \|v_j\|_2^2).$$

因为 v_i, v_j 是单位向量，所以上式等价于 $|\langle Av_i, Av_j \rangle - \langle v_i, v_j \rangle| < \epsilon$. □

注. *[lhy: 讨论渐近等分性、指数矩、大偏差理论之间的关系.]*

§4.4 J-L 引理的应用

回顾：J-L 引理描述的是对于 N 个向量，我们可以将它们降到 $\mathcal{O}(\log N)$ 维空间，并将相对距离的误差控制在一定范围内. 它的内容本身就和降维相关，所以最基本的应用就是直接作为降维方法. 许多其它算法例如局部敏感哈希（LSH）、随机 SVD，本质上也都依赖 J-L 引理. 除此之外，J-L 引理对机器学习模型中维度的选择提供了一些理论解释. 下面我们将介绍两个具体的应用案例.

例 4.1 (词向量维度) *在 NLP 的发展中产生了像 Word2Vec、GloVe 这样经典的词向量模型和基于注意力机制的各种大语言模型. 这里一个问题自然的问题是，当我们对 N 个单词进行建模，词向量的维度选择多少比较合适？如果维度过高，会使得后续的计算变得更加复杂. 如果维度过低，会无法完全表达出这些单词本身的信息. 对于这一问题，J-L 引理给出了一个比较直接的结论，$\mathcal{O}(\log N)$ 空间足以容纳下 N 个单词. 但是要注意，这一结论成立的前提是正态随机矩阵，然而单词的空间是否符合正态分布是不知道的，所以这一结果只是从理论上给了一个直观，选择什么样的 n 还是由具体的实验效果来决定.*

例 4.2 (多头注意力) *在注意力机制中，我们往往会先把* `head_size` *降低到 64 再做内积. 那么一个很自然的问题是，*`head_size` *为 64 的注意力机制是否足以拟合任何概率分布？具体来说，注意力的计算公式为*

$$a_{ij} = \frac{e^{\langle q_i, k_j \rangle}}{\sum_{j=1}^{L} e^{\langle q_i, k_j \rangle}}.$$

其中 $q_i, k_j \in \mathbb{R}^d$.

我们希望能够做到：给定任意的概率矩阵 (p_{ij})，上述 (a_{ij}) 都能够很好的逼近 (p_{ij}). 换言之，给定 (p_{ij}) 和维度 d，我们是否能找到一组 $q_1, \ldots, q_L, k_1, \ldots, k_L \in \mathbb{R}^d$，使得对应项 a_{ij} 与 p_{ij} 足够接近. 其实这就和词向量模型的维度选择问题是等价的. 词向量的维度变成了`head_size`*，词表大小变成了序列长度. J-L 引理告诉我们的答案依然是只需要 $\mathcal{O}(\log N)$ 的空间就足以容纳下 N 个单词，一个很粗糙的计算，*

[lhy: 这两个例子写得更详细一些.]

§4.5 习题

§4.6 章末注记

第五章　差分隐私

在本章我们关心数据的另一个维度：社会属性．机器学习需要大量的数据，这些数据从何而来？大部分时候，是通过收集个体的数据得到的．于是这里就涉及到了隐私的问题，如何保护个体的隐私，同时又能够让机器学习得到足够的数据？**差分隐私**就是解决这个问题的一个方法，本章将更详细地介绍差分隐私的概念和应用．

§5.1　数据隐私问题

许多科研工作的开展和推进都需要有大量真实有效的数据作为支撑．以医学为例，我们需要大量真实的病人提供病情数据．但这些数据可能都涉及到病人的隐私信息，例如一些*敏感数据*．因此，我们必须找到一种方法，既可以收集到这些数据，又保护病人的隐私信息．

保护病人的隐私信息的一种理解是不会让数据获得者将数据和人对应起来．那么最直观的想法就是将每条*数据匿名化*．比如，每条数据只包含患者的生日、性别、邮政编号（代表位置）、病情．这样做依然有问题：同一天生日、同一性别、相同邮政编号的人很少，而这些信息很容易被找到，比如公开的选民名册．于是通过一条数据的各种属性可以轻易定位到这个人，这样的匿名化是不安全的．

我们再看一个例子。2006 年，Netflix 举办了关于电影推荐系统的算法设计比赛．只公开了匿名代号和对应用户的观看电影名称、打分的数据集（这次连生日、性别都没有）．但这一数据集很快被破解，问题出在只要对某个用户稍微熟悉一些，就很容易对应出这个用户和观影数据．这种破解让部分用户感到焦虑，例如性少数群体害怕其他人可以从自己的观影记录中判断出自己的性取向．第二届 Netflix Prize 竞赛也因此停办．

这两个例子展现的是一种*去匿名化*的现象，也就是说匿名的数据实际上揭示了数据对应的那个个体。这去匿名化的出现都是因为数据具有*独特性*．比如我们看表 5.1，这是医院甲的病人数据表。56 岁的病人只有 Rebecca，所以假如我们知道 Rebecca 的年龄并了解到她去过这家医院，便立即得知她患有 HIV.

Name	Age	Gender	Zip Code	Smoker	Diagnosis
Richard	64	Male	19146	Y	Heart disease
Susan	61	Female	19118	N	Arthritis
Matthew	67	Male	19104	Y	Lung cancer
Alice	63	Female	19146	N	Crohn's disease
Thomas	69	Male	19115	Y	Lung cancer
Rebecca	**56**	**Female**	19103	N	HIV
Tony	52	Male	19146	Y	Lyme disease
Mohammed	59	Male	19130	Y	Seasonal allergies
Lisa	55	Female	19146	N	Ulcerative colitis

[lhy: 改成中文版]

表 5.1: 医院甲的病人数据表。

一种减少独特性的思想是k-匿名性：任何一个人的信息都不能和其他至少 $(k-1)$ 人区分开. 比如，可以不明确写出姓名、年龄和邮编，只给出模糊的范围，于是数据变成了下面的表 5.2.

Name	Age	Gender	Zip Code	Smoker	Diagnosis
*	60-70	Male	191**	Y	Heart disease
*	60-70	Female	191**	N	Arthritis
*	60-70	Male	191**	Y	Lung cancer
*	60-70	Female	191**	N	Crohn's disease
*	60-70	Male	191**	Y	Lung cancer
*	**50-60**	**Female**	191**	N	HIV
*	50-60	Male	191**	Y	Lyme disease
*	50-60	Male	191**	Y	Seasonal allergies
*	**50-60**	**Female**	191**	N	Ulcerative colitis

[lhy: 改成中文]

表 5.2: 医院甲的病人数据表，模糊了姓名、年龄和邮编。

这种方法仍然存在问题，因为我们不能把关键信息（病症信息）也模糊化. 如果我们还拿到了另一家医院乙的模糊之后的病人数据表（表 5.3），那么依然有办法定位到

Rebecca：这两张表上 50-60 岁的女性只有 HIV 是重合的，如果我们知道 Rebecca 的年龄并知道她同时去过两家医院，便立即得知她患有 HIV.

Name	**Age**	**Gender**	**Zip Code**	**Diagnosis**
*	**50-60**	**Female**	191**	HIV
*	**50-60**	**Female**	191**	Lupus
*	**50-60**	**Female**	191**	Hip fracture
*	60-70	Male	191**	Pancreatic cancer
*	60-70	Male	191**	Ulcerative colitis
*	60-70	Male	191**	Flu-like symptoms

[lhy: 改成中文]

表 5.3: 医院乙的病人数据表，模糊了姓名、年龄和邮编。

除了使用匿名化的手段，还有一种方法可以保护隐私：不再提供单人的数据，而是直接公布将数据集的总体信息，比如平均值. 但这种方法也不一定能保证不泄露单人数据. 例如：我们只公布一个机器学习模型（这算是一直非常抽象的总体信息）. 很多研究表明可以通过尝试不同的测试数据来判断出这个机器学习模型的训练集，这是因为机器学习模型总是会偏向于过拟合训练集数据. 所以如果对某个测试数据的结果很有自信，往往说明这一数据存在于训练集中. [lhy: 给一个 ref]

> **注.** 以上这些内容都说明**匿名化**很难保护个人隐私. 那么是否可以使用密码学的手段进行**加密**？其实加密和隐私在出发点上完全不同. 加密的目的是为了不让别人获取到真实数据. 而隐私是一个比简单地锁定数据更微妙的问题——我们希望我们的算法的结果能够释放有用的信息，而不是泄露私人信息. 因此我们这里讨论的其实是隐私保护问题而不是加密问题。

§5.2 差分隐私的定义与性质

我们上面探讨了隐私保护的必要性以及它的微妙之处，现在我们要给出一种合理的方案解决隐私保护的问题，这个方案就是*差分隐私*。要给出一个数学模型，不仅要知道什么情况下算是隐私泄漏，也需要知道什么情况下不算，所以我们再来看一个反面的例子。

Broky 是一位长期吸烟的男子，他参加了一项有关“吸烟与健康”的调查. 这项调查在不久后发布了一项结果，表明长期吸烟的人患上肺癌的几率更大. 伴随着这一结果的公

布，保险公司在出售相同保险时会对长期吸烟者索要更高的价格. Broky 当然也受到了这一政策的影响. 那我们是否可以认为这项研究泄露了 Broky（更有可能患病）的隐私呢?

我们的直观应该是不算泄露了隐私. 因为"长期吸烟的人患上肺癌的几率更大"这项结论并不依赖于 Broky 是否参加了调查. 考虑这样的对照，x 代表原来参加调查的人的集合，x' 代表其他人不变，只是 Broky 换成了另外一个人的集合. 如果是 x' 这些人参与了调查，结论是否会发生变化？大概率不会！

Broky 的例子告诉我们对于隐私的一种合理的衡量应该有以下性质：当数据集中包含 *Broky* 的信息，相比数据集中不包含 *Broky* 的信息，并不会**明显**增加损害 *Broky* 的利益的概率. 这一思想引出了*差分隐私*的概念，我们将在下面给出数学形式的定义.

考虑数据的空间 $\mathcal{X}$，其中的每一个元素都包含了个体的所有可能数据例如姓名、性别、年龄、国籍等. 考虑 n 个人的数据，形成了有序的数据集 $x = (x_1, \cdots, x_n) \in \mathcal{X}^n$. 设 A 是一种随机算法：在固定输入数据集 $x \in \mathcal{X}^n$ 下，$A(x)$ 是结果空间 $\mathcal{Y}$ 上的一个随机变量. 当我们改变（增加或删除）一个人的数据时，我们希望结果分布的变化可以控制在一定范围内. 为此，我们引入相邻数据集的概念：

定义 5.1 (k-相邻数据集) *设 $x, x' \in \mathcal{X}^n$，如果 x 和 x' 最多有 k 条数据不同，即至多存在 k 个不同的 $i_1, \ldots, i_k \in [n]$ 使得 $x_{i_j} = x'_{i_j}$ 对 $j \in [k]$ 成立，那么称 x 和 x' 是 k-相邻的.*

在之前的例子中，含有 Broky 的被调查者的数据集和把 Broky 换为任意一个其他人的数据集时 1-相邻数据集.

现在我们给出差分隐私的定义。

定义 5.2 (ϵ-DP) *考虑随机算法 $A : \mathcal{X}^n \to \mathcal{Y}$，如果对于任一对 1-相邻数据集 x, x'，对任意（可测）值集 $E \subseteq \mathcal{Y}$，有*

$$\Pr(A(x) \in E) \leq e^{\epsilon} \cdot \Pr(A(x') \in E),$$

*那么我们称 A 为*数据集大小为 n 的 ϵ-DP 算法.

需要注意的是，这一定义是针对随机算法的，并且是对称的，也就是 x 和 x' 的地位是平等的。直观上一个，ϵ-DP 算法的输出分布在相邻数据集上的变化不会太大. ϵ 衡量的是信息的泄漏量；ϵ 越大，算法泄漏的信息就越多，隐私保护效果当然也变差.

以上定义需要对所有的（可测）值集 E 都成立，这给验证带来了极大的困难，如果随机算法的输出分布是更加常规的，我们可以简化验证的过程.

对于离散型的输出，我们有如下等价定义：

命题 5.1 如果 $A(x)$ 对于任意 $x \in \mathcal{X}^n$ 都是离散型随机变量，那么随机算法 A 是数据集大小为 n 的 ϵ-*DP* 算法，当且仅当对于任意一对 1-相邻数据集 x, x' 和所有的 $y \in \mathcal{Y}$，有

$$\Pr(A(x) = y) \le e^{\epsilon} \cdot \Pr(A(x') = y).$$

证明 $\Longrightarrow$： 取 $E = \{y\}$ 即可证明.

$\Longleftarrow$： 假设 $E = \{y_1, \ldots, y_k, \ldots\}$. 对每一个 y_i，都有

$$\Pr(A(x) = y_i) \le e^{\epsilon} \cdot \Pr(A(x') = y_i).$$

因为 $A(\cdot) = y_i$ 对于不同的 i 是互斥事件，所以概率可以直接相加，于是:

$$\Pr(A(x) \in E) = \sum_i \Pr(A(x) = y_i) \le e^{\epsilon} \cdot \sum_i \Pr(A(x') = y_i) = e^{\epsilon} \cdot \Pr(A(x') \in E). \qquad \square$$

对连续型的输出，我们有如下等价定义:

命题 5.2 如果 $A(x)$ 对于任意 $x \in \mathcal{X}^n$ 都是连续型随机变量，那么它存在概率密度函数，记为 h_x. 此时随机算法 A 是数据集大小为 n 的 ϵ-*DP* 算法，当且仅当对于任意一对 1-相邻数据集 x, x' 和几乎所有的 $y \in \mathcal{Y}$，有

$$h_x(y) \le e^{\epsilon} \cdot h_{x'}(y).$$

证明 [1] $\Longrightarrow$： 根据概率密度函数的定义（实际上是 Lebesgue 微分定理），对 $x \in \mathcal{X}^n$，取 $E_\delta = (y - \delta, y + \delta)$，对几乎所有的 $y \in \mathcal{Y}$，有

$$\frac{\mathrm{d}}{\mathrm{d}\delta} \Pr(A(x) \in E_\delta) = h_x(y).$$

因此，对几乎所有的 $y \in \mathcal{Y}$，

$$\forall \delta > 0 \ \Pr(A(x) \in E_\delta) \le e^{\epsilon} \cdot \Pr(A(x') \in E_\delta) \implies h_x(y) \le e^{\epsilon} \cdot h_{x'}(y).$$

$\Longleftarrow$： 依然根据概率密度函数的定义，考虑 $x, x' \in \mathcal{X}^n$，对任意可测 $E \subseteq \mathcal{Y}$，有

$$\Pr(A(x) \in E) = \int_E h_x(y)\mathrm{d}y \le e^{\epsilon} \cdot \int_E h_{x'}(y)\mathrm{d}y = e^{\epsilon} \cdot \Pr(A(x') \in E). \qquad \square$$

接下来我们给出差分隐私的基本性质。

[1]这一部分的严格表述需要测度论的基础，所以这一证明从直观上理解即可，不需要考虑严格的定义。

命题 5.3 (复合性，两个算法的情形) A_1 和 A_2 是相互独立的随机算法，其中 $A_1 : X^n \to \mathcal{Y}_1$，$A_2 : \mathcal{Y}_1 \times \mathcal{X}^n \to \mathcal{Y}_2$. 假设 A_1 是 ϵ_1-*DP* 算法，A_2 是 ϵ_2-*DP* 算法.

令 $A : \mathcal{X}^n \to \mathcal{Y}_1 \times \mathcal{Y}_2$ 是随机算法，输出为 $A(x) = (y_1, y_2)$，其中 $y_1 = A_1(x)$，$y_2 = A_2(y_1, x)$，那么 A 是 $(\epsilon_1 + \epsilon_2)$-*DP* 算法.

证明 为了简化记号，这里我们只证明离散的情况. 令 x, x' 是 $\mathcal{X}^n$ 中的两个 1-相邻数据集，A 输出为 $y = (y_1, y_2) \in \mathcal{Y}_1 \times \mathcal{Y}_2$，那么根据定义和独立性，

$$\Pr(A(x) = (y_1, y_2)) = \Pr(A_1(x) = y_1) \cdot \Pr(A_2(y_1, x) = y_2).$$

由于 A_1 是 ϵ_1-DP 算法，A_2 是 ϵ_2-DP 算法，得到

$$\begin{aligned}\Pr(A(x) = (y_1, y_2)) &= \Pr(A_1(x) = y_1) \cdot \Pr(A_2(y_1, x) = y_2) \\ &\le e^{\epsilon_1} \Pr(A_1(x') = y_1) \cdot e^{\epsilon_2} \Pr(A_2(y_1, x') = y_2) \\ &= e^{\epsilon_1 + \epsilon_2} \cdot \Pr(A(x') = (y_1, y_2)).\end{aligned}$$

□

利用数学归纳法，很容易推广到多个随机算法的复合性：

命题 5.4 (复合性，多个算法的情形) 设 $A_1, A_2, \cdots, A_k$ 为一列相互独立的随机算法，

$$\begin{aligned}&A_1 : \mathcal{X}^n \to \mathcal{Y}_1, \\ &A_i : \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{i-1} \times \mathcal{X}^n \to \mathcal{Y}_i, \quad i = 2, 3, \cdots, k.\end{aligned}$$

也就是 A_i 将 $A_1, \cdots, A_{i-1}$ 的输出和 $\mathcal{X}^n$ 中的一个数据集作为输入元素. 对 $i = 1, \ldots, k$, A_i 是 ϵ_i-*DP* 算法.

依次运行算法 A_i 得到算法 $A : \mathcal{X}^n \to \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_k$，那么 A 是 ϵ-*DP*，其中 $\epsilon = \sum_{i=1}^{n} \epsilon_i$.

接下来的性质是说明操纵随机算法的输出不会影响隐私保护的效果：

命题 5.5 (后处理) 令 $A : \mathcal{X}^n \to \mathcal{Y}, B : \mathcal{Y} \to \mathcal{Z}$ 为相互独立的随机算法，其中 $\mathcal{X}$，$\mathcal{Y}$，$\mathcal{Z}$ 是任意集合. 如果 A 是 ϵ-*DP* 算法，那么组合算法 $B(A(\cdot))$ 也是 ϵ-*DP* 算法.

证明 我们仍然只考虑离散情形，采用定义的方法证明

$$\begin{aligned}\Pr(B(A(x)) = b) &= \sum_{y \in \mathcal{Y}} \Pr(A(x) = y) \Pr(B(y) = b) \\ &\le e^{\epsilon} \sum_{y \in \mathcal{Y}} \Pr(A(x') = y) \Pr(B(y) = b) \\ &= e^{\epsilon} \Pr(B(A(x')) = b).\end{aligned}$$

□

最后，我们讨论如果有多个人的数据都发生变化的时候，隐私保护的性质会发生什么变化。

命题 5.6 (群体隐私) *令 $x, x' \in \mathcal{X}^n$ 是 k-相邻数据集，$1 \le k \le n$. 如果 A 是 ϵ-DP 算法，那么对所有的值集 E，我们有*

$$\Pr(A(x) \in E) \le e^{k\epsilon} \Pr(A(x') \in E).$$

证明 考虑数据集 $x_0, x_1, \cdots, x_k$，其中 $x_0 = x$, $x_k = x'$，且 x_i 和 x_{i+1} 是 1-相邻数据集，$i = 0, \cdots, k-1$. 那么

$$\begin{aligned}\Pr(A(x) \in E) &\le e^{\epsilon} \Pr(A(x_1) \in E) \le e^{2\epsilon} \Pr(A(x_2) \in E) \\ &\le \cdots \le e^{k\epsilon} \Pr(A(x') \in E).\end{aligned}$$ □

换言之，k-相邻数据集上 ϵ-DP 算法的表现仿佛一个 $k\epsilon$-DP 算法。

这一性质还可以推出 ϵ 的含义。我们知道数据集 $x, x' \in \mathcal{X}^n$ 最多在 n 个位置不同. 所以对于一个 ϵ-DP 算法 A，一定有

$$\Pr(A(x) \in E) \le e^{n\epsilon} \Pr(A(x') \in E).$$

如果这里的 ϵ 太小，意味着这一算法对任何输入都有相似的输出. 换句话说，算法压根没有输出任何有意义的内容. 于是，我们更定量说明了，ϵ 还代表信息的泄露量. 因此，一个实用的 DP 算法不能让 ϵ 太小，否则输出没有意义；也不能让 ϵ 太大，否则隐私保护效果不好.

§5.3 差分隐私的应用

在这一部分，我们将会具体讨论三个差分隐私的算法。

5.3.1 随机反应算法

我们从一个具体场景开始。假设有一名老师想要调查班上的同学有多少人曾经在考试中作弊. 设班上一共有 n 名同学，每个人回答一个数字 $x_i \in \{0, 1\}$. 对于每个 i，独立地按照以下规则根据 x_i 得到对应的 y_i：

$$y_i = \begin{cases} x_i, & \text{以 } 2/3 \text{ 的概率}, \\ 1 - x_i, & \text{以 } 1/3 \text{ 的概率}. \end{cases}$$

并输出 $\sum_{i=1}^{n} y_i$.

我们称这一算法为*随机反应*（RR）算法. 当 $y_i = 1$ 时，学生 i 可以声称这是由于算法的随机机制造成的，而并非自己真的作弊过.

RR 算法在隐私保护上的表现由以下定理给出：

定理 5.1 *RR 算法是 $\log 2$-DP 算法.*

证明 记 Y_i 是 y_i 对应的随机变量. 我们知道 y_i 之间相互独立，所以

$$\Pr(A(x) = y) = \prod_{i=1}^{n} \Pr(Y_i = y_i \mid x_i).$$

对于 y_i，我们有

$$\frac{\Pr(Y_i = y_i \mid x_i = y_i)}{\Pr(Y_i = y_i \mid x_i = 1 - y_i)} = \frac{2/3}{1/3} = 2.$$

所以对于一对 1-相邻数据集 x, x' 和任意的 $y \in \mathcal{Y}$，假设 $x_j \neq x'_j$，有

$$\frac{1}{2} \leq \frac{\Pr(A(x) = y)}{\Pr(A(x') = y)} = \frac{\prod_{i=1}^{n} \Pr(Y_i = y_i \mid x_i)}{\prod_{i=1}^{n} \Pr(Y_i = y_i \mid x'_i)} = \frac{\Pr(Y_j = y_j \mid x_j)}{\Pr(Y_j = y_j \mid x'_j)} \leq 2.$$

由定义，RR 算法是 $\log 2$-DP 算法. □

另一方面，我们还关注 RR 算法得到的 $\sum_{i=1}^{n} y_i$ 是否能很好的估计出 $\sum_{i=1}^{n} x_i$. 为此，假设随机变量 $X_1, \cdots, X_n$ 独立服从参数为 p 的 Bernoulli 分布，即 $\Pr(X_i = 1) = p$ 而 $\Pr(X_i = 0) = 1 - p$. 于是，

$$\begin{aligned} q = \Pr(Y_i = 1) &= p \cdot \frac{2}{3} + (1 - p) \cdot \frac{1}{3} \\ \Longrightarrow \quad p &= 3q - 1. \end{aligned}$$

我们得到的 $\sum_{i=1}^{n} y_i$ 相当于是 q，真正的参数 p 和 q 存在上述关系. 设 $\hat{X} = \sum_{i=1}^{n} (3Y_i - 1)$，可得 $\mathbb{E}[\hat{X}] = \mathbb{E}[\sum_{i=1}^{n} X_i]$.

5.3.2 全局灵敏度与 Laplace 机制

从 RR 算法中获得灵感，我们可以在算法中添加随机性，比如向算法 f 的输出添加噪声. 那么就引出了另一个问题，需要添加多大的噪声？这和算法本身的性质有关，比如，当输入只改变一点时，算法的输出会改变多大？我们定义全局灵敏度来衡量这一性质.

定义 5.3 (全局灵敏度) *给定算法 $f : \mathcal{X}^n \to \mathbb{R}$，定义 f 的*全局灵敏度*为*

$$\mathrm{GS}_f = \sup_{x, x' \text{ 在 } \mathcal{X}^n \text{ 1-相邻}} |f(x) - f(x')|.$$

定义中的 1-相邻，可能会随着情景不同而改变. 全局灵敏度的定义是很直观的，就是改变一条数据会对算法输出带来的最大可能变化。

我们来计算一个简单的例子.

例 5.1 设 $f(x) = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$，$\phi : \mathcal{X} \to [0,1]$ 是满射. 那么，

$$\begin{aligned}\mathsf{GS}_f &= \sup_{x,x' \text{ 在 } \mathcal{X}^n \text{ 1-相邻}} |f(x) - f(x')| \\ &= \sup_{x,x' \text{ 在 } \mathcal{X}^n \text{ 1-相邻}} \frac{1}{n}\left|\sum_{i=1}^{n}\phi(x_i) - \sum_{i=1}^{n}\phi(x_i')\right| \\ &= \sup_{x,x' \text{ 在 } \mathcal{X}^n \text{ 只在 } j \text{ 不同}} \frac{1}{n}\left|\phi(x_j) - \phi(x_j')\right| = \frac{1}{n}.\end{aligned}$$

利用全局灵敏度的概念，我们实际上有一个一般的方法来构造差分隐私算法. 我们称这一方法为 *Laplace* 机制. 首先引入 Laplace 分布的概念.

定义 5.4 (Laplace 分布) 给定参数 $\mu \in \mathbb{R}$ 和 $\lambda > 0$，定义概率密度函数

$$h(x \mid \mu, \lambda) = \frac{1}{2\lambda}\exp\left(-\frac{|x-\mu|}{\lambda}\right).$$

我们称具有这一密度的分布为 Laplace 分布，记为 $\mathsf{Lap}(\mu, \lambda)$.

Laplace 分布是服从双边指数分布的随机变量进行线性变换后服从的分布. 具体来说，设 $X \sim \mathsf{DExp}(1)$，那么 $\lambda X + \mu \sim \mathsf{Lap}(\mu, \lambda)$. 在图 5.1 中，我们展示了不同参数的 Laplace 分布密度函数图像。更多关于 Laplace 分布的性质，我们留做习题. [lhy: 习题]

下面我们来陈述 Laplace 机制。我们想利用 Laplace 分布来创造一些随机性. 给定一个数据集 $x \in \mathcal{X}^n$ 和参数 ϵ. 对于一个算法 $f : \mathcal{X}^n \to \mathbb{R}$，先计算出 f 的全局灵敏度 GS_f。输出 $A_{\mathsf{Lap}}(\epsilon, x) = f(x) + Z$，其中 $Z \sim \mathsf{Lap}(0, \mathsf{GS}_f/\epsilon)$. 将随机算法 $A_{\mathsf{Lap}}(\epsilon, \cdot)$ 称为 *Laplace* 机制。我们有以下定理:

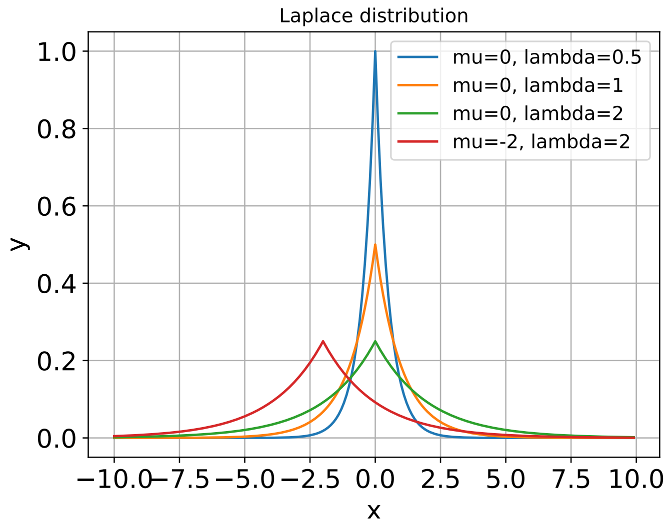
定理 5.2 对于任意的 $\epsilon > 0$，*Laplace* 机制是 *ϵ-DP* 算法.

证明 设 x，x' 是两个 1-相邻数据集，记 $\mu = f(x)$，$\mu' = f(x')$. 由 Laplace 分布的性质可知，$A_{\mathsf{Lap}}(\epsilon, x) \sim \mathsf{Lap}(\mu, GS_f/\epsilon)$，$A_{\mathsf{Lap}}(\epsilon, x') \sim \mathsf{Lap}(\mu', GS_f/\epsilon)$.

因此，对于任意的 $y \in \mathcal{Y}$，有

$$\begin{aligned}\frac{h_x(y)}{h_{x'}(y)} &= \exp\left(-\epsilon\frac{|\mu - y| - |\mu' - y|}{GS_f}\right) \\ &\le \exp\left(\epsilon\frac{|\mu - \mu'|}{GS_f}\right) \le \exp(\epsilon).\end{aligned}$$

根据命题 5.2，命题得证。 □



[lhy: 重画]

图 5.1: Laplace 分布的密度函数图像

5.3.3 DP 版本 Llyod 算法

作为一个 Laplace 机制的具体实例，我们将 k-均值聚类问题的经典算法 Llyod 算法改造成一个差分隐私算法.

k-均值聚类问题指的是给定一个数据集 x，找到 k 个点（中心）$\{c_i\} \subseteq \mathbb{R}^d$，使得 $\sum_{i\in[n]} \min_{j\in[k]} \|x_i - c_j\|^2$ 最小. 通俗来说，就是找到 k 个中心，使得数据集中每个点到最近的中心的距离之和最小. k-均值问题最常见的解决方法是使用迭代的启发式的 *Lloyd* 算法，其表述为下：[lhy: 改成算法模板]

- 输入：数据集 $x \in \mathcal{X}^n$，这里 $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_1 \le 1\}$，参数 k.
- 随机初始化 $c_1^{(0)}, c_2^{(0)}, \cdots, c_k^{(0)} \in \mathcal{X}$.
- for $t = 1$ to T
  - for $j = 1$ to k
    - 计算 $S_j = \{i : c_j^{(t-1)}$ 是 x_i 最近的中心$\}$.
    - 更新 $c_j^{(t)} = \frac{1}{|S_j|}\sum_{i\in S_j} x_i$.
- 输出：$c_1^{(T)}, c_2^{(T)}, \cdots, c_k^{(T)}$.

这个算法可以达到很好的效果，但它并不能保证 DP 性质. 我们希望对这一算法进行小规模的修改，让它具有 ϵ-DP 的性质. 我们给出如下的 DP 版本 Llyod 算法. [lhy: 改成算法模板]

- 输入：数据集 $x \in \mathcal{X}^n$，这里 $\mathcal{X} = \{x \in \mathbb{R}^d : ||x||_1 \leq 1\}$，参数 k，参数 ϵ.
- $\epsilon' = \frac{\epsilon}{2T}$，随机初始化 $c_1^{(0)}, c_2^{(0)}, \cdots, c_k^{(0)} \in \mathcal{X}$.
- for $t = 1$ to T
  - for $j = 1$ to k
    - 计算 $S_j = \{i : c_j^{(t-1)}$ 是x_i 最近的中心$\}$.
    - $n_j = |S_j|$.
    - $a_j = \sum_{i \in S_j} x_i$.
    - 计算 $\hat{n}_j = n_j + Y$，$Y \sim \mathsf{Lap}(0, 2/\epsilon')$.
    - 计算 $\hat{a}_j = a_j + (Z_1, \cdots, Z_d)$，$Z_i$ i.i.d. $\sim \mathsf{Lap}(0, 2/\epsilon')$.
    - 更新 $c_j^{(t)} = \begin{cases} \frac{\hat{a}_j}{\hat{n}_j}, & \hat{n}_j \geq 1, \\ \mathcal{X} \text{ 上的一个随机均匀采样}, & \hat{n}_j < 1. \end{cases}$
- 输出：$c_1^{(T)}, c_2^{(T)}, \cdots, c_k^{(T)}$.

以下定理表明上面的算法确实是一个 ϵ-DP 算法。

定理 5.3 *DP 版本的 Lloyd 算法是 ϵ-DP 算法.*

证明 (证明概要) 我们只在这里陈述证明的大致想法，将细节留到习题[lhy: 习题].

在第一步中，我们设置了 $\epsilon' = \epsilon/2T$. 我们把整体的算法拆成 T 个阶段，其中第 t 轮迭代 A_t 以 $c_1^{(t-1)}, c_2^{(t-1)}, \ldots, c_k^{(t-1)}$ 作为输入，输出 $c_1^{(t)}, c_2^{(t)}, \ldots, c_k^{(t)}$. 如果可以得到 A_t 是 $2\epsilon'$-DP 算法，那么由 DP 算法的复合性，就可以得到整个算法是 $T \cdot 2\epsilon'$-DP，也就是 ϵ-DP.

进一步，考虑证明每一个 A_t 是 $2\epsilon'$-DP 算法. 将 A_t 内循环的每一轮（以 j 为变量）视为输入为 n_j, a_j，输出为 $\hat{n}_j$，$\hat{a}_j$ 的算法. 分别证明这些算法符合 $2\epsilon'$-DP，然后再次借助 DP 算法的复合性. □

§5.4 差分隐私与信息论

[lhy: 讨论信息论约束下的差分隐私问题]

§5.5 习题

§5.6 章末注记

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 $a^\mathsf{T} x + b$）.

> **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 $(Ax - b)^\mathsf{T}(Ax - b)$，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：

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

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：
$$x_{(i_1,\ldots,i_n)} = \left(\frac{i_1}{p}, \ldots, \frac{i_n}{p}\right)^\top,$$
$i_k \in \{0, 1, \ldots, p\}$.
- 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 $B' = \{x \in [0,1]^n : \|x - x^*\|_\infty \leq \epsilon/L\}$ 内部. 因为 $1/(2p) \geq \epsilon/L$，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：

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

问题的下界：

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

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).

$$\begin{aligned}\Longleftarrow: \ f(y) &\geq (1-\alpha)^{-1}(f(\alpha x+(1-\alpha)y)-\alpha f(x)) \\ &= f(x)+(1-\alpha)^{-1}(f(x+(1-\alpha)(y-x))-f(x)).\end{aligned}$$

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 $\{x \in \mathbb{R}^n : Ax \leq b\}$，$A \in \mathbb{R}^{m\times n}$，$b \in \mathbb{R}^m$.
- 单纯形：$\Delta_n = \{x \in \mathbb{R}^n : x_1 + \cdots + x_n = 1, x_i \geq 0, \forall i\}$，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 $x \in C$，$a^\mathsf{T} x \leq b$.*

- 任意 $x \in D$，$a^\top x \geq b$.

由 $a^\top x = b$ 定义的超平面被称为**分离超平面**.

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 $f(x) = a^\top x - b$ 在 C 上非正在 D 上非负. 对称地，只证明在 D 上非负.

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

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

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！ □

第七章　对偶理论

在本章中，我们考虑带约束的规划问题。它的一般形式是

$$
\begin{aligned}
\min \quad & f(x) \\
\text{s.t.} \quad & h_i(x) = 0, \qquad i = 1, \ldots, m, \\
& g_j(x) \le 0, \qquad j = 1, \ldots, p, \\
& x \in \Omega \subseteq \mathbb{R}^n.
\end{aligned}
$$

其中，$m \le n$，函数 f, h_i, g_j 都是连续的，且通常假设它们拥有连续的二阶导.

为简化记号，们用向量形式的函数，即 $h = (h_1, h_2, \ldots, h_m)$ 和 $g = (g_1, g_2, \ldots, g_p)$，把问题的形式重写为:

$$
\begin{aligned}
\min \quad & f(x) \\
\text{s.t.} \quad & h(x) = 0, \\
& g(x) \le 0, \\
& x \in \Omega.
\end{aligned}
$$

约束 $h(x) = 0, g(x) \le 0$ 被称作**函数约束**. $x \in \Omega$ 是**集合约束**. 我们并不强调集合约束，因此假设在大部分情况下 Ω 就是整个 $\mathbb{R}^n$ 的空间，或者问题的解就在 Ω 的内部.

一个满足所有函数约束的点 $x \in \Omega$ 被称作**可行解**，而使得 f 取得最小值的可行解叫做**最优解**. 有时候优化问题的目标可能是最大化 f，此时相应的最优解就是使得 f 取得最大值的可行解。本章的任务是讨论各种情况下最优值的必要条件，这些必要条件最终形成了所谓的**对偶理论**。

§7.1 条件极值与 Lagrange 乘子法

我们现在先只考虑等式约束

$$
\begin{aligned}
\min \quad & f(x) \\
\text{s.t.} \quad & h(x) = 0, \\
& x \in \Omega.
\end{aligned}
\tag{7.1}
$$

这些约束定义了一个 $\mathbb{R}^n$ 的子集，可以被看作一个曲面. 在恰当的条件下，这个曲面是 $n-m$ 维的（类比线性空间）. 如果函数 $h_i, i = 1, 2, \ldots, m$ 有一阶连续导数（记为属于 C^1），那么他们定义的曲面就是*光滑*的. 曲面上可以定义*切空间*.

为了引入切空间，我们先介绍曲线，然后曲线的定义可以导出切空间的定义.

定义 7.1 (曲线与切空间)

- *超平面 S 上的一条* **曲线** *是一系列点的集合：$x(t) \in S$，它们以 t 为参数，$a \le t \le b$ 且在该区间上连续.*
- *称曲线是* **可微的**，*如果 $\dot{x} = \mathrm{d}(x(t))/\mathrm{d}t$ 存在.*
- *称曲线 $x(t)$* **经过点** *x^*，如果存在 $t^* \in [a, b]$ 使得 $x^* = x(t^*)$.*
- *曲线在 x^* 的* **导数** *被定义为 $\dot{x}(t^*)$，该导数是 $\mathbb{R}^n$ 内的一个向量，这个向量可以看作沿着曲线 t 在 x^* 处的切向量.*
- *考虑所有 S 内经过点 x^* 的可微曲线. 点 x^* 处的* **切空间** *$T_{x^*}(S)$ 被定义为这些曲线在点 x^* 处的导数的集合.*

切空间的重要特点是，它是一个线性空间.

引理 7.1 *切空间是一个线性空间。*

既然切空间是一个线性空间，我们的一个主要目标就是给出切空间的显示表达，比如给出它的一组基向量。考虑一条曲线 $x(t)$，如果它在 $h_i(x) = 0$ 形成的曲面上，那么应该有

$$
\frac{\mathrm{d}}{\mathrm{d}t} h_i(x(t)) = 0 \iff \nabla_x h_i(x(t)) \dot{x}(t) = 0.
$$

因此 $x(t)$ 的切向量和该点处函数 $h_i(x(t))$ 的导数正交. 于是，如果 $x(t)$ 在 $h(x) = 0$ 形成的曲面上，那么 $x(t)$ 处的导数 $\nabla h(x(t))$ 是切平面的法向量. 这一数学推导的示意图见图 7.1.

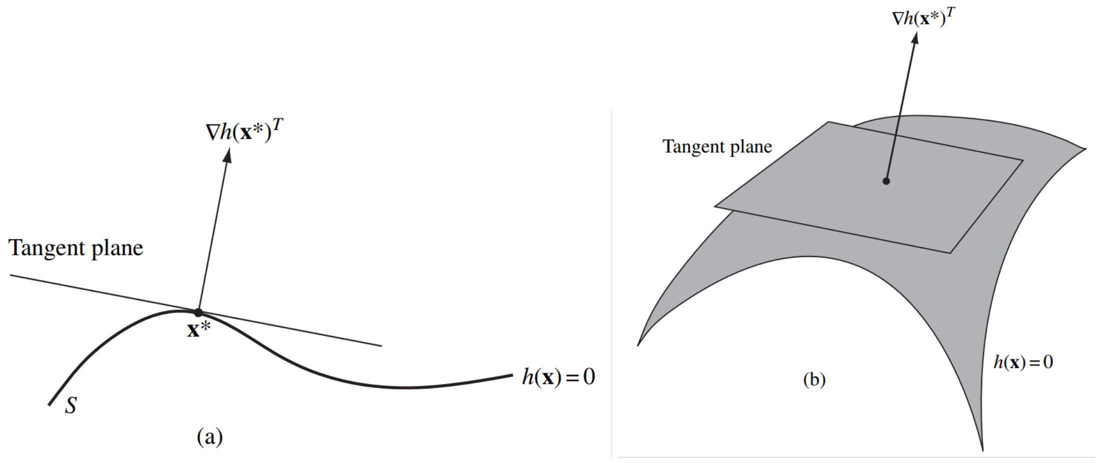


图 7.1: 切空间的示意图.

我们把刚刚得到的垂直于 $\nabla h(x^*)y$ 的子空间（即正交补空间）记作

$$M = \{y \in \mathbb{R}^n : \nabla h(x^*)y = 0\}.$$

我们已经证明 $T_{x^*}(S) \subseteq M$。反过来，在什么条件下会有 $M = T_{x^*}(S)$？为此，我们引入正规点的概念。

定义 7.2 (正规点) 考虑优化问题 (7.1)，当一个点 $x^* \in \Omega$ 满足约束 $h(x^*) = 0$，且梯度向量 $\nabla h_1(x^*), \nabla h_2(x^*), \ldots, \nabla h_m(x^*)$ 线性无关时，它被称作该约束的**正规点**.

直观上来说，正规点上每一条约束都起到了实际的作用，因此梯度向量 $\nabla h_i(x^*)$ 形成了一个线性无关的集合，张成了空间 $M^\perp$. 此时，切空间恰好完全垂直于 $M^\perp$，即 $T_{x^*}(S) = M$. 这一几何直观见图 7.2，点 x^* 处的两个等式约束共同确定了该点的切空间. 因此，在正规点，用约束函数的梯度来描述切空间是可行的.

定理 7.1 (正规点切空间刻画定理) 设曲面 $S \subseteq \mathbb{R}^n$ 由约束 $h(x) = 0$ 定义，$x^* \in S$ 是正规点，那么，

$$T_{x^*}(S) = M = \{y : \nabla h(x^*)y = 0\}.$$

该定理的证明需要隐函数定理，对微积分要求较高，我们这里略去.

有了切空间的准备，现在我们要对正规点推导带约束的优化问题的极值条件. 考虑优化(7.1)，设 x^* 是一个约束 $h(x) = 0$ 一个正规点，同时也是函数 f 的一个在可行域中的极值点.

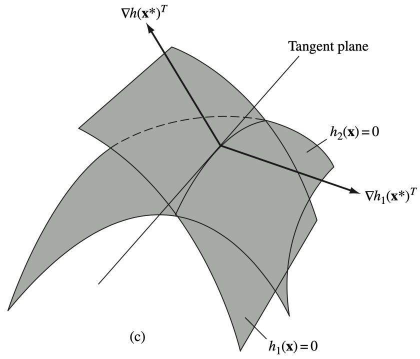


图 7.2: 正规点示意图。

引理 7.2 对 $y \in \mathbb{R}^n$，如果 $\nabla h(x^*)y = 0$，那么 $\nabla f(x^*)y = 0$.

证明 因为 x^* 是正规点，根据正规点切空间刻画定理，$\nabla h(x^*)y = 0$ 等价于 y 是 x^* 处的切空间中的向量. 根据定义，存在该约束曲面内的某个光滑曲线 $x(t)$，经过点 x^*，并且以 y 为切向量. 那么，$x(0) = x^*$，$\dot{x}(0) = y$，且 $h(x(t)) = 0$ 在区间 $-a \le t \le a$ 上成立（对某个正数 a）. 因为点 x^* 是一个函数 f 的受等式约束的极值点，我们有

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t))\bigg|_{t=0} = 0 \iff \nabla f(x^*)y = 0.$$

□

引理 7.2 对任意 $y \in \mathbb{R}^n$ 都成立，根据线性代数零空间的性质，这等价于 $\nabla f(x^*)$ 是 $\nabla h_i(x^*)$ 的线性组合，即

$$\nabla f(x^*) = \sum_i \lambda_i \nabla h_i(x^*).$$

据此，我们得到条件极值的一阶必要条件：

定理 7.2 (条件极值的一阶必要条件) 令 x^* 是一个 f 的满足约束 $h(x) = 0$ 的正规极值点. 那么存在一个 $\lambda \in \mathbb{R}^m$ 使得

$$\nabla f(x^*) + \lambda^\mathsf{T} \nabla h(x^*) = 0.$$

一阶必要条件 $\nabla f(x^*) + \lambda^\mathsf{T} \nabla h(x^*) = 0$ 以及约束 $h(x^*) = 0$ 给出了 $n + m$ 个等式，包含 x^*, λ 在内的 $n + m$ 个变量. 因此在非退化的情况下，他们给出了一个唯一解.

引入与这个约束问题对应的 Lagrange 函数：

$$l(x, \lambda) = f(x) + \lambda^\mathsf{T} h(x).$$

λ 被称为 *Lagrange* 乘子. 必要条件可以被写作:

$$\nabla_x l(x, \lambda) = 0,$$
$$\nabla_\lambda l(x, \lambda) = 0.$$

例 7.1 (最大熵) 考虑一个离散的概率分布，其分布列为 $p_i = \Pr(X = x_i), i = 1, \ldots, n$. 该分布的熵为

$$\epsilon = -\sum_{i=1}^{n} p_i \log p_i.$$

该分布的均值为 $\sum_{i=1}^{n} x_i p_i$.

如果均值固定为 m，求解使熵最大化的参数可以被转化成以下问题:

$$\begin{aligned}
\max \quad & -\sum_{i=1}^{n} p_i \log p_i \\
s.t. \quad & \sum_{i=1}^{n} p_i = 1, \\
& \sum_{i=1}^{n} x_i p_i = m, \\
& p_i \geq 0, \qquad i = 1, 2, \ldots, n.
\end{aligned}$$

我们先忽略非负约束，假设这些约束不会被触发. 引入两个 *Lagrange* 乘子，λ 和 μ，则 *Lagrange* 函数为

$$l = \sum_{i=1}^{n} (-p_i \log p_i + \lambda p_i + \mu x_i p_i) - \lambda - \mu m.$$

由一阶必要条件，$-\log p_i - 1 + \lambda + \mu x_i = 0$，$i = 1, 2, \ldots, n$. 因此，

$$p_i = \exp((\lambda - 1) + \mu x_i), \quad i = 1, 2, \ldots, n.$$

注意 $p_i > 0$，所以非负约束确实没有被触发. *Lagrange* 乘子 λ 和 μ 是两个用来保证等式约束被满足的参数.

§7.2 Karush–Kuhn–Tucker 条件

现在加入不等式约束，考虑以下形式的问题:

$$\begin{aligned}
\min \quad & f(x) \\
\text{s.t.} \quad & h(x) = 0, \\
& g(x) \leq 0.
\end{aligned} \tag{7.2}$$

假设 f 和 h 和前面一样，g 是一个 p 维的函数，$f,h,g \in C^1$.

我们推广正规点 x^* 的定义为：

定义 7.3 (正规点) 考虑优化问题 (7.2)，点 x^* 被称为**正规点**，如果

- 它满足约束：$h(x^*) = 0, g(x^*) \le 0$.
- 令 J 为满足 $g_j(x^*) = 0$ 的下标 j 的集合（激活的约束）. 那么，梯度向量 $\nabla h_i(x^*)$，$\nabla g_j(x^*)$，$1 \le i \le m$，$j \in J$ 是线性无关的.

换言之，此时的正规点不仅考虑等式约束，还要考虑起作用的或者说被激活的不等式约束，这些不等式约束相当于等式约束。类似 Lagrange 乘子法，此时的一阶必要条件为：

定理 7.3 (Karush-Kuhn-Tucker 条件) 令 x^* 为优化问题 (7.2) 的正规极小值点，那么，存在向量 $\lambda \in \mathbb{R}^m$ 和向量 $\mu \in \mathbb{R}^p$ 且 $\mu \ge 0$ 使得

$$\nabla f(x^*) + \lambda^\mathsf{T} \nabla h(x^*) + \mu^\mathsf{T} \nabla g(x^*) = 0, \tag{7.3}$$

$$\mu^\mathsf{T} g(x^*) = 0. \tag{7.4}$$

证明 首先，因为 $\mu \ge 0$ 且 $g(x^*) \le 0$，(7.4) 等价于：μ 的一个分量非零仅当对应的约束被激活（即取到等号）. 这是一个互补松弛条件，即 $g(x^*)_i < 0$ 可得出 $\mu_i = 0$，以及 $\mu_i > 0$ 可得出 $g(x^*)_i = 0$.

设被激活的下标为 J. 因为 x^* 是约束集合上的一个极小点，它也是满足等式约束 $h(x) = 0, g_i(x) = 0, i \in J$ 的极小点. 因此，在新的等式约束问题中，x^* 的邻域中存在 Lagrange 乘子，满足一阶必要条件. 我们得出结论：一阶必要条件 (7.3) 成立，且若 $g_j(x^*) \neq 0$，则 $\mu_j = 0$.（于是也有 (7.4) 成立）

现在还需要证明 $\mu \ge 0$. 用反证法，假设 $\mu_k < 0$ 对某个 $k \in J$ 成立. 设 S 为其他所有被激活的约束在 x^* 处定义的曲面，$M = T_{x^*}(S)$. 因为 x^* 是正规的，存在 $y \in M$ 且 $\nabla g_k(x^*) y < 0$. 令 $x(t)$ 为一条在 S 内且经过 x^*（此处 $t = 0$）的曲线，且有 $\dot{x}(0) = y$. 则对于充分小的 $t \ge 0$，$x(t)$ 是可行的，由 (7.3) 以及 $y \in M$，

$$\begin{aligned} \left.\frac{\mathrm{d} f(x(t))}{\mathrm{d} t}\right|_{t=0} &= \nabla f(x^*) y \\ &= -\lambda^\mathsf{T} \nabla h(x^*) y - \mu^\mathsf{T} \nabla g(x^*) y \\ &= -\mu_k \nabla g_k(x^*) y < 0. \end{aligned}$$

这与 x^* 是极小点矛盾. □

注. 这一证明具有很强的几何直观，关键在于找一个可行的方向使得函数值下降. 非常需要注意的是，这一证明并不能用于否证 $\mu_k > 0$. 此时需要取 $y \in M$ 使得 $\nabla g_k(x^*)y > 0$. 然而此时对应的 $x(t)$ 不再可行，因为对充分小的 $t > 0$，$g_k(x(t)) > 0$，违背了约束的条件.

下面我们来看一个运用 KKT 条件的例子：

例 7.2 考虑问题

$$\begin{aligned} \min \quad & 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \\ s.t. \quad & x_1^2 + x_2^2 \le 5, \\ & 3x_1 + x_2 \le 6. \end{aligned}$$

KKT 条件为（注意，一阶必要条件还需要加入问题中的约束条件）

$$\begin{aligned} 4x_1 + 2x_2 - 10 + 2\mu_1 x_1 + 3\mu_2 &= 0, \\ 2x_1 + 2x_2 - 10 + 2\mu_1 x_2 + \mu_2 &= 0, \\ \mu_1(x_1^2 + x_2^2 - 5) &= 0, \\ \mu_2(3x_1 + x_2 - 6) &= 0, \\ \mu_i &\ge 0, \quad i = 1, 2. \end{aligned}$$

为了求解此类问题，我们假设一些约束被激活，然后检查所得出的 *Lagrange* 乘子的符号正负. 在这个问题中，我们可以尝试假设有 *0*，*1*，*2* 个约束被激活.

假设第一个约束被激活，第二个约束没有被激活，得出等式

$$\begin{aligned} 4x_1 + 2x_2 - 10 + 2\mu_1 x_1 &= 0, \\ 2x_1 + 2x_2 - 10 + 2\mu_1 x_2 &= 0, \\ x_1^2 + x_2^2 &= 5. \end{aligned}$$

可得解 $x_1 = 1, x_2 = 2, \mu_1 = 1$.

由于 $3x_1 + x_2 = 5$，因此第二个约束也被满足了. 因此，因为 $\mu_1 > 0$，我们得出结论，这个解满足一阶必要条件.

§7.3 Lagrange 对偶

7.3.1 Lagrange 定理

现在，我们不再假设函数可微，我们考虑极值点的零阶必要条件，首先考虑只有等式约束的情形：

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & h(x) = 0, \\ & x \in \Omega. \end{aligned} \tag{7.5}$$

如果函数 f 是凸函数，m 维函数 h 是仿射的，并且集合 $\Omega \subset \mathbb{R}^n$ 是是凸的，那么这个规划问题是一个凸规划问题.

为了给这样的问题一个一阶必要条件，我们依然需要引入正规性条件. 此时正规性不再仅仅只对一个点，而是对仿射函数 h.

定义 7.4 (正规性条件) 一个仿射函数 h 关于集合 Ω 是**正规的**，指的是像集 $h(\Omega) = \{y : \exists x \in \Omega\, h(x) = y\}$ 包含 0 处的一个开球邻域. 也就是说，$h(\Omega)$ 包含一个形如 $\{y : \|y\| < \epsilon\}$（对某个 $\epsilon > 0$）的集合.

注. 这个条件是一阶正规点定义的推广. 如果 h 在点 x^* 有连续的导数，那么一阶正规性条件意味着 $\nabla h(x^*)$ 是满秩的，并且由隐函数定理可知存在一个 $\epsilon > 0$ 使得对于任意满足 $\|y - h(x^*)\| < \epsilon$ 的 y，都有一个 x 使得 $h(x) = y$. 换言之，存在一个 $y^* = h(x^*)$ 周围的开球.

我们可以用 Lagrange 乘子来表述零阶必要条件：

定理 7.4 (零阶必要条件，等式约束情形) 假设 $\Omega \subset \mathbb{R}^n$ 是凸的，f 是 Ω 上的凸函数，h 是一个 Ω 上的 m 维仿射函数. 假设 h 是对于 Ω 正规的. 如果 x^* 是 (7.5) 的解，那么存在 $\lambda \in \mathbb{R}^m$ 使得 x^* 是以下 *Lagrange* 问题的解：

$$\begin{aligned} \min \quad & f(x) + \lambda^\mathsf{T} h(x) \\ s.t. \quad & x \in \Omega. \end{aligned}$$

这一定理证明的关键在于引入原始函数. 对应于问题(7.5) 的**原始函数**是：

$$\omega(y) = \inf\{f(x) : h(x) = y, x \in \Omega\}, \quad y \in h(\Omega).$$

证明 (零阶必要条件的证明) 令 $f^* = f(x^*)$. 定义 $\mathbb{R}^m \times \mathbb{R}$ 内的集合 A 和 B 为:

$$A = \{(y, r) : r \geq \omega(y), y \in h(\Omega)\},$$
$$B = \{(y, r) : r \leq f^*, y = 0\}.$$

A 是 ω 的上图，B 是 f^* 向下延申并与原点对齐的垂线. A 和 B 都是凸集. 他们唯一的公共点是 $(0, f^*)$. 由超平面分离定理可知，存在一个超平面分离 A 和 B. 这个超平面可以被表示成一个在 $\mathbb{R}^m \times \mathbb{R}$ 内的形如 $(\lambda, s), \lambda \in \mathbb{R}^m$ 的非零向量，还有一个分离常数 c. 分离条件是

$$sr + \lambda^\mathsf{T} y \geq c, \quad \forall (y, r) \in A, \quad sr + \lambda^\mathsf{T} y \leq c, \quad \forall (y, r) \in B.$$

这一过程的示意图见图 7.3.

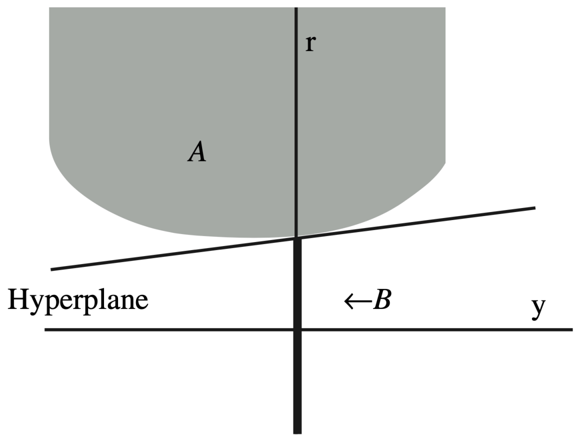


图 7.3: 证明示意图。

注意到 $s \geq 0$，否则取 $|r|$ 非常大的负数 r，点 $(r, 0) \in B$ 违反第二个分离不等式. 几何上看，若 $s = 0$，超平面将垂直. 我们来证明 $s \neq 0$. 假设 $s = 0$，因为 s 和 λ 不能都是 0，$\lambda \neq 0$. 因为分离超平面必须包含点 $(f^*, 0)$，从第二个分离不等式得 $c = 0$. 由 h 的正规性，以 $0 \in h(\Omega)$ 为中心的某个球包含在 $h(\Omega)$ 中，任取 y 属于这个开球. 第一个分离不等式左侧为 $\lambda^\mathsf{T} y$，它对于某些 y 来说是负的. 这违背第一个分离不等式. 因此 $s \neq 0$，继而 $s > 0$.

不失一般性，可以假设 $s = 1$. 假设 $x \in \Omega$. 那么 $(h(x), f(x)) \in A$ 且 $(0, f(x^*)) \in B$. 因此，由分离不等式可知，我们有

$$f(x) + \lambda^\mathsf{T} h(x) \geq f(x^*) = f(x^*) + \lambda^\mathsf{T} h(x^*).$$

因此 x^* 是优化问题 (7.5) 解. □

我们再考虑只有不等式约束的模型

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \le 0, \\ & x \in \Omega. \end{aligned} \tag{7.6}$$

其中，g 是一个 p 维的函数.

然后我们引入正规性条件。对于不等式约束来说，正规性条件也被称为做 *Slater* 条件.

定义 7.5 (Slater 条件) *考虑优化问题 (7.6)，令*

$$D = \{z \in \mathbb{R}^p : \exists x \in \Omega \, g(x) \le z\}.$$

正规性条件（Slater 条件）为：存在一个 $z' \in D$ 使得 $z' < 0$.

直观来说，Slater 条件指的是存在满足约束的内点.

类似地，我们可以用 Lagrange 乘子来表述零阶必要条件：

定理 7.5 (零阶必要条件，不等式情形) *假设 Ω 是一个 $\mathbb{R}^n$ 的凸子集，且 f 和 g 是凸函数. 假设优化问题 (7.6) 满足正规性条件，x^* 是该问题的解，那么存在一个向量 $\mu \in \mathbb{R}^p$ 满足 $\mu \ge 0$ 使得 x^* 是下述 Lagrange 问题的解：*

$$\begin{aligned} \min \quad & f(x^*) + \mu^\top g(x) \\ s.t. \quad & x \in \Omega. \end{aligned}$$

此外，$\mu^\top g(x^*) = 0$.

这一定理的证明类似于定理 7.4 的证明。首先还是引入原始函数。问题 (7.6) 对应的原始函数为：

$$\omega(z) = \inf\{f(x) : g(x) \le z, x \in \Omega\}, z \in D.$$

证明 (证明概要) 令 $f^* = f(x^*)$. 在 $\mathbb{R}^p \times \mathbb{R}$ 内定义两个集合

$$\begin{aligned} A &= \{(z, r) : r \ge \omega(z), z \in D\}, \\ B &= \{(z, r) : r \le f^*, z \le 0\}. \end{aligned}$$

A 和 B 都是凸的. 证明依然是构造 A, B 的分离超平面，正规性条件保证了超平面不会是垂直的. 这个过程的示意见图图 7.4.

条件 $\mu^\top g(x^*) = 0$ 是互补松弛条件，这一讨论类似 KKT 条件. □

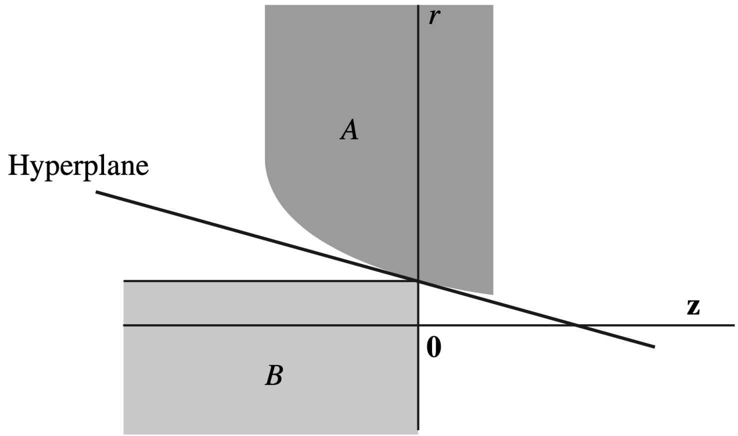


图 7.4: 证明示意图。

现在，我们考虑一般情形，

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & h(x) = 0, \\ & g(x) \leq 0, \\ & x \in \Omega. \end{aligned} \tag{7.7}$$

组合以上两个零阶必要条件，我们得到一般情形的 Lagrange 定理.

定理 7.6 (Lagrange，零阶必要条件，混合情形) 假设 $\Omega \subset \mathbb{R}^n$ 是凸集. f 和 g 是一维和 p 维的凸函数，h 是维数为 m 的仿射函数. 假设 h 满足对于 Ω 的正规性条件，且 g 在 (7.7) 的可行域上满足正规性条件. 假设 x^* 是问题 (7.7) 的解. 那么存在向量 $\lambda \in \mathbb{R}^m$ 和 $\mu \in \mathbb{R}^p$ 满足 $\mu \geq 0$ 使得 x^* 是以下 *Lagrange* 问题的解:

$$\begin{aligned} \min \quad & f(x) + \lambda^\top h(x) + \mu^\top g(x) \\ s.t. \quad & x \in \Omega. \end{aligned}$$

此外，$\mu^\top g(x^*) = 0$.

[lhy: 举个例子]

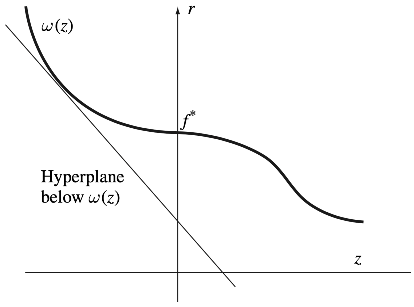


图 7.5: 纵截距的示意图.

7.3.2 弱对偶定理，强对偶定理

Lagrange 定理有非常强的几何直观，这一直观最终导致了优化中的对偶理论. 先考虑不等式约束的情形：

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \le 0, \\ & x \in \Omega. \end{aligned} \tag{7.8}$$

$\Omega \subset \mathbb{R}^n$ 是凸集，函数 f 和 g 定义在 Ω 上. 函数 g 是 p 维的.

回忆原始函数的定义：

$$\omega(z) = \inf\{f(x) : g(x) \le z, x \in \Omega\}.$$

设 x^* 是 (7.8) 的解，$f^* = f(x^*)$，那么函数 $\omega(z)$ 与纵轴的交点是 f^*. 如果 (7.8) 没有解，那么 $f^* = \inf\{f(x) : g(x) \le 0, x \in \Omega\}$ 就是纵轴与 $\omega(z)$ 的交点. 考虑在 $\omega(z)$ 以下的超平面，关注其纵截距（见图图 7.5），我们用它产生对偶原理.

为了刻画超平面以及其纵截距，我们引入对偶函数. 在 $\mathbb{R}^p_{\ge 0}$ 上定义**对偶函数**为：

$$\varphi(\mu) = \inf\{f(x) + \mu^\mathsf{T} g(x) : x \in \Omega\}.$$

定义其最大值为

$$\varphi^* = \sup\{\varphi(\mu), \mu \ge 0\}.$$

我们很容易可以证明以下定理：

定理 7.7 (弱对偶定理) $\varphi^* \le f^*$.

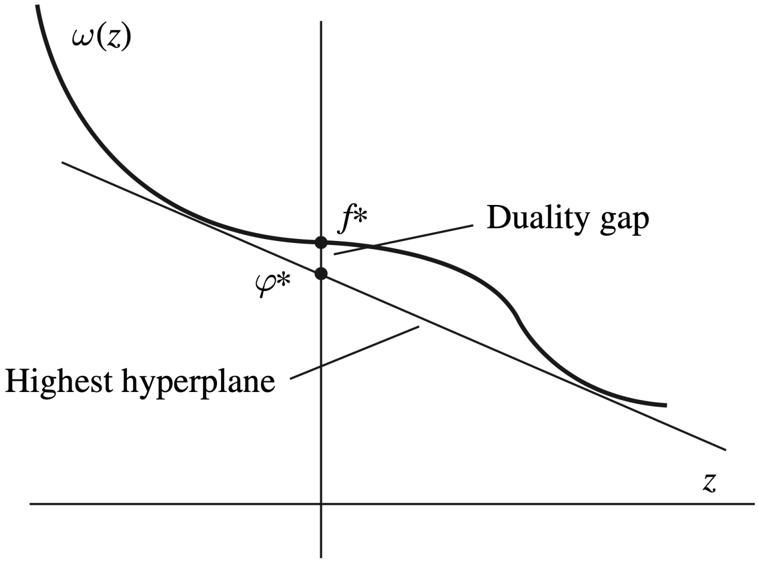


图 7.6: 对偶间距的示意图.

证明 对任意 $\mu \geq 0$ 我们有

$$\begin{aligned}\varphi(\mu) &= \inf\{f(x) + \mu^\mathsf{T} g(x) : x \in \Omega\} \\ &\leq \inf\{f(x) + \mu^\mathsf{T} g(x) : g(x) \leq 0, x \in \Omega\} \\ &\leq \inf\{f(x) : g(x) \leq 0, x \in \Omega\} = f^*.\end{aligned}$$

由此，$\varphi^* \leq f^*$. □

弱对偶定理也有非常几何的解释。考虑向量 $(\mu, 1) \in \mathbb{R}^p \times \mathbb{R}$，$\mu \geq 0$ 和一个常数 c. 关于 (z, r) 的方程 $(\mu, 1)^\mathsf{T}(z, r) = r + \mu^\mathsf{T} z = c$ 定义了一个 $\mathbb{R}^p \times \mathbb{R}$ 内的超平面. 不同的 c 得到不同的超平面，他们都是平行的. 对于给定的 $(\mu, 1)$（即平行的超平面），选取一个最低的超平面，使得它刚刚碰到了原始函数上图边界. 假设 x_1 是这个触点，有 $r = f(x_1)$ 和 $z = g(x_1)$. 那么 $c = f(x_1) + \mu^\mathsf{T} g(x_1) = \varphi(\mu)$. 注意到此时 $c = \varphi(\mu)$ 就是截距，这就是 $\varphi(\mu)$ 的几何含义.

另一方面，求截距 c（对偶函数值）的最大值 φ^*，就是求位于原始函数之下的超平面的最大截距. 因此至少有 $\varphi^* \leq f^*$，差 $f^* - \varphi^*$ 被称为对偶间距. 这就是弱对偶定理，图示参见图 7.6.

由此可以得到对偶性原理：位于 ω 之下的超平面的最大截距等于刚刚碰到 ω 的超平面的最小截距.

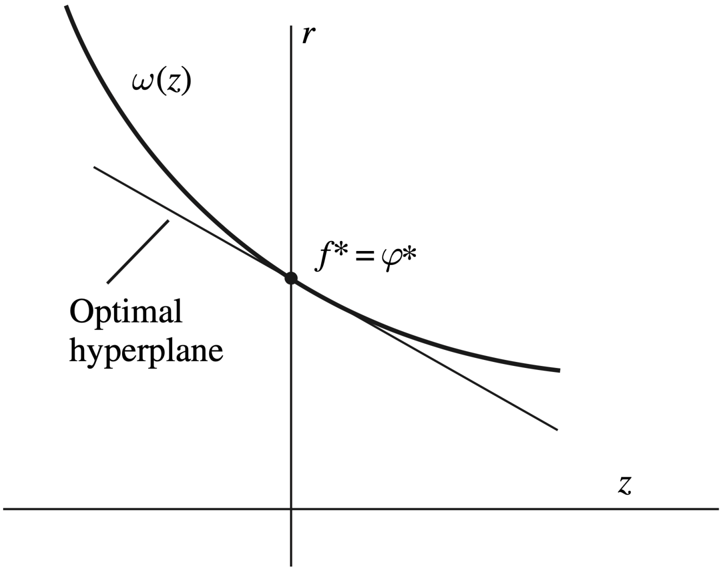


图 7.7: 强对偶定理的示意图.

如果原始函数 ω 是凸的，那么弱对偶定理可以被加强到强对偶定理，此时 φ^* 和 f^* 之间不再存在对偶间距，图 7.6 变成了图 7.7.

下面我们叙述并证明强对偶定理。我们直接考虑一般的优化问题.

$$
\begin{aligned}
\min \quad & f(x) \\
\text{s.t.} \quad & h(x) = 0, \\
& g(x) \leq 0, \\
& x \in \Omega.
\end{aligned}
\tag{7.9}
$$

其中，h 是 m 维仿射函数，g 是 p 维凸函数，$\Omega \subseteq \mathbb{R}^n$ 是凸集.

原始函数可以写作

$$
\omega(y, z) = \inf\{f(x) : \exists x \in \Omega\, h(x) = y, g(x) \leq z\}.
$$

对偶函数定义为：

$$
\varphi(\lambda, \mu) = \inf\{f(x) + \lambda^\mathsf{T} h(x) + \mu^\mathsf{T} g(x) : x \in \Omega\}.
$$

它的最大值记为

$$
\varphi^* = \sup\{\varphi(\lambda, \mu) : \lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p, \mu \geq 0\}.
$$

利用以上定义，我们可以表述强对偶定理如下：

	原料 1	原料 2	原料 3	售价（万元/吨）
清洁剂 A	0.25	0.50	0.25	12
清洁剂 B	0.50	0.50		15
存量（吨）	120	150	50	

表 7.1: 清洁剂原料价格存量表。

定理 7.8 (强对偶定理) *在问题 (7.9) 中，假设 h 是对于 Ω 正规的，在可行域内 g 满足正规性条件. 假设 x^* 是问题 (7.9) 的解，设 $f(x^*) = f^*$. 那么对每个 λ 和 $\mu \geq 0$ 都有*

$$\varphi(\lambda, \mu) \leq f^*.$$

另外，存在 $\lambda, \mu \geq 0$ 使得

$$\varphi(\lambda, \mu) = f^*.$$

因此 $\varphi^* = f^*$. 与此同时，λ, μ 是该问题的 Lagrange 乘子.

证明 由 Lagrange 零阶条件定理（定理 7.6）可知：

$$\begin{aligned} f^* &= \min\{f(x) + \lambda^{\mathsf{T}} h(x) + \mu^{\mathsf{T}} g(x) : x \in \Omega\} \\ &= \varphi(\lambda, \mu) \leq \varphi^* \leq f^*. \end{aligned}$$

因此，$\varphi^* = f^*$，并且取等号的 λ, μ 是 Lagrange 乘子. □

从对偶原理我们可以写出对偶规划的一般形式：

$$\begin{array}{llcll} \multicolumn{2}{c}{\text{原始问题}} & \qquad & \multicolumn{2}{c}{\text{对偶问题}} \\ \min & \omega(y, z) & & \max & \varphi(\lambda, \mu) \\ \text{s.t.} & y = 0, & & \text{s.t.} & \lambda \in \mathbb{R}^m, \\ & z \leq 0. & & & \mu \geq 0. \end{array}$$

作为例子，下面我们给一个对偶规划的经济学解释.

例 7.3 (线性规划的经济学解释) *表 7.1 描述了公司甲用原料生产清洁剂的价格与存量表。*

甲用 3 种原料混合成 2 种清洁剂. 2 种清洁剂应该如何配制，使总价值最大?

设清洁剂 A 和 B 分别配制 x_1 和 x_2，我们可以把甲的目标写成一个规划问题：

$$\begin{array}{lrcl}\max & z = 12x_1 + 15x_2 & & \\ s.t. & 0.25x_1 + 0.50x_2 & \leq & 120, \\ & 0.50x_1 + 0.50x_2 & \leq & 150, \\ & 0.25x_1 & \leq & 50, \\ & x_1 & \geq & 0, \\ & x_2 & \geq & 0. \end{array}$$

现在有一个公司乙需要这 3 种原料，打算向甲购买，应付出多少钱?

乙向甲购买 3 种原料，出价分别为每吨 y_1, y_2, y_3 万元. 希望总价格尽量小，但不能低于甲用原料生产清洁剂所产生的价值，因此写出规划问题为：

$$\begin{array}{lrcl}\min & w = 120y_1 + 150y_2 + 50y_3 & & \\ s.t. & 0.25y_1 + 0.50y_2 + 0.25y_3 & \geq & 12, \\ & 0.50y_1 + 0.50y_2 & \geq & 15, \\ & y_1 & \geq & 0, \\ & y_2 & \geq & 0, \\ & y_3 & \geq & 0. \end{array}$$

注意到，以上两个规划问题恰好互为对偶问题.

§7.4 应用：支持向量机（SVM）

作为前面极值必要条件的一个具体应用，我们考虑一个经典的机器学习分类器：支持向量机（SVM）.

考虑二分类问题，输入 $x \in \mathbb{R}^n$，函数 f 输出一个 $\{-1, 1\}$ 中的值. 二分类问题的学习问题指的是给定训练集 $\{(x_i, y_i)\}_{i=1}^N$，找到 f 使得 $f(x_i) = y_i$. 假设训练集是线性可分的，例如，存在某个 $w \in \mathbb{R}^n$ 和 $b \in \mathbb{R}$ 使得

$$f(x) = \begin{cases} 1, & w^\mathsf{T} x + b > 0, \\ -1, & w^\mathsf{T} x + b < 0. \end{cases}$$

学习问题的首要目标是找到正确的以及最优的 w 和 b。本质上说，这就是一个找分离超平面的过程。那么，什么才叫最优呢? 从几何视角来看，一个自然的想法是最大化分离距离，即训练集中所有点到分离超平面的距离和的最小值，见图 7.8.

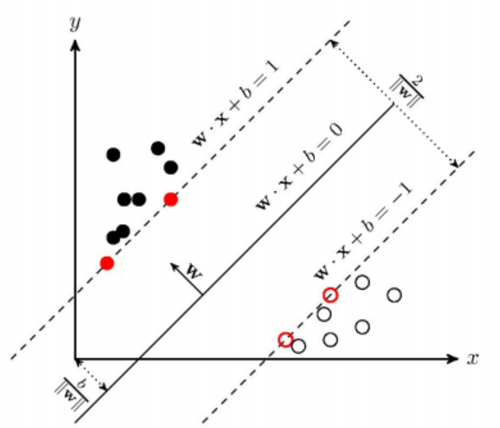


图 7.8: 分离距离示意图.

采样点 x_i 到分离超平面的归一化距离为

$$\gamma_i = y_i \left(\left(\frac{w}{\|w\|_2} \right)^\mathsf{T} x + \frac{b}{\|w\|_2} \right).$$

$\gamma = \min_i \gamma_i$ 是最小的归一化距离. 于是我们的任务变成了最大化 γ. 等价地，我们求解如下优化问题

$$\begin{aligned} \max_{w,b} \quad & \gamma \\ \text{s.t.} \quad & \gamma \le \gamma_i, \quad i = 1, 2, \ldots, N. \end{aligned}$$

$\gamma \le \gamma_i$ 等价于

$$y_i \left(\left(\frac{w}{\gamma \|w\|_2} \right)^\mathsf{T} x + \frac{b}{\gamma \|w\|_2} \right) \ge 1.$$

简洁起见，把 w 替换成 $\frac{w}{\gamma\|w\|_2}$，把 b 替换成 $\frac{b}{\gamma\|w\|_2}$，我们有

$$y_i(w^\mathsf{T} x + b) \ge 1.$$

那么最大化 $\gamma = \frac{1}{\|w\|_2}$ 等价于最小化 $\|w\|_2^2$.

我们得到以下凸规划问题:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & y_i(w^\mathsf{T} x_i + b) \ge 1, \quad i = 1, 2, \cdots, N. \end{aligned}$$

如何解决这个问题? 利用上面的对偶理论，我们有如下步骤:

- 第一步，用 Lagrange 乘子法，转化成 Lagrange 问题（min-max）.
- 第二步，写出对偶问题（max-min），验证强对偶定理的正规性条件，于是只需要求解对偶规划.
- 第三步，写出 KKT 条件，将对偶规划解为一个二次规划（min），用优化算法求解二次规划.

第八章　不动点理论

第四部分

逻辑与博弈

第九章　动态博弈

第十章　静态博弈

第五部分

认知逻辑

第十一章　模态逻辑基础

第十二章　认知逻辑与共同知识

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