标题 **title**

作者 *author*

2024 年 8 月 17 日

前言

目录

第一部分

AI 的逻辑

第二部分

信息与数据

第四章　高维几何，Johnson-Lindenstrauss 引理

作为生活在三维空间中的人，一个令我们倍感着迷的问题是：如果我们的世界是四维、五维甚至更高维的，我们会看到什么样的景象？Christopher Nolan 导演的电影《星际穿越》（Interstellar）给出了一个美妙的想象。

在近未来的地球上，资源匮乏和环境恶化使人类濒临灭绝。为了寻找新的生存空间，一群科学家和宇航员开始了一次前所未有的宇宙冒险。他们的目标是穿越一个神秘的虫洞，探索另一片星系中的适居行星。

然而，当主人公 Cooper 和他的团队成功穿越虫洞时，他们发现自己面对的并不仅仅是遥远的星系，还有更加不可思议的挑战。在探索的过程中，Cooper 最终进入了一个被称为 tesseract 的空间——一个超越我们三维世界的五维空间。在这个空间中，时间变得像空间一样可以自由导航，过去和未来不再是固定不变的线性进程，而是可以被观察和影响的维度。

通过操控时间维度，库珀能够在这个超立方体般的结构中，跨越时间的界限，影响他女儿 Murph 的命运，从而拯救全人类。这一情节不仅带给观众震撼的视觉体验，也揭示了一个深刻的物理与几何真理：在高维空间中，我们的直觉常常失效，必须依赖数学工具来理解和探索这些新维度的性质。

高维空间不仅是科学幻想中的概念，也是现实中的客观存在：当我们描述一个人的时候，我们会考虑他的年龄、身高、体重、学历、职业等多个属性，这些属性构成了一个多维的空间。在这个空间中，每个人都是一个点，而这些点之间的距离和关系，构成了我们对这个人的认知和理解。

在计算机的世界中，世界的一切都被表示为了数据，而且往往是高维数据。因此，如何理解和处理高维数据已经成为人工智能领域的一个重要问题。本章将探讨高维空间的几何性质，揭示出那些超越我们日常经验的奇异特性。

其中特别重要的是 Johnson-Lindenstrauss 引理，它揭示了高维空间“稀疏”的特性，

因而被广泛用于数据压缩. 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 高维几何

§4.1.1 高维球体

首先，我们探讨高维空间中“球体”的特殊性。按照微积分的方式（见附录 C.1.1），我们可以定义 n 维空间中的体积和表面积。定义

$$B_n = \{x \in \mathbb{R}^n : \|x\| \leq 1\}.$$

这是一个 n 维空间中的单位球体.

第一个反直觉的事实是，随着 n 趋于无穷，B_n 的体积和表面积都会趋于零！首先，我们给出 n 维单位球体的体积和表面积的计算公式：

定理 4.1 *记 V_n 为 n 维单位球体的体积，S_n 为 n 维单位球体的表面积。那么，*

$$V_n = \frac{\pi^{n/2}}{\Gamma(1+n/2)},$$
$$S_n = \frac{2\pi^{n/2}}{\Gamma(n/2)}.$$

其中 Γ 是 Gamma 函数，对自然数 n，它定义为

$$\Gamma\left(\frac{n}{2}\right) = \begin{cases} (m-1)!, & n = 2m, \\ \frac{(2m)!}{4^m m!}\sqrt{\pi}, & n = 2m+1. \end{cases}$$

这一定理的证明对积分的技巧要求较高，我们这里不给出，感兴趣的读者见习题[lhy: 出一下]。

这一定理的推论是，随着 n 的增大，V_n 和 S_n 都会趋于零：

推论 4.1

$$\lim_{n\to\infty} V_n = 0,$$
$$\lim_{n\to\infty} S_n = 0.$$

证明. 当 n 趋于无穷时，由 Stirling 公式，我们有

$$\Gamma(n) \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

因此，当 $n \to \infty$，

$$V_n \sim \frac{\pi^{n/2}}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} = \frac{1}{\sqrt{2\pi n}}\left(\frac{e}{n\sqrt{\pi}}\right)^n \to 0.$$

同理，S_n 的极限也为 0. □

接下来，我们说明，高维空间中的球体质量分布是极其不均匀的，大部分质量都集中在球体的边界上。定义一个半径为 r 的 n 维球为

$$B_n(r) = \{x \in \mathbb{R}^n : \|x\| \le r\}.$$

那么我们有：

命题 4.1 *对任意 $\epsilon \in (0,1)$，*

$$\frac{\lambda(B_n(1-\epsilon))}{\lambda(B_n(1))} = (1-\epsilon)^n.$$

其中 λ 是 $\mathbb{R}^n$ 上的 Lebesgue 测度（体积）。

因此，当 n 趋于无穷时，这一比值以指数速度趋于零。

证明. 利用体积（Lebesgue 测度）的性质（见附录 C.1.1），

$$\lambda(B_n(r)) = r^n \lambda(B_n(1)).$$

代入 $r = 1-\epsilon$，即可得证. □

接下来，我们再说明，如果把 n 维球和 n 维超立方体放在一起看，他们的质量分布也是非常反直觉的。对于 $n \in \mathbb{N}$，$\epsilon > 0$，定义我们定义一个 n 维的“球壳”$A_{n,\epsilon}$：

$$A_{n,\epsilon} = \{x \in \mathbb{R}^n : (1-\epsilon)\sqrt{n/3} < \|x\| < (1+\epsilon)\sqrt{n/3}\}.$$

这是一个半径为 $\sqrt{n/3}$ 的 n 维“球壳”，厚度为 2ϵ。

我们有如下定理：

定理 4.2

$$\lim_{n\to\infty} \frac{\lambda(A_{n,\epsilon} \cap (-1,1)^n)}{\lambda((-1,1)^n)} = 1,$$

其中 λ 是 $\mathbb{R}^n$ 上的 Lebesgue 测度（体积）。

我们来解释这一定理反直觉的地方。

- 如果我们只看方向 $(1,0,\ldots,0)$（或者其他“正方向”），半径为 $\sqrt{n/3}$ 的球壳应该远远超出了 $(-1,1)$ 的范围。然而，这一定理告诉我们，球壳在超立方体内占据了几乎全部的体积，这说明在其他的方向，球体以不可思议的方式被“压扁”了。
- 当 n 很大时，超立方体 $(-1,1)^n$ 的绝大部分体积都是由一个厚度为 ϵ，半径为 $\sqrt{n/3}$ 的 n 维球壳提供的，当 ϵ 很小时，这个球壳非常薄。一层薄球壳占据了一个实心立方体的绝大部分体积，这个在二维和三维空间中也是难以想象的。

接下来，我们证明这一定理。

证明. 为了证明这一定理，我们可以考虑 n 维随机变量的分布。

设 $X_1, X_2, \ldots, X_n$ 是独立同分布（i.i.d.）的随机变量，且服从均匀分布 $\mathcal{U}(-1,1)$。$Z_n = (X_1, X_2, \ldots, X_n)$ 服从均匀分布 $\mathcal{U}((-1,1)^n)$。对任意集合 $A \subseteq \mathbb{R}^n$，有

$$\Pr(Z_n \in A) = \frac{\lambda(A \cap (-1,1)^n)}{\lambda((-1,1)^n)}.$$

我们来计算 $\Pr(Z_n \in A_{n,\epsilon})$。

$Y_i = X_i^2$ 也是 i.i.d. 的，并且有

$$\mathbb{E}[Y_i] = \int_{-1}^{1} \frac{1}{2} \cdot x^2 \mathrm{d}x = \frac{1}{3}.$$

因为 $\mathsf{Var}[Y_i] \le \mathbb{E}[Y_i^2] \le 1$，由弱大数定律，$\sum_{i=1}^n X_i^2 / n$ 偏离期望 $1/3$ 某个值的概率会随着 n 趋于无穷而趋于零。更精确来说，当 $n \to \infty$ 时，

$$\Pr\left[\left|\frac{\sum_{i=1}^n X_i^2}{n} - \frac{1}{3}\right| > \epsilon\right] \to 0.$$

变形得

$$\underbrace{\Pr\left[(1-\epsilon)\sqrt{\frac{n}{3}} \le \sqrt{\sum_{i=1}^n X_i^2} \le (1+\epsilon)\sqrt{\frac{n}{3}}\right]}_{=\Pr[Z_n \in A_{n,\epsilon}]} \to 1.$$

于是，

$$\frac{\lambda(A_{n,\epsilon} \cap (-1,1)^n)}{\lambda((-1,1)^n)} = \Pr\left[Z_n \in A_{n,\epsilon}\right] \to 1.$$

这就完成了证明。 □

§4.1.2 Stein 悖论

接下来，我们转向更加抽象的高维空间，先考虑一维空间。假设 $X_1 \sim \mathcal{N}(\mu, 1)$，但我们并不知道 μ 是什么。通过随机采样得到了一个样本 $x_1 = 7$，怎样合理地估计 μ？既然没有多余的信息，我们不妨就猜 $\hat{\mu} = 7$，这是一个符合直觉的估计。

然后转向二维空间，假设 $(X_1, X_2) \sim \mathcal{N}(\mu, \mathbf{1}_2), \mu = (\mu_1, \mu_2)$，我们还是不知道 μ 是什么。同样，随机采样得到样本 $x_1 = 7, x_2 = 6$，怎样合理地估计 μ？我们似乎依然没有多余的选择，$\hat{\mu}_1 = 7, \hat{\mu}_2 = 6$ 看起来也是一个“好的”估计。

现在，转向一般的 n 维空间，$n \geq 3$. 假设

$$(X_1, X_2, \ldots, X_n) \sim \mathcal{N}(\mu, \mathbf{1}_n), \quad \mu = (\mu_1, \mu_2, \ldots, \mu_n),$$

μ 未知。随机采样得到样本 $x_1, x_2, \ldots, x_n$，怎样对 μ 进行估计?

直观上看，这似乎与一维、二维空间的情况并无区别。我们除了直接取

$$\hat{\mu} = (x_1, x_2, \ldots, x_n), \tag{4.1}$$

似乎没有更好的选择，我们把这种估计称为*朴素估计*。然而，在这部分我们将看到，在高维空间中，存在比朴素估计更好的估计方法！这就是 *Stein 悖论*。

为了衡量一个估计量方法的优劣程度，可以定义损失函数（均方误差）:

$$\ell = \mathbb{E}\left[\|\hat{\mu} - \mu\|^2\right].$$

均方误差越小，我们认为估计量越好。

然而，好坏这件事情似乎并没有那么简单，在一维的情况下，考虑如下估计量:

1. 一个估计量 $\hat{\mu}_1$ 是令 μ 等于得到的样本点，那么

$$\mathbb{E}\left[\|\hat{\mu}_1 - \mu\|^2\right] = \mathbb{E}\left[(x - \mu)^2\right] = \mathsf{Var}[x] = 1.$$

2. 另一个估计量 $\hat{\mu}_2$ 是令 μ 等于一个固定的值，比如 $\mu = 7$，那么

$$\mathbb{E}\left[\|\hat{\mu}_2 - \mu\|^2\right] = \mathbb{E}\left[(7 - \mu)^2\right] = (7 - \mu)^2.$$

我们不能明确说明哪一种估计量更好，因为如果 μ 在 7 附近，第二种方法会更好；但是如果 μ 在 0 附近，第一种方法会更好。

上面的例子表明，如果一个模型的参数 μ，我们很可能无法判断哪一种方法更好。但有一种情况，我们是可以明确说明一个方法 A 一定不好：有另外一个估计量在任何 μ 下都比它好。这就是如下定义:

定义 4.1 (可接受性) 考虑对参数 μ 的估计量方法 A，如果存在估计量方法 B，在任意的 μ 下都成立

$$\ell_B > \ell_A,$$

我们就称 A 方法是**不可接受的**。否则，我们称 A 方法是**可接受的**。

有了评判估计量好坏的标准，我们就可以引入 James-Stein 估计量了。

定义 4.2 (James-Stein 估计量) 假设采样得到的数据点是 $x_1, x_2, \ldots, x_n$，对参数

$$\mu = (\mu_1, \mu_2, \ldots, \mu_n)^\mathsf{T},$$

定义 **James-Stein 估计量**为：

$$\hat{\mu} = \left(1 - \frac{n-2}{x_1^2 + x_2^2 + \cdots + x_n^2}\right) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}. \tag{4.2}$$

定理 4.3 (Stein 悖论) *对于 $n \geq 3$，朴素估计 (4.1) 是不可接受的，具体来说，James-Stein 估计量 (4.2) 在任意 μ 下都比朴素估计更好。*

这一定理的证明十分具有技巧性，我们这里不给出证明，感兴趣的读者参阅第 4.5 节。

比起证明这个定理，更重要的问题是，为什么 James-Stein 估计量会比朴素估计更好？答案正是在于高维空间的反直觉性。

如图 4.1 所示，坐标轴上有一个圆心为 c 的单位圆，圆内随机选取一个点 x，那么

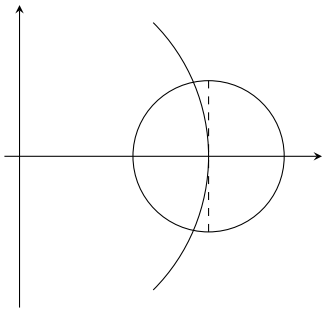
$$\Pr(\|x\| > \|c\|) > \frac{1}{2}.$$

如果不是二维的圆而是高维的球，这一不等式依然成立，并且这种效应随维数的增加而变强（见习题[lhy: 出一下]）。另一方面，圆心 c 离中心越远，这一概率越接近 $1/2$。

现在，我们回到最早的估计问题，我们把 μ 看作是圆心 c，而随机采样的点就看作是样本点。上面的概率不等式意味着，随着维数变高，朴素估计量 x 与真正的 μ 之间的的差距会越来越大。不仅如此，朴素估计量对 μ 的估计会偏大。

直觉上，要想更加精确估计 μ，我们需要比样本点 x 更接近圆心 c。仔细观察 (4.2)，James-Stein 估计量的确是这样做的。下面，我们详细介绍这一估计量的几何推导。

由于坐标系的选取是随意的，可以设一条坐标轴和 μ 的方向相同，其他 $n-2$ 根坐标轴方向随意，但正交。在新坐标系下，$\mu = (\|\mu\|, 0, \ldots, 0)^\mathsf{T}$。



[lhy: 重画]

图 4.1: 高维空间中的采样点

设样本是 $x = (x_1, x_2, \ldots, x_n)^\mathsf{T}$。损失函数可以被分解成两个部分的和：

$$\ell = (x_1 - \|\mu\|)^2 + \sum_{i=2}^{n} x_i^2. \tag{4.3}$$

令

$$\rho = \sqrt{\sum_{i=2}^{n} x_i^2}.$$

因为 x 是分量相互独立的 Gauss 向量，坐标轴旋转不改变分量之间的独立性，因此 ρ 服从自由度为 $n-1$ 的 χ 分布（可参见附录 C.4）。

假设样本点恰好满足 $x_1 = \|\mu\|$，而 x_i 以概率 1 都不为 0，可以画出图 4.2。

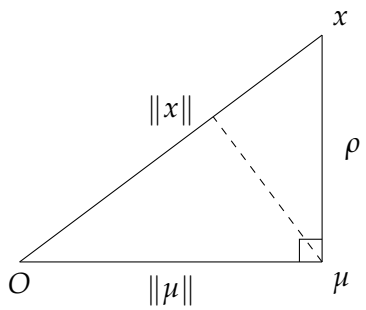


图 4.2: x 与 μ 形成的直角三角形

如果我们直接用样本点 x 作为估计量 $\hat{\mu}$，那么我们产生的偏差在于直边上，因此会有 ρ^2 的损失。现在，我们尝试移动 $\hat{\mu}$，来减少损失。

我们除了样本点 x 和原点 O 之外，其他任何信息都没有。因此，盲目的移动反而会带来更大的损失。结合前面的讨论，我们想要将 $\hat{\mu}$ 靠近原点，因此，一个合理的办法就是沿着斜边 Ox 向原点移动。

根据直角三角形的性质，当 $\hat{\mu}-\mu$ 与斜边垂直的时候，损失最小。我们来推导 $\hat{\mu}$ 的表达式。设 $\hat{\mu}=\alpha x$。根据三角形相似的原理，我们有

$$\begin{aligned}
&\frac{|O\hat{\mu}|}{|O\mu|}=\frac{|O\mu|}{|Ox|},\\
\iff &\frac{\alpha\|x\|}{\|\mu\|}=\frac{\|\mu\|}{\|x\|}\\
\iff &\alpha=\frac{\|\mu\|^2}{\|x\|^2}=1-\frac{\rho^2}{\|x\|^2}.
\end{aligned}$$

因此，新的估计量是

$$\hat{\mu}=\left(1-\frac{\rho^2}{\|x\|^2}\right)x=\left(1-\frac{\rho^2}{x_1^2+x_2^2+\cdots+x_n^2}\right)x.$$

然而，这一新估计量是不可计算的：因为我们只知道 O 和旋转之前的 x，所以我们没有办法计算 ρ. 为了得到 James-Stein 估计量，我们用一些数字特征来代替 ρ。对于自由度为 $k>1$ 的 χ 分布，其众数是 $\sqrt{k-1}$（见习题[lhy: 出一下]）。用众数来代替 ρ，就得到了 James-Stein 估计量。

最后，我们给出一些关于 Stein 悖论的讨论。

- 存在比 James-Stein 估计量更好的估计量。直观上，当样本 x 过于靠近原点的时候，$\|x\|$ 接近零，因此 James-Stein 估计量会穿过原点，往反方向跑到很远地方。这自然会带来很大的损失。因此，修正这一行为可以得到更好的估计量，比如

  $$\hat{\mu}=\mathsf{ReLU}\left(1-\frac{n-2}{x_1^2+x_2^2+\cdots+x_n^2}\right)x,$$

  其中

  $$\mathsf{ReLU}(x)=\begin{cases}x, & x>0,\\ 0, & x\le 0.\end{cases}$$

- 我们也可以从偏差-方差权衡的角度来理解 James-Stein 估计量。“距离”这一概念，在代数上，可以做如下分解：

  $$\begin{aligned}
  \mathbb{E}\left[(\hat{\mu}-\mu)^2\right]&=\mathsf{Var}\left[\hat{\mu}-\mu\right]+\left(\mathbb{E}\left[\hat{\mu}-\mu\right]\right)^2\\
  &=\mathsf{Var}[\hat{\mu}]+\left(\mathbb{E}\left[\hat{\mu}-\mu\right]\right)^2.
  \end{aligned}$$

  前半部分是方差（距离差的平方的期望），后半部分是偏差（距离差的期望的平方）。朴素估计是无偏差估计，但会引入很大的方差。通过适当引入偏差可能会减小方差，从而减小总体的预测误差，这就是 James-Stein 估计量的原理。

不过，如果你仔细观察会发现，实际上 (4.3) 的第一项就是方差，第二项是偏差。因而，偏差-方差权衡（代数直观）和几何直观其实在说同一件事情。

- 为什么 James-Stein 估计量在一维和二维空间中不起效？Lawrence Brown 证明了如下的定理：n 维空间中的朴素估计量是可接受的当且仅当 n 维空间中的简单对称随机游走以概率 1 无限次返回原点。这一惊人的联系揭示了这一问题的答案。

  一维和二维空间中的随机游走都会以概率 1 无限次返回原点，因此朴素估计量是可接受的，所以我们不可能找到更好的估计量。而更高维的空间中，随机游走以概率 1 无限次返回原点的概率是 0，因此朴素估计量是不可接受的，所以 James-Stein 估计量就起效了！

- Stein 悖论不意味着“中国茶叶的价格可以帮助预测墨尔本的降雨概率”。尽管我们总是可以把毫不相关的随机事件捆绑在一起，然后利用 James-Stein 估计量减少总体的预测误差，但是这不意味着对其中任何一个事件的预测会更准确。

§4.1.3 为什么我们要正则化？远有潜龙，勿用

在前两节中，我们用了一些简单、理想化的模型说明了高维空间的一些奇异性质。尽管真实的机器学习问题远比这些讨论要复杂，但他们所带来的启示不容忽视。

在机器学习中，我们可以把问题都归结为参数估计问题。虽然参数空间的“原点”看起来并没有什么不同，但原点这一概念本身确实带着人类先验的知识。例如，如果一个神经网络的所有参数都是零，无论输入是什么，它都会输出全零。同样，在 Stein 悖论中，靠近原点就是会产生更好的估计量。

在机器学习中，我们同样有偏好“原点”的倾向，例如使用 L^2 正则化这样的技术可以使模型变得更简单，因此不太可能过拟合。本节的内容通过极端的例子展示了正则化背后的原理：在高维空间中，离原点较远的地方体积远大于靠近原点的地方。因此，在高维空间中，向原点收缩一点就能减少大量的参数空间。

换句话说，对于一个大型机器学习模型来说，过拟合的方式远多于欠拟合的方式，所以我们倾向于让模型更偏向于欠拟合：欠拟合只会带来少部分问题，而过拟合带来的是数不胜数、千奇百怪的问题。

模型越远离原点，它的行为就越难以控制和解释。高维几何与 Stein 悖论给我们的启示是，远离原点就会有危险，而在高维空间中，稍微远离原点就会引入大量危险。化用《周易》的一句话：

> “远有潜龙，勿用。”[1]

[1]原句出自《周易 · 乾卦》，“初九：潜龙勿用。”孔子对这句话的解释是如果身居下位，时机还没有成熟，

§4.2 集中不等式

我们在前一节中阐述了高维空间中怪诞反直觉的性质。从本节开始，我们将阐述高维空间中随机变量的另一重属性：集中不等式。集中不等式说明的是，尽管整个空间非常庞大、难以理解，但如果随机变量具有某些性质，那么它们的取值就会集中在某个非常小的区域内，因而并没有我们所设想的那么复杂。利用这一原理，我们可以将非常高维的数据压缩到一个较为低维的空间中，从而可以驾驭他们。

接下来，我们先做一些准备工作，更详细的讨论参见附录 C。我们先引入示性函数的概念.

定义 4.3 (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 4.2 *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.

应当像潜藏的龙一样不要施展你的才干。这里，复杂的、远离原点模型就像是潜龙，隐藏着巨大的力量，但是现在人类对他们的理解还远远不够，因此我们应当保持谨慎，不要轻易使用。

- 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 4.3 *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（见命题 C.10）：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 4.4 (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：

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

□

注意，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 4.2 (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 4.5 (Hoeffding 不等式) *设 $X_1, \dots, X_n$ 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 $a = (a_1, \dots, a_n) \in \mathbb{R}^n$，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 $X_1, \dots, X_n$ 服从对称 Bernoulli 分布，得到（习题）[lhy: 习题]

$$e^{-\lambda t} \prod_i \mathbb{E}\left[\exp(\lambda a_i X_i)\right] \leq \exp\left(-\lambda t + \frac{\lambda^2}{2} \sum_i a_i^2\right). \tag{4.4}$$

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（见习题[lhy: 出一下]）.

定理 4.6 (Hoeffding 不等式，一般情形) *设 $X_1, \dots, X_n$ 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 4.7 (Chernoff 不等式) 设 $X_1, \ldots, X_n$ 是相互独立的随机变量，分别服从于参数为 $p_1, \ldots, p_n$ 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 4.8 (Johnson-Lindenstrauss 引理) 给定 N 个单位向量 $v_1, \dots, v_N \in \mathbb{R}^m$ 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.

通常来说，相对距离编码了很多重要的信息。例如，如果两个人的年龄、身高、体重等属性相差很小，那么他们也应该更相似。在这种观点下，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.

值得注意的是，J-L 引理中压缩空间的维数并不依赖于原始空间的维数，而只依赖于数据的数量. 因此，这对于一些抽象空间中数据的降维是非常有用的，见第 4.4 节.

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（参考附录 C.4）。

命题 4.4 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 4.4，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。不过，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 $u = \frac{v_i - v_j}{\|v_i - v_j\|_2}$ 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 $\mathbb{E}\left[e^{\lambda u_i^2}\right] = \sqrt{n/(n-2\lambda)}$（见习题[lhy: 出一下]），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

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

换言之，对任意 i,j，$\left|\left\|\frac{A(v_i-v_j)}{\|v_i-v_j\|_2}\right\|_2^2-1\right|<\epsilon$ 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 4.9 (J-L 引理，内积形式) 给定 N 个单位向量 $v_1,\dots,v_N\in\mathbb{R}^m$ 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：

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

我们将第一行乘 -1 加到第二行可以得到

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

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

§4.4 J-L 引理的应用

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（或内积）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. 下面我们将介绍两个具体的应用案例说明 J-L 引理如何指导我们在深度学习中选择合适的维度.

例 4.1 (词向量维度) 在为自然语言建立深度学习模型的时候，我们面对的首要问题就是如何在计算机中表示语言。我们以中文为例。如果我们看一句话，中文组成的基本单元是词。比如，这句话就可以被分解为如下成分：

比如 / ，/ 这句话 / 就 / 可以 / 被 / 分解 / 为 / 如下 / 成分

任何的中文句子都可以被这样分解，变成一个词的序列。这一过程被称为分词. 于是，为了表示一段中文，我们只需要能够表示所有的词。

于是，一个基本的任务就是如何表示一个词. 词向量就是这样的一个表示方法. 它的想法很简单：我们用一个向量空间来表示所有的词，每个词对应一个向量，我们用 $v(w)$ 来表示词 w 对应的向量.

很多时候，这些向量之间的关系可以反映出词之间的语义关系. 例如，“男人”和“女人”之间差异应该和“国王”和“女王”之间差异是类似的. 也就是说，我们希望

$$v(\text{男人}) - v(\text{女人}) \approx v(\text{国王}) - v(\text{女王}).$$

这里，$\approx$ 表示两个向量之间比较相似。具体来说，我们通常会用向量的内积来衡量相似度. 也就是说，给定一个词 w，我们希望

$$\langle v(w), v(\text{男人}) - v(\text{女人}) \rangle \approx \langle v(w), v(\text{国王}) - v(\text{女王}) \rangle .$$

我们可以想象，在真实的世界中，所有的词也构成了一个抽象向量空间，以类似的方式来表示词之间的关系，但这个向量空间的维数我们完全不得而知。对于计算机来说，我们需要选择一个确定的、压缩过的空间来表示词向量。

为了表示这种相似性，在压缩过的词向量空间中，所有词之间的内积应该尽量保持词本身的相似度。这正是内积形式的 J-L 引理（定理 4.9）所描述的情况。于是，J-L 引理告诉我们，$\mathcal{O}(\log N)$ 维空间足以容纳下 N 个单词，还保持了单词之间的相似性。

到此时，我们要十分警惕“理论指导实践”这一表述的理解。J-L 引理（理论）所指导的结论，成立的前提是“正态随机矩阵”，然而我们完全不清楚单词的空间是否符合这一条件。所以，我们不能说 J-L 引理直接给了我们合适的词向量维度，我们只能说 J-L 引理给了词向量选择的一个直觉。具体应该用多少维度的词向量，还需要实验来验证. □

例 4.2 (多头注意力) *注意力机制*是现代深度学习架构中最核心的模块之一。注意力机制的一种理解方式是将其看作一个键-值存储。想象我们把数据都存储到了一个数据库中，它的存储方式是 $\{k_i : v_i\}$，其中 k_i 是键，v_i 是值，例如“性别：男”就是一种典型的键值对，其中“性别”是键，而“男”是值。

现在，假设数据库中的所有键、值对是 $\{k_i : v_i\}_{i=1}^N$，我们希望从中找到与某个查询 q 最接近的键对应的值. 在深度学习中，键、值、查询都可以按照例 4.1 的方式表示成向量。类似地，我们可以用内积来衡量键和查询的相似度。

假设这些查询和键的向量都处在 $\mathbb{R}^d$ 中，那么注意力的计算公式为

$$a_j = \frac{e^{\langle q,k_j\rangle}}{\sum_{j=1}^N e^{\langle q,k_j\rangle}}.$$

换言之，我们把相似性转化为了概率分布，相似度越高的键，被选中的概率越大. 我们把 d 成为*注意力头大小*，在很多深度学习框架中，它被记为 `head_size`.

在很多场景下，一个数据库只能查询一种键-值对的关系，对于一些复杂的问题，我们可能需要查询多种不同的键-值对. 例如，对于中文来说，两个词的意思是否相近可以形成一个数据库，而两个词的词性是否相近又可以形成另一个数据库. 因此，我们需要多个注意力机制来查询不同的数据库，这就是*多头注意力*.

在简化的场景下，我们可以假设所有数据库里的 $k_i : v_i$ 对都是一致的，只是查询 q_i 不同. 那么，多头注意力就是要计算

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

类似例 4.1 的问题，如果真实世界中 a_{ij} 是 p_{ij}，我们应该如何选择向量维数 d 才能保证 a_{ij} 能够足够好地逼近 p_{ij} 呢？这个问题和例 4.1 是一样的.

在这个例子中，词向量的维度变成了 d，词表大小变成了数据库中的键值对数量 N. 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 个键值对，还能保持多组查询与键之间的相似性。

更为重要的是，这个压缩空间的维度 d 和查询的数量无关。这说明，如果有同样多的参数，头很大的单头注意力机制并不如头很小的多头注意力机制。此外，这也说明无论多少个头，多头注意力的 d 并不需要随着头的数量增加而显著增加.

同样地，J-L 引理只是给了我们一个直觉，具体的维度选择还需要实验来验证. □

§4.5 附录：Stein 悖论的证明

在本节，我们将给出 Stein 悖论（定理 4.3）的证明。

[lhy: TODO]

§4.6 习题

[lhy: TODO]

§4.7 章末注记

[lhy: TODO]

第三部分

决策与优化

第四部分

逻辑与博弈

第五部分

认知逻辑

第六部分

附录：预备知识

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