序言

编辑：詹好，赵志民，王茂霖

关于《机器学习理论导引》

近年来，机器学习领域发展迅猛，相关的课程与教材层出不穷。国内的经典教材如周志华的 **《机器学习》** 和李航的 **《统计学习方法》**，为许多学子提供了机器学习的入门指引。而在国外，Mitchell 的 *Machine Learning*、Duda 等人的 *Pattern Classification*、Alpaydin 的 *Introduction to Machine Learning* 等书籍则提供了更为系统的学习路径。对于希望深入学习的读者，Bishop 的 *Pattern Recognition and Machine Learning*、Murphy 的 *Machine Learning - A Probabilistic Perspective*、Hastie 等人的 *The Elements of Statistical Learning* 等著作也能提供详尽的理论指导。这些书籍无论在国内外，都成为了学习机器学习的重要资源。

然而，从**机器学习理论**的角度来看，现有的学习材料仍存在不足之处。相比于聚焦机器学习算法的著作，专注于机器学习理论的书籍未得到足够的重视。尽管上述一些经典著作中涉及到理论探讨，但篇幅有限，往往仅以独立章节或片段呈现，难以满足深入研究的需求。

以往的机器学习理论经典教材大多为英文撰写。上世纪末围绕统计学习理论展开的讨论，催生了诸如 Vapnik 的 *The Nature of Statistical Learning Theory* 和 *Statistical Learning Theory*，以及 Devroye 等人的 *A Probabilistic Theory of Pattern Recognition* 等经典文献。近年来，Shalev-Shwartz 和 Ben-David 的 *Understanding Machine Learning*，以及 Mohri 等人的 *Foundations of Machine Learning* 进一步推进了这一领域的发展。虽然部分经典著作已有高质量的中文译本，但由中文作者撰写的机器学习理论入门书籍仍显不足。

如今，周志华、王魏、高尉、张利军等老师合著的 **《机器学习理论导引》**（以下简称《导引》）填补了这一空白。该书以通俗易懂的语言，为有志于学习和研究机器学习理论的读者提供了良好的入门指引。全书涵盖了 **可学性、假设空间复杂度、泛化界、稳定性、一致性、收敛率、遗憾界** 七个重要的概念和理论工具。

尽管学习机器学习理论可能不像学习算法那样能够立即应用，但只要持之以恒，深入探究，必将能够领悟到机器学习中的重要思想，并体会其中的深邃奥妙。

-- 詹好

关于《机器学习理论导引》讲解笔记

《导引》的讲解笔记在团队内部被亲切地称为《钥匙书》。“钥匙”寓意着帮助读者开启知识之门，解答学习中的疑惑。

《导引》作为一本理论性较强的著作，涵盖了大量数学定理和证明。尽管作者团队已尽力降低学习难度，但由于机器学习理论本身的复杂性，读者仍需具备较高的数学基础。这可能导致部分读者在学习过程中感到困惑，影响学习效果。此外，由于篇幅限制，书中对某些概念和理论的实例说明不足，也增加了理解的难度。

基于以上原因，我们决定编辑这本《钥匙书》作为参考笔记，对《导引》进行深入的注解和补充。其目的是帮助读者更快理解并掌握书中内容，同时记录我们在学习过程中的思考和心得。

《钥匙书》主要包含以下四个部分：

1. **概念解释**：介绍书中涉及但未详细阐释的相关概念。
2. **证明补充**：详细解释部分证明的思路，并补充书中省略的证明过程。
3. **案例分享**：增加相关实例，帮助读者加深对抽象概念的理解。

鉴于《导引》第一章的内容简明易懂，《钥匙书》从第二章开始详细展开。

对我个人而言，《机器学习理论导引》与*Understanding Machine Learning*和*Foundations of Machine Learning*一样，都是既“无用”又“有用”的书籍。“无用”在于目前的经典机器学习理论尚难全面解释深度学习，尤其是现代生成式大模型的惊人表现。然而，我坚信未来的理论突破将基于现有研究成果，开创新的篇章。因此，分析结论可能并非最重要，真正宝贵的是其中蕴含的思想和分析

思路。数学作为一种强有力的工具，能够帮助我们更深入地理解和探索。我期望未来的深度学习能够拥有更多坚实的理论支撑，从而更好地指导实践。正如费曼所言：“What I cannot create, I do not understand.”——“凡我不能创造，我就不能理解。”希望大家能从这些理论中获得启发，创造出更有意义的成果。

另一方面，这本书也让我认识到自身的不足。不同于传统的机器学习算法教材，本书要求读者具备良好的数学功底，通过数学工具从更抽象的角度分析机器学习算法的性质，而非算法本身。学习之路或许漫长，但正如《牧羊少年的奇幻漂流》中所言：“每个人的寻梦过程都是以‘新手的运气’为开端，又总是以‘对远征者的考验’收尾。”希望大家能坚持经历考验，最终实现自己的梦想。

自《钥匙书》v1.0 版本发布以来，受到了众多学习者的关注。我们也收到了许多关于教材内容的疑问。为进一步深入理解相关知识，并记录团队对机器学习理论相关书籍的学习过程，我们将持续对《钥匙书》进行不定期更新，期待大家的关注。

-- 王茂霖

关于机器学习理论与实践

随着机器学习的蓬勃发展，**SOTA（State-of-the-art，最先进技术）**几乎成了评判算法优劣的唯一标准。这种对表面表现的单一追求，常常忽视了支撑其背后的基础理论。正如硅谷投资人吴军曾指出的，最顶尖的科学家通过理论设定学科的边界，赋予未来研究者方向和框架。1936年，图灵在其著名的论文中为可计算性理论奠定了基础，定义了哪些问题可以通过算法解决。同样，机器学习领域的研究者只有具备深厚的理论根基，才能在实践中面对瓶颈时不至于迷失，而是继续探索，甚至开拓新的领域。

没有免费午餐定理（No Free Lunch Theorem）便是一个鲜明的例子。该定理告诉我们，不存在一种能够应对所有问题的通用算法。尽管许多算法在特定领域或时间点看似“无敌”，如神经网络的兴起，但每个算法的优势往往局限于特定的任务和环境。因此，过度依赖某一种算法的短期成功，可能导致长期陷入困境。通过理论学习，研究者能够意识到这种局限，并避免在实践中过分追逐SOTA，而忽视更为长远的技术路线。

当然，理论和实践之间的差距依然存在。许多理论假设在现实应用中并不完全成立，尤其是在面对大数据和复杂模型时，理论的指导可能显得力不从心。但这并不意味着理论无用，恰恰相反，这正是学科发展的驱动力。正如机器学习的发展史所示，当实践进展遇到瓶颈时，往往是理论创新引领了新的突破。例如，在早期，受限于数据和算力，机器学习中的理论研究主导了整个领域的发展；而到了互联网时代，随着数据量的指数级增长和计算资源的提升，实践逐渐超越了理论。如今，面对数据、能源和算力等问题的挑战，理论又重新成为了优化模型效率和算法性能的焦点。

一个鲜明的例子是，近期在 ICLR 2024 大会上，斯隆奖得主马腾宇及其团队通过数学方法证明了Transformer 模型具备模拟任意多项式规模数字电路的能力。这一成果表明，随着**思维链（Chain of Thought, CoT）**的不断延展，Transformer能够有效地处理更为复杂的问题。 这项研究不仅展示了理论在推动前沿技术进步中的重要性，还让我们认识到，尽管外界对数据质量不足、模型的算力需求以及能源消耗提出了诸多质疑，但这些问题并非不可逾越。通过深入学习机器学习理论，我们可以更好地理解这些挑战，意识到它们实际上是迈向通用人工智能（AGI）**过程中必须面对和解决的关键节点。

不仅如此，理论学习还有助于我们建立对算法泛化能力的深刻理解。通过对机器学习理论的深入研究，我们能够推导出在不同假设条件下，算法的性能极限。比如，我们可以评估某一算法的收敛速度，预测其在不同数据量和模型复杂度下的表现。这些理论工具不仅提高了研究的严谨性，还为实际应用提供了有力的指导。例如，正是通过理论推导，我们能够理解大规模语言模型的训练为何需要如此庞大的数据集，同时又能预见在某些任务上微调模型的效果。

最后，掌握机器学习理论不仅能够为初学者奠定坚实的基础，增强他们的信心，还能帮助他们在面对外界质疑时保持理性和清醒的判断。无论是在研究中追求算法的提升，还是在实践中应对现实的挑战，理论的力量都不可或缺。在本书的编撰中，我们特别对部分证明进行了必要的诠释和展开，主要集中在原书中存在流畅性不足的内容，或那些虽提供了参考文献但证明篇幅不超过5页的论述。对于超出5页的文献，我们建议读者直接参阅原文，以便进行更深入的探究；此类情况在本书中出现频率较低，约不超过五处。

-- 赵志民

项目成员贡献与特别鸣谢

詹好负责了项目的初期规划与统筹，并参与了第一版的编辑和审核；赵志民主导了项目二期的更新与维护，并负责全书最终编辑和校验；李一飞参与了第1-5章内容的编辑；王茂霖参与了第2-6章内容的编辑。

另外，特别鸣谢了谢文睿和杨昱文，他们共同提供了本书的在线阅读功能；张雨对第2章的早期内容进行了修订，各成员的协作确保了本书高质量的编写和顺利完成。

主要符号表

x 标量
x 向量
A 矩阵
I 单位阵
$\mathcal{X}$ 样本空间或状态空间
$\mathcal{H}$ 假设空间
$\mathcal{D}$ 概率分布
D 数据样本(数据集)
$\mathbb{R}$ 实数集
$\mathbb{R}^+$ 正实数集
$\mathfrak{L}$ 学习算法
$(\cdot,\cdot,\cdot)$ 行向量
$(;;;)$ 列向量
$(\cdot)^T$ 向量或矩阵转置
$\cdots$ 集合
$[m]$ 集合 $\{1,\ldots,m\}$
$|\cdots|$ 集合 $\cdots$ 中元素的个数
$\|\cdot\|_p$ 范数, p 缺省时为 L_2 范数
$P()$, $P(\cdot\,|\,\cdot)$ 概率质量函数, 条件概率质量函数
$p(\cdot)$, $p(\cdot\,|\,\cdot)$ 概率密度函数, 条件概率密度函数
$E_{\cdot\sim\mathcal{D}}[f(\cdot)]$ 函数 $f(\cdot)$ 对 $\cdot$ 在分布 D 下的数学期望, 意义明确时将省略 D 和(或)$\cdot$
$\sup(\cdot)$ 上确界
$\inf(\cdot)$ 下确界
$\mathbb{I}(\cdot)$ 指示函数, 在 $\cdot$ 为真和假时分别取值为 $1,0$
$\mathrm{sign}(\cdot)$ 符号函数, 在 $\cdot<0,=0,>0$ 时分别取值为 $-1,0,1$

第1章：预备定理

编辑：赵志民, 李一飞

本章将对书中出现或用到的重要定理进行回顾，并简要解释其证明和应用场景。对于可能不熟悉相关基础知识的读者，建议参考附录中的基础知识部分。通过这些定理的阐述，希望帮助读者更好地理解数学推导的核心原理，并为后续章节的学习打下坚实基础。大数定律（Law of Large Numbers）和集中不等式（Concentration Inequality）密切相关，二者共同揭示了随机变量偏离其期望值的行为。大数定律说明，当样本量足够大时，样本均值会以概率收敛于总体的期望值，反映了长期平均结果的稳定性。而集中不等式（定理 1.8 至 1.18）则更进一步，为随机变量在有限样本量下偏离其期望值的可能性提供了精确的上界。这些不等式描述了随机变量偏离期望值的程度有多大，通过对概率的约束，确保这种偏离发生的概率较小，从而为各种随机现象提供了更细致的控制。集中不等式在大数定律的基础上提供了有力的工具，用于分析有限样本中的波动。

1.1 Jensen 不等式

对于任意凸函数 f，则有：

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)] \tag{1}$$

成立。

证明

设 $p(x)$ 为 X 的概率密度函数。由 Taylor 展开式及 f 的凸性，可知 $\exists \xi$ 使得：

$$\begin{aligned} f(x) &= f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(x - \mathbb{E}[X]) + \frac{f''(\xi)}{2}(x - \mathbb{E}[X])^2 \\ &\geq f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(x - \mathbb{E}[X]) \end{aligned} \tag{2}$$

对上式取期望，得到：

$$\begin{aligned} \mathbb{E}[f(X)] &= \int p(x) f(x)\, dx \\ &\geq f(\mathbb{E}[X]) \int p(x)\, dx + f'(\mathbb{E}[X]) \int p(x)(x - \mathbb{E}[X])\, dx \\ &= f(\mathbb{E}[X]) \end{aligned} \tag{3}$$

因此，原不等式得证。

如果 f 是凹函数，则 Jensen 不等式变为：

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)] \tag{4}$$

这一结论可以通过将上述证明中的 f 替换为 $-f$ 得到。□

1.2 Hölder 不等式

对于任意 $p, q \in \mathbb{R}^+$，且满足 $\frac{1}{p} + \frac{1}{q} = 1$，则有：

$$\mathbb{E}[|XY|] \leq (\mathbb{E}[|X|^p])^{\frac{1}{p}} (\mathbb{E}[|Y|^q])^{\frac{1}{q}} \tag{5}$$

成立。

证明

设 $f(x)$ 和 $g(y)$ 分别为 X 和 Y 的概率密度函数，定义：

$$M = \frac{|x|}{(\int_X |x|^p f(x)\, dx)^{\frac{1}{p}}}, \quad N = \frac{|y|}{(\int_Y |y|^q g(y)\, dy)^{\frac{1}{q}}} \tag{6}$$

代入 Young 不等式：

$$MN \leq \frac{1}{p} M^p + \frac{1}{q} N^q \tag{7}$$

对该不等式两边同时取期望：

$$\begin{aligned} \frac{\mathbb{E}[|XY|]}{(\mathbb{E}[|X|^p])^{\frac{1}{p}} (\mathbb{E}[|Y|^q])^{\frac{1}{q}}} &= \frac{\int_{XY} |xy| f(x) g(y)\, dx\, dy}{(\int_X |x|^p f(x)\, dx)^{\frac{1}{p}} (\int_Y |y|^q g(y)\, dy)^{\frac{1}{q}}} \\ &\leq \frac{\int_X |x|^p f(x)\, dx}{p \int_X |x|^p f(x)\, dx} + \frac{\int_Y |y|^q g(y)\, dy}{q \int_Y |y|^q g(y)\, dy} \\ &= \frac{1}{p} + \frac{1}{q} \\ &= 1 \end{aligned} \tag{8}$$

因此，Hölder 不等式得证。□

1.3 Cauchy-Schwarz 不等式

当 $p = q = 2$ 时，Hölder 不等式退化为 Cauchy-Schwarz 不等式：

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]} \tag{9}$$

1.4 Lyapunov 不等式

对于任意 $0 < r \leq s$，有：

$$\sqrt[r]{\mathbb{E}[|X|^r]} \leq \sqrt[s]{\mathbb{E}[|X|^s]} \tag{10}$$

证明

由 Hölder 不等式： 对任意 $p \geq 1$，有：

$$\begin{aligned} \mathbb{E}[|X|^r] &= \mathbb{E}[|X \cdot 1|^r] \\ &\leq (\mathbb{E}[|X|^{rp}])^{\frac{1}{p}} \cdot (\mathbb{E}[1^q])^{\frac{1}{q}} \\ &= (\mathbb{E}[|X|^{rp}])^{\frac{1}{p}} \end{aligned} \tag{11}$$

记 $s = rp \geq r$，则：

$$\mathbb{E}[|X|^r] \leq (\mathbb{E}[|X|^s])^{\frac{r}{s}} \tag{12}$$

因此，原不等式得证。□

1.5 Minkowski 不等式

对于任意 $p \geq 1$，有：

$$\sqrt[p]{\mathbb{E}[|X+Y|^p]} \leq \sqrt[p]{\mathbb{E}[|X|^p]} + \sqrt[p]{\mathbb{E}[|Y|^p]} \tag{13}$$

证明

由三角不等式和 Hölder 不等式，可得：

$$\begin{aligned} \mathbb{E}[|X+Y|^p] &\leq \mathbb{E}[(|X|+|Y|)|X+Y|^{p-1}] \\ &= \mathbb{E}[|X||X+Y|^{p-1}] + \mathbb{E}[|Y||X+Y|^{p-1}] \\ &\leq (\mathbb{E}[|X|^p])^{\frac{1}{p}} (\mathbb{E}[|X+Y|^{(p-1)q}])^{\frac{1}{q}} + (\mathbb{E}[|Y|^p])^{\frac{1}{p}} (\mathbb{E}[|X+Y|^{(p-1)q}])^{\frac{1}{q}} \\ &= [(\mathbb{E}[|X|^p])^{\frac{1}{p}} + (\mathbb{E}[|Y|^p])^{\frac{1}{p}}] \cdot \frac{\mathbb{E}[|X+Y|^p]}{(\mathbb{E}[|X+Y|^p])^{\frac{1}{p}}} \end{aligned} \tag{14}$$

化简后即得证。□

1.6 Bhatia-Davis 不等式

对 $X \in [a, b]$，有：

$$\mathbb{V}[X] \leq (b - \mathbb{E}[X])(\mathbb{E}[X] - a) \leq \frac{(b-a)^2}{4} \tag{15}$$

证明

因为 $a \leq X \leq b$，所以有：

$$\begin{aligned} 0 &\leq \mathbb{E}[(b-X)(X-a)] \\ &= -\mathbb{E}[X^2] - ab + (a+b)\mathbb{E}[X] \end{aligned} \tag{16}$$

因此，

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &\leq -ab + (a+b)\mathbb{E}[X] - \mathbb{E}[X^2] \\ &= (b - \mathbb{E}[X])(\mathbb{E}[X] - a)\end{aligned} \tag{17}$$

考虑 AM-GM 不等式：

$$xy \leq (\frac{x+y}{2})^2 \tag{18}$$

将 $x = b - \mathbb{E}[X]$ 和 $y = \mathbb{E}[X] - a$ 带入并化简即得证。□

1.7 Union Bound（Boole's）不等式

对于任意事件 X 和 Y，有：

$$P(X \cup Y) \leq P(X) + P(Y) \tag{19}$$

证明

根据概率的加法公式：

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y) \leq P(X) + P(Y) \tag{20}$$

由于 $P(X \cap Y) \geq 0$，因此不等式得证。□

1.8 Markov 不等式

若 $X \geq 0$，则对于任意 $\varepsilon > 0$，有：

$$P(X \geq \varepsilon) \leq \frac{\mathbb{E}[X]}{\varepsilon} \tag{21}$$

证明

由定义可得：

$$\mathbb{E}[X] = \int_0^\infty x p(x)\,dx \geq \int_\varepsilon^\infty x p(x)\,dx \geq \varepsilon \int_\varepsilon^\infty p(x)\,dx = \varepsilon P(X \geq \varepsilon) \tag{22}$$

因此，原不等式得证。□

1.9 Chebyshev 不等式

对于任意 $\varepsilon > 0$，有：

$$P(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathbb{V}[X]}{\varepsilon^2} \tag{23}$$

证明

利用 Markov 不等式，得到：

$$P(|X - \mathbb{E}[X]| \geq \varepsilon) = P((X - \mathbb{E}[X])^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{\varepsilon^2} = \frac{\mathbb{V}[X]}{\varepsilon^2} \tag{24}$$

因此，Chebyshev 不等式得证。□

1.10 Cantelli 不等式

对于任意 $\varepsilon > 0$，有：

$$P(X - \mathbb{E}[X] \geq \varepsilon) \leq \frac{\mathbb{V}[X]}{\mathbb{V}[X] + \varepsilon^2} \tag{25}$$

证明

设 $Y = X - \mathbb{E}[X]$，则对于任意 $\lambda \geq 0$，有：

$$\begin{aligned} P(X - \mathbb{E}[X] \geq \varepsilon) &= P(Y \geq \varepsilon) \\ &= P(Y + \lambda \geq \varepsilon + \lambda) \\ &= P((Y + \lambda)^2 \geq (\varepsilon + \lambda)^2) \\ &\leq \frac{\mathbb{E}[(Y + \lambda)^2]}{(\varepsilon + \lambda)^2} = \frac{\mathbb{V}[X] + \lambda^2}{(\varepsilon + \lambda)^2} \end{aligned} \tag{26}$$

通过对 λ 求导，得右端在 $\lambda = \frac{\mathbb{V}[X]}{\varepsilon}$ 时取得最小值 $\frac{\mathbb{V}[X]}{\mathbb{V}[X] + \varepsilon^2}$，因此：

$$P(X - \mathbb{E}[X] \geq \varepsilon) \leq \frac{\mathbb{V}[X]}{\mathbb{V}[X] + \varepsilon^2} \tag{27}$$

原不等式得证。□

值得注意的是，Cantelli 不等式是 Chebyshev 不等式的加强版，也称为单边 Chebyshev 不等式。通过类似的构造方法，可以推导出比 Cantelli 不等式更严格的上界。

1.11 Chernoff 界（Chernoff-Cramér 界）

对于任意 $\lambda > 0, \varepsilon > 0$，有：

$$P(X \geq \varepsilon) \leq \min_{\lambda > 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \varepsilon}} \tag{28}$$

对于任意 $\lambda < 0, \varepsilon > 0$，有：

$$P(X \leq \varepsilon) \leq \min_{\lambda < 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \varepsilon}} \tag{29}$$

证明

应用 Markov 不等式，有：

$$P(X \geq \varepsilon) = P(e^{\lambda X} \geq e^{\lambda \varepsilon}) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \varepsilon}}, \quad \lambda > 0, \varepsilon > 0 \tag{30}$$

同理，

$$P(X \leq \varepsilon) = P(e^{\lambda X} \leq e^{\lambda \varepsilon}) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \varepsilon}}, \quad \lambda < 0, \varepsilon > 0 \tag{31}$$

因此，Chernoff 界得证。□

基于上述 Chernoff 界的技术，我们可以进一步定义次高斯性：

定义 1 (随机变量的次高斯性)：若一个期望为零的随机变量 X 的矩母函数满足 $\forall \lambda \in \mathbb{R}^+$：

$$\mathbb{E}[e^{\lambda X}] \leq \exp(\frac{\sigma^2 \lambda^2}{2}) \tag{32}$$

则称 X 服从参数为 σ 的次高斯分布。

实际上，Hoeffding 引理中的随机变量 X 服从 $\frac{(b-a)}{2}$ 的次高斯分布。Hoeffding 引理也是次高斯分布的直接体现。次高斯性还有一系列等价定义，这里不作详细讨论。

次高斯分布有一个直接的性质：假设两个独立的随机变量 X_1, X_2 都是次高斯分布的，分别服从参数 σ_1, σ_2，那么 $X_1 + X_2$ 就是服从参数为 $\sqrt{\sigma_1^2 + \sigma_2^2}$ 的次高斯分布。这个结果的证明可以直接利用定义来完成。

显然，并非所有常见的随机变量都是次高斯的，例如指数分布。为此可以扩大定义：

定义 2 (随机变量的次指数性)：若非负的随机变量 X 的矩母函数满足 $\forall \lambda \in (0, a)$：

$$\mathbb{E}[e^{\lambda X}] \leq \frac{a}{a - \lambda} \tag{33}$$

则称 X 服从参数为 $(\mathbb{V}[X], 1/a)$ 的次指数分布。

同样地，次指数性也有一系列等价定义。一种不直观但更常用的定义如下：存在 (σ^2, b)，使得 $\forall |s| < 1/b$：

$$\mathbb{E}[e^{s(X - \mathbb{E}[X])}] \leq \exp(\frac{s^2\sigma^2}{2}) \tag{34}$$

常见的次指数分布包括：指数分布，Gamma 分布，以及**任何有界随机变量**。

类似地，次指数分布对于加法也是封闭的：如果 X_1, X_2 分别是服从 (σ_1^2, b_1) 和 (σ_2^2, b_2) 的次指数分布，那么 $X_1 + X_2$ 是服从 $(\sigma_1^2 + \sigma_2^2, \max(b_1, b_2))$ 的次指数分布。在高维统计问题中，次高斯分布和次指数分布的尾端控制能得到一些重要的结论。

1.12 Chernoff 不等式（乘积形式）

对于 m 个独立同分布的随机变量 $x_i \in [0, 1], i \in [m]$，设 $X = \sum_{i=1}^{m} X_i$，$\mu > 0$ 且 $r \leq 1$。若对所有 $i \leq m$ 都有 $\mathbb{E}[x_i] \leq \mu$，则：

$$\begin{aligned} P(X \geq (1 + r)\mu m) &\leq e^{-\frac{r^2 \mu m}{3}}, \quad r \geq 0 \\ P(X \leq (1 - r)\mu m) &\leq e^{-\frac{r^2 \mu m}{2}}, \quad r \geq 0 \end{aligned} \tag{35}$$

证明

应用 Markov 不等式，有：

$$P(X \geq (1 + r)\mu m) = P((1 + r)^X \geq (1 + r)^{(1+r)\mu m}) \leq \frac{\mathbb{E}[(1 + r)^X]}{(1 + r)^{(1+r)\mu m}} \tag{36}$$

由于 x_i 之间是独立的，可得：

$$\mathbb{E}[(1 + r)^X] = \prod_{i=1}^{m} \mathbb{E}[(1 + r)^{x_i}] \leq \prod_{i=1}^{m} \mathbb{E}[1 + r x_i] \leq \prod_{i=1}^{m} (1 + r\mu) \leq e^{r\mu m} \tag{37}$$

其中，第二步使用了 $\forall x \in [0, 1]$ 都有 $(1 + r)^x \leq 1 + rx$，第三步使用了 $\mathbb{E}[x_i] \leq \mu$，第四步使用了 $\forall x \in [0, 1]$ 都有 $1 + x \leq e^x$。

又由于 $\forall r \in [0, 1]$，有 $\frac{e^r}{(1+r)^{1+r}} \leq e^{-\frac{r^2}{3}}$，综上所述：

$$P(X \geq (1 + r)\mu m) \leq (\frac{e^r}{(1 + r)^{(1+r)}})^{\mu m} \leq e^{-\frac{r^2 \mu m}{3}} \tag{38}$$

当我们将 r 替换为 $-r$ 时，根据之前的推导，并利用 $\forall r \in [0, 1]$ 有 $\frac{e^r}{(1-r)^{1-r}} \leq e^{-\frac{r^2}{2}}$，可得第二个不等式的证明。□

1.13 最优 Chernoff 界

如果 X 是一个随机变量，并且 $\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq e^{\phi(\lambda)}$ 对于所有 $\lambda \geq 0$ 成立，则有以下结论：

$$P(X - \mathbb{E}X \geq \varepsilon) \leq e^{-\phi^*(\varepsilon)}, \quad \varepsilon \geq 0 \tag{39}$$

或

$$P(X - \mathbb{E}X \leq (\phi^*)^{-1}(\ln(1/\delta))) \geq 1 - \delta, \quad \delta \in [0, 1] \tag{40}$$

其中，ϕ^* 是 ϕ 的凸共轭函数，即 $\phi^*(x) = \sup_{\lambda \geq 0}(\lambda x - \phi(\lambda))$。

证明

根据 Chernoff 不等式，有：

$$\begin{aligned} P(X - \mathbb{E}X \geq \varepsilon) &\leq \inf_{\lambda \geq 0} e^{-\lambda\varepsilon}\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \\ &\leq \inf_{\lambda \geq 0} e^{\phi(\lambda)-\lambda\varepsilon} \\ &= e^{-\sup_{\lambda\geq 0}(\lambda\varepsilon-\phi(\lambda))} \\ &= e^{-\phi^*(\varepsilon)} \end{aligned} \tag{41}$$

因此，最优 Chernoff 界得证。□

1.14 Hoeffding 不等式

设有 m 个独立随机变量 $X_i \in [a_i, b_i]$，令 $\bar{X}$ 为 X_i 的均值。Hoeffding 不等式表示：

$$P(\bar{X} - \mathbb{E}[\bar{X}] \geq \varepsilon) \leq \exp(-\frac{2m^2\varepsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}) \tag{42}$$

证明

首先，我们引入一个引理 (Hoeffding 定理)：

对于 $\mathbb{E}[X] = 0$ 且 $X \in [a, b]$ 的随机变量，对于任意 $\lambda \in \mathbb{R}$，有：

$$\mathbb{E}[e^{\lambda X}] \leq \exp(\frac{\lambda^2(b-a)^2}{8}) \tag{43}$$

由于 e^x 是凸函数，对于任意 $x \in [a, b]$，可以写为：

$$e^{\lambda x} \leq \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b} \tag{44}$$

对上式取期望，得到：

$$\mathbb{E}[e^{\lambda X}] \leq \frac{b-\mathbb{E}[X]}{b-a}e^{\lambda a} + \frac{\mathbb{E}[X]-a}{b-a}e^{\lambda b} = \frac{be^{\lambda a} - ae^{\lambda b}}{b-a} \tag{45}$$

记 $\theta = -\frac{a}{b-a}$，$h = \lambda(b-a)$，则：

$$\frac{be^{\lambda a} - ae^{\lambda b}}{b-a} = [1 - \theta + \theta e^h]e^{-\theta h} = e^{\ln(1-\theta+\theta e^h)}e^{-\theta h} = e^{\ln(1-\theta+\theta e^h)-\theta h} \tag{46}$$

定义函数 $\varphi(\theta, h) = \ln(1 - \theta + \theta e^h) - \theta h$。注意到 θ 实际上与 h 无关。对 h 求偏导数：

$$\frac{\partial \varphi}{\partial h} = \frac{\theta e^h}{1 - \theta + \theta e^h} - \theta \tag{47}$$

显然有 $\frac{\partial \varphi}{\partial h}\big|_{h=0^+}=0$。同理，利用链式法则可得：

$$\frac{\partial^2 \varphi}{\partial h^2}=\frac{\theta e^h(1-\theta+\theta e^h)-\theta^2 e^{2h}}{(1-\theta+\theta e^h)^2}=\frac{\theta e^h}{1-\theta+\theta e^h}(1-\frac{\theta e^h}{1-\theta+\theta e^h})\leq \frac{1}{4} \tag{48}$$

根据泰勒展开式，可以得到：

$$\varphi(\theta,h)\leq \frac{h^2}{8}=\frac{\lambda^2(b-a)^2}{8} \tag{49}$$

由 Markov 不等式可知，对于任意 $\lambda>0$：

$$P(\bar{X}-\mathbb{E}[\bar{X}]\geq \varepsilon)=P(e^{\lambda(\bar{X}-\mathbb{E}[\bar{X}])}\geq e^{\lambda\varepsilon})\leq \frac{\mathbb{E}[e^{\lambda(\bar{X}-\mathbb{E}[\bar{X}])}]}{e^{\lambda\varepsilon}} \tag{50}$$

利用随机变量的独立性及 Hoeffding 引理，有：

$$\frac{\mathbb{E}[e^{\lambda(\bar{X}-\mathbb{E}[\bar{X}])}]}{e^{\lambda\varepsilon}}=e^{-\lambda\varepsilon}\prod_{i=1}^{m}\mathbb{E}[e^{\lambda(X_i-\mathbb{E}[X_i])/m}]\leq e^{-\lambda\varepsilon}\prod_{i=1}^{m}\exp(\frac{\lambda^2(b_i-a_i)^2}{8m^2}) \tag{51}$$

考虑二次函数 $g(\lambda)=-\lambda\varepsilon+\frac{\lambda^2}{8m^2}\sum_{i=1}^{m}(b_i-a_i)^2$，其最小值为 $-\frac{2m^2\varepsilon^2}{\sum_{i=1}^{m}(b_i-a_i)^2}$。

因此可以得到：

$$P(\bar{X}-\mathbb{E}[\bar{X}]\geq \varepsilon)\leq \exp(-\frac{2m^2\varepsilon^2}{\sum_{i=1}^{m}(b_i-a_i)^2}) \tag{52}$$

□

注意，这里并未要求随机变量同分布，因此Hoeffding 不等式常用来解释集成学习的基本原理。

1.15 McDiarmid 不等式

对于 m 个独立随机变量 $X_i\in\mathcal{X}$，若函数 f 是差有界的，则对于任意 $\varepsilon>0$，有：

$$P(f(X_1,\cdots,X_m)-\mathbb{E}[f(X_1,\cdots,X_m)]\geq\varepsilon)\leq\exp(-\frac{\varepsilon^2}{2\sum_{i=1}^{m}c_i^2}) \tag{53}$$

证明

构造一个鞅差序列：

$$D_j=\mathbb{E}[f(X)\mid X_1,\cdots,X_j]-\mathbb{E}[f(X)\mid X_1,\cdots,X_{j-1}] \tag{54}$$

容易验证：

$$f(X)-\mathbb{E}[f(X)]=\sum_{i=1}^{m}D_i \tag{55}$$

由于 f 是差有界的，因此满足 Azuma-Hoeffding 引理。代入后可得：

$$P(f(X_1,\cdots,X_m)-\mathbb{E}[f(X_1,\cdots,X_m)]\geq\varepsilon)\leq\exp(-\frac{\varepsilon^2}{2\sum_{i=1}^{m}c_i^2}) \tag{56}$$

原不等式得证。□

1.16 Bennett 不等式

对于 m 个独立随机变量 X_i，令 $\bar{X}$ 为 X_i 的均值，若存在 $b > 0$，使得 $|X_i - \mathbb{E}[X_i]| < b$，则有：

$$P(\bar{X} - \mathbb{E}[\bar{X}] \geq \varepsilon) \leq \exp(-\frac{m\varepsilon^2}{2(\sum_{i=1}^m \mathbb{V}[X_i]/m + b\varepsilon/3)}) \tag{57}$$

证明

首先，Bennett 不等式是 Hoeffding 不等式的一个加强版，对于独立随机变量的条件可以放宽为弱独立条件，结论仍然成立。

这些 Bernstein 类的集中不等式更多地反映了在非渐近观点下的大数定律表现，即它们刻画了样本均值如何集中在总体均值附近。

如果将样本均值看作是样本（数据点的函数），即令 $f(X_1, \cdots, X_m) = \sum_{i=1}^m X_i/m$，那么 Bernstein 类不等式刻画了如下的概率：

$$P(f(X_1, \cdots, X_m) - \mathbb{E}[f(X_1, \cdots, X_m)] \geq \varepsilon) \tag{58}$$

为了在某些泛函上也具有类似 Bernstein 类的集中不等式形式，显然 f 需要满足某些特定性质。差有界性是一种常见的约束条件。

定义 3: 差有界性

函数 $f : \mathcal{X}^m \to \mathbb{R}$ 满足对于每个 i，存在常数 $c_i < \infty$，使得：

$$|f(x_1, \cdots, x_i, \cdots, x_m) - f(x_1, \cdots, x_i', \cdots, x_m)| \leq c_i \tag{59}$$

则称 f 是差有界的。

为了证明这些结果，需要引入一些新的数学工具。

定义 4: 离散鞅

若离散随机变量序列（随机过程）Z_m 满足：

1. $\mathbb{E}[|Z_i|] < \infty$
2. $\mathbb{E}[Z_{m+1} \mid Z_1, \cdots, Z_m] = \mathbb{E}[Z_{m+1} \mid \mathcal{F}_m] = Z_m$

则称序列 Z_i 为离散鞅。

引理 2: Azuma-Hoeffding 定理

对于鞅 Z_i，若 $\mathbb{E}[Z_i] = \mu, Z_1 = \mu_\circ$，则构造鞅差序列 $X_i = Z_i - Z_{i-1}$，且 $|X_i| \leq c_i$，则对于任意 $\varepsilon > 0$，有：

$$P(Z_m - \mu \geq \varepsilon) = P(\sum_{i=1}^m X_i \geq \varepsilon) \leq \exp(-\frac{\varepsilon^2}{2\sum_{i=1}^m c_i^2}) \tag{60}$$

证明

首先，若 $\mathbb{E}[X \mid Y] = 0$，则有 $\forall \lambda > 0$：

$$\mathbb{E}[e^{\lambda X} \mid Y] \leq \mathbb{E}[e^{\lambda X}] \tag{61}$$

因此，由恒等式 $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$ 及 Chernoff 一般性技巧，对于任意 $\lambda > 0$：

$$\begin{aligned} P(Z_m - \mu \geq \varepsilon) &\geq e^{-\lambda\varepsilon}\mathbb{E}[e^{\lambda(Z_m - \mu)}] \\ &= e^{-\lambda\varepsilon}\mathbb{E}[\mathbb{E}[e^{\lambda(Z_m - \mu)} \mid \mathcal{F}_{m-1}]] \\ &= e^{-\lambda\varepsilon}\mathbb{E}[e^{\lambda(Z_{m-1} - \mu)}\mathbb{E}[e^{\lambda(Z_m - Z_{m-1})} \mid \mathcal{F}_{m-1}]] \end{aligned} \tag{62}$$

由于 $\{X_i\}$ 是鞅差序列，因此 $\mathbb{E}[X_m \mid \mathcal{F}_{m-1}] = 0, \mathbb{E}[X_i] = 0$。再结合不等式 $\mathbb{E}[e^{\lambda X} \mid Y] \leq \mathbb{E}[e^{\lambda X}]$ 及 Hoeffding 引理，有：

$$
\begin{aligned}
P(Z_m - \mu \geq \varepsilon) &\leq e^{-\lambda\varepsilon}\mathbb{E}[e^{\lambda(Z_{m-1}-\mu)}]\mathbb{E}[e^{\lambda X_n}] \\
&\leq e^{-\lambda\varepsilon}\mathbb{E}[e^{\lambda(Z_{m-1}-\mu)}]\exp(\frac{\lambda^2 c_m^2}{2})
\end{aligned} \tag{63}
$$

迭代上不等式可得：

$$
P(Z_m - \mu \geq \varepsilon) \leq e^{-\lambda\varepsilon}\prod_{i=1}^{m}\exp(\frac{\lambda^2 c_i^2}{2}) \tag{64}
$$

当 $\lambda = \frac{\varepsilon}{\sum_{i=1}^m c_i^2}$ 时，上式右端取得极小值：

$$
P(Z_m - \mu \geq \varepsilon) \leq \exp(-\frac{\varepsilon^2}{2\sum_{i=1}^m c_i^2}) \tag{65}
$$

原不等式得证。□

1.17 Bernstein 不等式

考虑 m 个独立同分布的随机变量 $X_i, i \in [m]$。令 $\bar{X} = \frac{\sum_{i=1}^m X_i}{m}$。若存在常数 $b > 0$，使得对所有 $k \geq 2$，第 k 阶矩满足 $\mathbb{E}[|X_i|^k] \leq \frac{k!b^{k-2}}{2}\mathbb{V}[X_1]$，则该不等式成立：

$$
\mathbb{P}(\bar{X} \geq \mathbb{E}[\bar{X}] + \epsilon) \leq \exp\left(\frac{-m\epsilon^2}{2\mathbb{V}[X_1] + 2b\epsilon}\right) \tag{66}
$$

证明

首先，我们需要将**矩条件**（moment condition）转换为**亚指数条件**（sub-exponential condition），以便进一步推导，即：

- **矩条件：** 对于随机变量 X，其 k-阶中心矩 满足如下条件：

$$
\mathbb{E}\left[|X - \mathbb{E}[X]|^k\right] \leq \frac{k!\, b^{k-2}}{2}\mathbb{V}[X], \quad \forall k \geq 2 \tag{67}
$$

  其中：

  1. **中心矩**：随机变量 X 的 k 阶中心矩为 $\mathbb{E}\left[|X - \mathbb{E}[X]|^k\right]$，表示 X 偏离其期望值的 k 次幂的期望值。中心矩用于衡量随机变量的分布形状，尤其是描述其尾部行为。当 $k = 2$ 时，中心矩即为随机变量的方差。
  2. $\frac{k!}{2}$ 是阶乘项，随着 k 增大迅速增长。
  3. b^{k-2} 是一个修正因子，其中 b 为常数，用以控制高阶矩的增长速率。
  4. $\mathbb{V}[X]$ 表示随机变量 X 的方差，它作为标准的离散度量来标定中心矩的大小。

- **亚指数条件**： 给定随机变量 X，其均值为 $\mathbb{E}[X]$，方差为 $\mathbb{V}[X]$，则其偏离均值的随机变量 $X - \mathbb{E}[X]$ 的矩母函数（MGF）满足如下不等式：

$$
\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\mathbb{V}[X]\lambda^2}{2(1-b\lambda)}\right), \quad \forall \lambda \in \left[0, \frac{1}{b}\right) \tag{68}
$$

  其中：

  1. **矩母函数**：这是一个重要的工具，用于控制随机变量的尾部概率。矩母函数的形式是 $\mathbb{E}[e^{\lambda X}]$，它通过调整 λ 来捕捉不同程度的偏差行为。
  2. **方差主导项**：不等式右边的表达式包含一个方差主导的项 $\frac{\mathbb{V}[X]\lambda^2}{2}$，类似于高斯分布的尾部特性，表明当 λ 较小时，X 的偏差行为主要由其方差控制，尾部概率呈现指数衰减。
  3. **修正项** $(1 - b\lambda)$：该项显示，当 λ 接近 $\frac{1}{b}$ 时，尾部偏差的控制变得更加复杂。这种形式通常出现在亚指数条件中，意味着随机变量的尾部行为介于高斯分布和重尾分布之间，尾部衰减较慢但仍比重尾分布快。

- **步骤 1：中心化随机变量**

设：

$$Y = X - \mathbb{E}[X] \tag{69}$$

我们的目标是对 Y 的矩母函数（MGF）进行上界：

$$\mathbb{E}\left[e^{\lambda Y}\right] \tag{70}$$

- **步骤 2：展开指数矩**

将 MGF 展开为幂级数（Taylor展开）：

$$\mathbb{E}\left[e^{\lambda Y}\right] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(\lambda Y)^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[Y^k] \tag{71}$$

由于 $\mathbb{E}[Y] = 0$，故 $k = 1$ 项消失：

$$\mathbb{E}\left[e^{\lambda Y}\right] = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[Y^k] \tag{72}$$

- **步骤 3：使用矩条件对中心矩进行上界**

根据矩条件：

$$\mathbb{E}\left[|Y|^k\right] \leq \frac{k!\, b^{k-2}}{2} \mathbb{V}[X] \tag{73}$$

因此：

$$|\mathbb{E}[Y^k]| \leq \mathbb{E}\left[|Y|^k\right] \leq \frac{k!\, b^{k-2}}{2} \mathbb{V}[X] \tag{74}$$

- **步骤 4：代入 MGF 展开式**

将上界代入 MGF 展开式：

$$\mathbb{E}\left[e^{\lambda Y}\right] \leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} \cdot \frac{k!\, b^{k-2}}{2} \mathbb{V}[X] = 1 + \frac{\mathbb{V}[X]}{2} \sum_{k=2}^{\infty} (b\lambda)^{k-2} \lambda^2 \tag{75}$$

通过令 $j = k - 2$ 进行简化：

$$\mathbb{E}\left[e^{\lambda Y}\right] \leq 1 + \frac{\mathbb{V}[X]\lambda^2}{2} \sum_{j=0}^{\infty} (b\lambda)^j \tag{76}$$

- **步骤 5：求解几何级数的和**

当 $b\lambda < 1$ 时，几何级数收敛：

$$\sum_{j=0}^{\infty} (b\lambda)^j = \frac{1}{1 - b\lambda} \tag{77}$$

因此：

$$\mathbb{E}\left[e^{\lambda Y}\right] \leq 1 + \frac{\mathbb{V}[X]\lambda^2}{2(1 - b\lambda)} \tag{78}$$

- **步骤 6：应用指数不等式**

使用不等式 $1 + x \leq e^x$ 对所有实数 x 成立：

$$\mathbb{E}\left[e^{\lambda Y}\right] \leq \exp\left(\frac{\mathbb{V}[X]\lambda^2}{2(1-b\lambda)}\right) \tag{79}$$

这与**亚指数条件**相符：

$$\mathbb{E}\left[e^{\lambda Y}\right] \leq \exp\left(\frac{\mathbb{V}[X]\lambda^2}{2(1-b\lambda)}\right), \quad \forall \lambda \in \left[0, \frac{1}{b}\right) \tag{80}$$

接下来我们完成在给定矩条件下的**Bernstein 不等式**的证明，即：

陈述：

给定 m 个独立同分布的随机变量 $X_i, i \in [m]$，令 $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$。若存在常数 $b > 0$，使得对所有 $k \geq 2$，

$$\mathbb{E}\left[|X_i - \mathbb{E}[X_i]|^k\right] \leq \frac{k!\, b^{k-2}}{2}\, \mathbb{V}[X_1], \tag{81}$$

则对于任意 $\epsilon > 0$，

$$\mathbb{P}\left(\bar{X} \geq \mathbb{E}[\bar{X}] + \epsilon\right) \leq \exp\left(\frac{-m\epsilon^2}{2\mathbb{V}[X_1] + 2b\epsilon}\right) \tag{82}$$

- **步骤 1：定义单侧 Bernstein 条件**

首先，回顾对于参数 $b > 0$ 的**单侧 Bernstein 条件**：

$$\mathbb{E}\left[e^{\lambda(Y)}\right] \leq \exp\left(\frac{\mathbb{V}[Y]\lambda^2/2}{1-b\lambda}\right), \quad \forall \lambda \in \left[0, \frac{1}{b}\right) \tag{83}$$

其中 $Y = X - \mathbb{E}[X]$。

根据**矩条件**，我们已经证明 Y 满足**亚指数条件**：

$$\mathbb{E}\left[e^{\lambda Y}\right] \leq \exp\left(\frac{\mathbb{V}[Y]\lambda^2}{2(1-b\lambda)}\right), \quad \forall \lambda \in \left[0, \frac{1}{b}\right) \tag{84}$$

因此，Y 满足**单侧 Bernstein 条件**，且 $\mathbb{V}[Y] = \mathbb{V}[X]$。

- **步骤 2：应用 Chernoff 界**

考虑 m 个独立同分布随机变量 $Y_i = X_i - \mathbb{E}[X_i]$ 的和：

$$S_m = \sum_{i=1}^{m} Y_i = m(\bar{X} - \mathbb{E}[\bar{X}]) \tag{85}$$

我们的目标是对概率 $\mathbb{P}(S_m \geq m\epsilon)$ 进行上界，这等价于 $\mathbb{P}(\bar{X} \geq \mathbb{E}[\bar{X}] + \epsilon)$。

使用**Chernoff 界**：

$$\mathbb{P}(S_m \geq m\epsilon) \leq \inf_{\lambda > 0} \exp(-\lambda m\epsilon)\mathbb{E}\left[e^{\lambda S_m}\right] \tag{86}$$

- **步骤 3：对和的矩母函数进行上界**

由于 Y_i 是独立的：

$$\mathbb{E}\left[e^{\lambda S_m}\right] = \prod_{i=1}^{m} \mathbb{E}\left[e^{\lambda Y_i}\right] \leq \left[\exp\left(\frac{\mathbb{V}[Y_i]\lambda^2}{2(1-b\lambda)}\right)\right]^m = \exp\left(\frac{m\mathbb{V}[Y]\lambda^2}{2(1-b\lambda)}\right) \tag{87}$$

因此：

$$\mathbb{P}(S_m \geq m\epsilon) \leq \inf_{\lambda>0} \exp\left(-\lambda m\epsilon + \frac{m\mathbb{V}[Y]\lambda^2}{2(1-b\lambda)}\right) \tag{88}$$

- **步骤 4：对 λ 进行优化**

为了找到最紧的界，我们需要对 λ 进行优化。最优的 λ 是使指数最小的值：

$$-\lambda m\epsilon + \frac{m\mathbb{V}[Y]\lambda^2}{2(1-b\lambda)} \tag{89}$$

对 λ 求导并令其为零：

$$-\epsilon + \frac{\mathbb{V}[Y]\lambda}{1-b\lambda} + \frac{\mathbb{V}[Y]\lambda^2 b}{2(1-b\lambda)^2} = 0 \tag{90}$$

然而，直接求解该方程较为复杂。我们可以选择：

$$\lambda = \frac{\epsilon}{\mathbb{V}[Y] + b\epsilon} \tag{91}$$

此时 λ 满足 $\left[0, \frac{1}{b}\right)$ 的范围，因为：

$$\lambda b = \frac{b\epsilon}{\mathbb{V}[Y] + b\epsilon} < 1 \tag{92}$$

- **步骤 5：将最优的 λ 代入界中**

将 $\lambda = \frac{\epsilon}{\mathbb{V}[Y]+b\epsilon}$ 代入指数中：

$$-\lambda m\epsilon + \frac{m\mathbb{V}[Y]\lambda^2}{2(1-b\lambda)} = -\frac{m\epsilon^2}{\mathbb{V}[Y]+b\epsilon} + \frac{m\mathbb{V}[Y]\left(\frac{\epsilon}{\mathbb{V}[Y]+b\epsilon}\right)^2}{2\left(1-\frac{b\epsilon}{\mathbb{V}[Y]+b\epsilon}\right)} \tag{93}$$

在第二项中简化分母：

$$1 - b\lambda = 1 - \frac{b\epsilon}{\mathbb{V}[Y]+b\epsilon} = \frac{\mathbb{V}[Y]}{\mathbb{V}[Y]+b\epsilon} \tag{94}$$

现在，代入回去：

$$-\frac{m\epsilon^2}{\mathbb{V}[Y]+b\epsilon} + \frac{m\epsilon^2}{2(\mathbb{V}[Y]+b\epsilon)} = -\frac{m\epsilon^2}{2(\mathbb{V}[Y]+b\epsilon)} \tag{95}$$

因此：

$$\mathbb{P}(S_m \geq m\epsilon) \leq \exp\left(-\frac{m\epsilon^2}{2(\mathbb{V}[Y]+b\epsilon)}\right) \tag{96}$$

- **步骤 6：回到样本均值**

回忆：

$$S_m = m(\bar{X} - \mathbb{E}[\bar{X}]) \tag{97}$$

因此：

$$\mathbb{P}\left(\bar{X} - \mathbb{E}[\bar{X}] \geq \epsilon\right) = \mathbb{P}(S_m \geq m\epsilon) \leq \exp\left(-\frac{m\epsilon^2}{2(\mathbb{V}[Y]+b\epsilon)}\right) \tag{98}$$

由于 $\mathbb{V}[Y] = \mathbb{V}[X]$，我们得到：

$$\mathbb{P}\left(\bar{X} \geq \mathbb{E}[\bar{X}] + \epsilon\right) \leq \exp\left(-\frac{m\epsilon^2}{2(\mathbb{V}[X]+b\epsilon)}\right) \tag{99}$$

□

1.18 Azuma–Hoeffding（Azuma）不等式

对于均值为 $Z_0 = \mu$ 的鞅差序列 $\{Z_m, m \geq 1\}$，若 $|Z_i - Z_{i-1}| \leq c_i$，其中$c_i > 0$为已知常数，则对于任意 $\varepsilon > 0$，有：

$$\begin{aligned} P(Z_m - \mu \geq \varepsilon) &\leq \exp(-\frac{\varepsilon^2}{2\sum_{i=1}^m c_i^2}) \\ P(Z_m - \mu \leq -\varepsilon) &\leq \exp(-\frac{\varepsilon^2}{2\sum_{i=1}^m c_i^2}) \end{aligned} \tag{100}$$

证明

1. **构造指数鞅**

   考虑参数 $s > 0$，构造如下的指数鞅：

   $$M_m = \exp(s(Z_m - \mu) - \frac{s^2}{2}\sum_{i=1}^m c_i^2) \tag{101}$$

   我们需要证明 $\{M_m\}_{m \geq 0}$ 是一个超鞅。

2. **验证鞅性质**

   对于任意 $m \geq 1$，有

   $$\mathbb{E}[M_m \mid \mathcal{F}_{m-1}] = \mathbb{E}[\exp(s(Z_m - Z_{m-1})) \mid \mathcal{F}_{m-1}] \cdot \exp(s(Z_{m-1} - \mu) - \frac{s^2}{2}\sum_{i=1}^m c_i^2) \tag{102}$$

   由于 $|Z_m - Z_{m-1}| \leq c_m$，并且 $\mathbb{E}[Z_m - Z_{m-1} \mid \mathcal{F}_{m-1}] = 0$（鞅性质），可以应用 Hoeffding 引理得到：

   $$\mathbb{E}[\exp(s(Z_m - Z_{m-1})) \mid \mathcal{F}_{m-1}] \leq \exp(s\mathbb{E}[Z_m - Z_{m-1} \mid \mathcal{F}_{m-1}] + \frac{s^2(c_m - (-c_m))^2}{8}) = \exp(\frac{s^2 c_m^2}{2}) \tag{103}$$

   因此，

   $$\mathbb{E}[M_m \mid \mathcal{F}_{m-1}] \leq \exp(\frac{s^2 c_m^2}{2}) \cdot \exp(s(Z_{m-1} - \mu) - \frac{s^2}{2}\sum_{i=1}^m c_i^2) = M_{m-1} \tag{104}$$

   这表明 $\{M_m\}$ 是一个超鞅。

3. **应用鞅不等式**

   由于 $\{M_m\}$ 是一个超鞅，且 $M_0 = \exp(0) = 1$，根据超鞅的性质，有

   $$\mathbb{E}[M_m] \leq M_0 = 1 \tag{105}$$

   对于事件 $\{Z_m - \mu \geq \varepsilon\}$，有

   $$M_m = \exp(s(Z_m - \mu) - \frac{s^2}{2}\sum_{i=1}^m c_i^2) \geq \exp(s\varepsilon - \frac{s^2}{2}\sum_{i=1}^m c_i^2) \tag{106}$$

   我们令 $a = \exp\left(s\varepsilon - \frac{s^2}{2}\sum_{i=1}^m c_i^2\right)$，由于 $\{Z_m - \mu \geq \varepsilon\}$ 蕴含了 $\{M_m \geq a\}$，所以：

   $$P\left(Z_m - \mu \geq \varepsilon\right) \leq P\left(M_m \geq a\right) \tag{107}$$

   结合已知的 $\mathbb{E}[M_m] \leq 1$，应用 Markov 不等式可得：

   $$P\left(M_m \geq a\right) \leq \frac{1}{a} = \exp\left(-s\varepsilon + \frac{s^2}{2}\sum_{i=1}^m c_i^2\right) \tag{108}$$

因此，我们得到：

$$P\left(Z_m-\mu\geq\varepsilon\right)\leq\exp\left(-s\varepsilon+\frac{s^2}{2}\sum_{i=1}^m c_i^2\right) \tag{109}$$

4. **优化参数 s**

   为了得到最优的上界，选择 s 使得表达式 $-s\varepsilon+\frac{s^2}{2}\sum c_i^2$ 最小化。对 s 求导并取零：

$$-\varepsilon+s\sum_{i=1}^m c_i^2=0\quad\Rightarrow\quad s=\frac{\varepsilon}{\sum_{i=1}^m c_i^2} \tag{110}$$

   代入得：

$$P(Z_m-\mu\geq\varepsilon)\leq\exp(-\frac{\varepsilon^2}{2\sum_{i=1}^m c_i^2}) \tag{111}$$

   这即是 Azuma 不等式的上侧不等式。

5. **下侧不等式的证明**

   对于下侧不等式，可以类似地考虑 $-Z_m$ 作为鞅，应用相同的方法得到：

$$P(Z_m-\mu\leq-\varepsilon)\leq\exp(-\frac{\varepsilon^2}{2\sum_{i=1}^m c_i^2}) \tag{112}$$

   因此，Azuma 不等式得证。□

1.19 Slud 不等式

若 $X\sim B(m,p)$，则有：

$$P(\frac{X}{m}\geq\frac{1}{2})\geq\frac{1}{2}[1-\sqrt{1-\exp(-\frac{m\varepsilon^2}{1-\varepsilon^2})}] \tag{113}$$

其中 $p=\frac{1-\varepsilon}{2}$。

证明

二项随机变量 X 表示在 m 次独立伯努利试验中成功的次数，成功概率为 p。对于大的 m，二项分布 $B(m,p)$ 可以近似为均值 $\mu=mp$ 和方差 $\sigma^2=mp(1-p)$ 的正态分布：

$$\begin{aligned}\mu&=\frac{m(1-\varepsilon)}{2}\\\sigma^2&=\frac{m(1-\varepsilon^2)}{4}\end{aligned} \tag{114}$$

令 $Z=\frac{X-\mu}{\sigma}$，代入 μ 和 σ，有：

$$P[\frac{X}{m}\geq\frac{1}{2}]=P[Z\geq\frac{\frac{m}{2}-\mu}{\sigma}]=P[Z\geq\frac{\varepsilon\sqrt{m}}{\sqrt{1-\varepsilon^2}}] \tag{115}$$

根据正态分布不等式（定理 21），有：

$$P[Z\geq x]\geq\frac{1}{2}[1-\sqrt{1-\exp(-\frac{2x^2}{\pi})}]\geq\frac{1}{2}[1-\sqrt{1-\exp(-x^2)}] \tag{116}$$

代入可得：

$$P[Z\geq\frac{\varepsilon\sqrt{m}}{\sqrt{1-\varepsilon^2}}]\geq\frac{1}{2}[1-\sqrt{1-\exp(-\frac{m\varepsilon^2}{1-\varepsilon^2})}] \tag{117}$$

□

1.20 上界不等式之加性公式

若 $\sup(f)$ 和 $\sup(g)$ 分别为函数 f 和 g 的上界，则有：

$$\sup(f+g) \leq \sup(f) + \sup(g) \tag{118}$$

证明

假设 f, g 分别有相同的定义域 D_f, D_g。根据上确界的定义，对于每一个 $x \in D_f \cap D_g$，我们有

$$g(x) \leq \sup_{y \in D_g} g(y), \tag{119}$$

从而

$$f(x) + g(x) \leq f(x) + \sup_{y \in D_g} g(y). \tag{120}$$

因为这对于每一个 $x \in D_f \cap D_g$ 都成立，我们可以在不等式的两边取上确界，得到：

$$\sup_{x \in D_f \cap D_g} (f(x) + g(x)) \leq \sup_{x \in D_f \cap D_g} f(x) + \sup_{y \in D_g} g(y) \leq \sup_{z \in D_f} f(z) + \sup_{y \in D_g} g(y). \tag{121}$$

这里我们使用了 $\sup_{x \in D_f \cap D_g} f(x) \leq \sup_{z \in D_f} f(z)$，因为 $D_f \cap D_g \subset D_f$。□

值得注意的是，该不等式在（4.33）中利用过两次，且原推导并没有用到 Jensen 不等式的任何性质。

另外，加性公式有几个常见的变形，例如：

$$\sup(f-g) - \sup(f-k) \leq \sup(k-g) \tag{122}$$

该不等式在（4.29）中出现过。

1.21 正态分布不等式

若 X 是一个服从标准正态分布的随机变量，那么对于任意 $u \geq 0$，有：

$$\mathbb{P}[X \leq u] \leq \frac{1}{2}\sqrt{1 - e^{-\frac{2}{\pi}u^2}} \tag{123}$$

证明

令 $G(u) = \mathbb{P}[X \leq u]$，则有：

$$2G(u) = \int_{-u}^{u} (2\pi)^{-1/2} e^{-x^2/2}\, dx = \int_{-u}^{u} (2\pi)^{-1/2} e^{-y^2/2}\, dy \tag{124}$$

因此：

$$2\pi[2G(u)]^2 = \int_{-u}^{u}\int_{-u}^{u} e^{-(x^2+y^2)/2}\, dx\, dy \tag{125}$$

让我们考虑更一般的积分形式：

$$2\pi[2G(u)]^2 = \iint_{R} e^{-(x^2+y^2)/2}\, dx\, dy \tag{126}$$

此时 R 为任意面积为 $4u^2$ 的区域。通过反证法可以证明，只有当 R 为以原点为中心的圆形区域 R_0 时，积分值最大：

$$R_0 = \{(x, y) : \pi(x^2 + y^2) \leq 4u^2\} \tag{127}$$

此时，有：

$$\begin{aligned} 2\pi[2G(u)]^2 &\leq \iint_{R_0} e^{-(x^2+y^2)/2}\, dx\, dy \\ &= \int_0^{2\pi} \int_0^{2u\pi^{-1/2}} e^{-r^2/2} r\, dr\, d\varphi \\ &= 2\pi(1 - e^{-2u^2/\pi}) \end{aligned} \tag{128}$$

因此，有：

$$G(u) = \mathbb{P}[X \leq u] \leq \frac{1}{2}\sqrt{1 - e^{-\frac{2}{\pi}u^2}} \tag{129}$$

进一步，我们可以得到：

$$\mathbb{P}[X \geq u] \geq \frac{1}{2}(1 - \sqrt{1 - e^{-\frac{2}{\pi}u^2}}) \tag{130}$$

□

1.22 AM-GM 不等式

算术平均数和几何平均数的不等式，简称 AM-GM 不等式。该不等式指出非负实数序列的算术平均数大于等于该序列的几何平均数，当且仅当序列中的每个数相同时，等号成立。形式上，对于非负实数序列 $\{x_n\}$，其算术平均值定义为：

$$A_n = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{131}$$

其几何平均值定义为：

$$G_n = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{132}$$

则 AM-GM 不等式成立：

$$A_n \geq G_n \tag{133}$$

证明

我们可以通过 Jensen 不等式来证明 AM-GM 不等式。首先，我们考虑函数 $f(x) = -\ln x$，该函数是凸函数，因此有：

$$\frac{1}{n}\sum_{i=1}^{n} -\ln x_i \geq -\ln(\frac{1}{n}\sum_{i=1}^{n} x_i) \tag{134}$$

即：

$$\begin{aligned} &\ln(\frac{1}{n}\sum_{i=1}^{n} x_i) \geq \frac{1}{n}\sum_{i=1}^{n} \ln x_i = \ln(\sqrt[n]{\prod_{i=1}^{n} x_i}) \\ &\Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i \geq \sqrt[n]{\prod_{i=1}^{n} x_i} \end{aligned} \tag{135}$$

当取 $x_1 = x_2 = \cdots = x_n$ 时，等号成立。特别地，当 $n = 2$ 时，我们有：

$$\frac{x_1+x_2}{2} \geq \sqrt{x_1 x_2} \tag{136}$$

□

1.23 Young 不等式

对于任意 $a, b \geq 0$ 且 $p, q > 1$，若 $\frac{1}{p} + \frac{1}{q} = 1$，则有：

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \tag{137}$$

当且仅当 $a^p = b^q$ 时，等号成立。

证明

我们可以通过 Jensen 不等式来证明 Young 不等式。首先，当 $ab = 0$ 时，该不等式显然成立。当 $a, b > 0$ 时，我们令 $t = 1/p, 1-t = 1/q$，根据 $\ln(x)$ 的凹性，我们有：

$$\begin{aligned} \ln(ta^p + (1-t)b^q) &\geq t\ln(a^p) + (1-t)\ln(b^q) \\ &= \ln(a) + \ln(b) \\ &= \ln(ab) \end{aligned} \tag{138}$$

当且仅当 $a^p = b^q$ 时，等号成立。□

1.24 Bayes 定理

贝叶斯定理是概率论中的一个重要定理，它描述了在已知某些条件下更新事件概率的数学方法。贝叶斯定理的公式为：

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{139}$$

其中：

- $P(A|B)$ 是在事件 B 发生的情况下事件 A 发生的后验概率。
- $P(B|A)$ 是在事件 A 发生的情况下事件 B 发生的似然函数。
- $P(A)$ 是事件 A 的先验概率。
- $P(B)$ 是事件 B 的边缘概率。

证明

根据条件概率的定义，事件 A 在事件 B 发生下的条件概率 $P(A|B)$ 表示为：

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{140}$$

同样地，事件 B 在事件 A 发生下的条件概率 $P(B|A)$ 表示为：

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{141}$$

通过这两个公式可以得到联合概率 $P(A \cap B)$ 的两种表示方式：

$$P(A \cap B) = P(A|B) \cdot P(B) \tag{142}$$

以及：

$$P(A \cap B) = P(B|A) \cdot P(A) \tag{143}$$

由于联合概率的性质，我们可以将上述两个等式等同：

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \tag{144}$$

将上述等式两边同时除以 $P(B)$，得到贝叶斯定理：

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{145}$$

□

通过先验和后验的更新过程，贝叶斯统计提供了一种动态的、不断修正认知的不确定性量化方法。

1.25 广义二项式定理

广义二项式定理（Generalized Binomial Theorem）是二项式定理的扩展：

$$(x+y)^r = \sum_{k=0}^{\infty} \binom{r}{k} x^{r-k} y^k, \quad |x| < |y|, \quad k \in \mathbb{N}, \quad r \in \mathbb{R} \tag{146}$$

其中我们令 $\binom{r}{k} := \frac{(r)_k}{k!}$，$(r)_k = r(r-1)\cdots(r-k+1)$ 为递降阶乘（falling factorial）。

证明

首先代入定义，易证：

$$(r-k)\binom{r}{k} + (r-(k-1))\binom{r}{k-1} = r\binom{r}{k} \tag{147}$$

我们从特殊情况 $y=1$ 开始。首先我们证明只要 $|x|<1$，后者级数就会收敛。

通过使用幂级数收敛半径的商式来证明这一点，由于绝对值的连续性使我们可以先在绝对值内部计算极限，可得：

$$\lim_{k\to\infty} \frac{|a_k|}{|a_{k+1}|} = \lim_{k\to\infty} |\frac{k+1}{r-k}| = |-1| = 1 \tag{148}$$

因此我们有一个为 1 的收敛半径。这种收敛使我们能够在 $|x|<1$ 的收敛区域内应用逐项求导，得到：

$$\frac{d}{dx}\sum_{k=0}^{\infty}\binom{r}{k}x^k = \sum_{k=1}^{\infty}(r-(k-1))\binom{r}{k-1}x^{k-1} \tag{149}$$

如果我们将我们正在考虑的级数定义的函数记为 $g(x)$，我们得到：

$$\begin{aligned}
(1+x)\frac{d}{dx}g(x) &= \sum_{k=1}^{\infty}(r-(k-1))\binom{r}{k-1}x^{k-1} + \sum_{k=1}^{\infty}(r-(k-1))\binom{r}{k-1}x^k \\
&= r + \sum_{k=1}^{\infty}((r-k)\binom{r}{k} + (r-(k-1))\binom{r}{k-1})x^k \\
&= r + r\sum_{k=1}^{\infty}\binom{r}{k}x^k \\
&= rg(x),
\end{aligned} \tag{150}$$

上式的推导使用了前述引理。

现在定义 $f(x) = (1+x)^r$，我们通过通常的求导规则得到：

$$\frac{d}{dx}(\frac{g(x)}{f(x)}) = \frac{g'(x)f(x) - f'(x)g(x)}{f(x)^2} = \frac{r\frac{g(x)}{x+1}(1+x)^r - rg(x)(1+x)^{r-1}}{f(x)^2} = 0 \tag{151}$$

$|x|<1$ 意味着 $f(x)\neq 0$，因此 g/f 为常数。又 $f(0)=g(0)=1$ 可得 $f(x)=g(x)$。

对于一般的 $x,y\in\mathbb{R}$ 且 $|x|<|y|$，我们有：

$$\frac{(x+y)^r}{y^r}=(\frac{x}{y}+1)^r=\sum_{k=0}^{\infty}\binom{r}{k}(\frac{x}{y})^k; \tag{152}$$

收敛性由假设 $|x/y|<1$ 保证。为了得到原定理的形式，我们只需乘以 y^r 即可。□

1.26 Stirling 公式

Stirling 公式是用于近似计算阶乘的一种公式，即使在 n 很小时也有很高的精度。Stirling 公式的一种形式为：

$$n!=\sqrt{2\pi}n^{n+1/2}e^{-n}e^{r_n} \tag{153}$$

其中，$\frac{1}{12n+1}<r_n<\frac{1}{12n}$。

证明

我们令：

$$S_n=\ln(n!)=\sum_{p=1}^{n-1}\ln(p+1) \tag{154}$$

且

$$\ln(p+1)=A_p+b_p-\varepsilon_p \tag{155}$$

其中：

$$\begin{aligned}A_p&=\int_p^{p+1}\ln x\,dx\\ b_p&=\frac{1}{2}[\ln(p+1)-\ln(p)]\\ \varepsilon_p&=\int_p^{p+1}\ln x\,dx-\frac{1}{2}[\ln(p+1)+\ln(p)]\end{aligned} \tag{156}$$

此时：

$$S_n=\sum_{p=1}^{n-1}(A_p+b_p-\varepsilon_p)=\int_1^n\ln x\,dx+\frac{1}{2}\ln n-\sum_{p=1}^{n-1}\varepsilon_p \tag{157}$$

易证 $\int\ln x\,dx=x\ln x-x+C,\ C\in\mathbb{R}$，故：

$$S_n=(n+1/2)\ln n-n+1-\sum_{p=1}^{n-1}\varepsilon_p \tag{158}$$

此时：

$$\varepsilon_p=\frac{2p+1}{2}\ln(\frac{p+1}{p})-1 \tag{159}$$

接下来我们对 $\ln(\frac{p+1}{p})$ 进行级数展开，根据广义二项式定理，即：

令 $a=-1,\ t=\frac{1}{p},\ t\in(-1,1)$，则有：

$$\frac{1}{1+t}=1-t+t^2-t^3+t^4-\cdots \tag{160}$$

对上式两边同时进行积分，我们有：

$$\ln(1+t)=t-\frac{1}{2}t^2+\frac{1}{3}t^3-\frac{1}{4}t^4+\cdots \tag{161}$$

如果我们令 $-t$ 来代替 t，则有：

$$\ln\frac{1}{1-t}=t+\frac{1}{2}t^2+\frac{1}{3}t^3+\frac{1}{4}t^4+\cdots \tag{162}$$

将两式相加，我们有：

$$\frac{1}{2}\ln\frac{1+t}{1-t}=t+\frac{1}{3}t^3+\frac{1}{5}t^5+\cdots \tag{163}$$

回到我们的问题，我们令 $t=(2p+1)^{-1}\in(0,1)$，如此才满足 $\frac{1+t}{1-t}=\frac{p+1}{p}$，带入前式：

$$\varepsilon_p=\frac{1}{3(2p+1)^2}+\frac{1}{5(2p+1)^4}+\frac{1}{7(2p+1)^6}+\cdots \tag{164}$$

因此：

$$\varepsilon_p<\frac{1}{3(2p+1)^2}\sum_{i=0}^{\infty}\frac{1}{(2p+1)^{2i}}=\frac{1}{3(2p+1)^2}\frac{1}{1-\frac{1}{(2p+1)^2}}=\frac{1}{3[(2p+1)^2-1]}=\frac{1}{12}(\frac{1}{p}-\frac{1}{p+1}) \tag{165}$$

且

$$\varepsilon_p>\frac{1}{3(2p+1)^2}\sum_{i=0}^{\infty}\frac{1}{[3(2p+1)^2]^i}=\frac{1}{3(2p+1)^2}\frac{1}{1-\frac{1}{3(2p+1)^2}}=\frac{1}{3(2p+1)^2-1} \tag{166}$$

易证

$$(p+\frac{1}{12})(p+1+\frac{1}{12})=p^2+\frac{7}{6}p+\frac{13}{144}>p^2+p+\frac{1}{6}=\frac{1}{12}[3(2p+1)^2-1],\quad p\in\mathbb{N}^+ \tag{167}$$

因此：

$$\varepsilon_p>\frac{1}{12}(\frac{1}{p+\frac{1}{12}}-\frac{1}{p+1+\frac{1}{12}}) \tag{168}$$

我们令：

$$B=\sum_{p=1}^{\infty}\varepsilon_p,\quad r_n=\sum_{p=n}^{\infty}\varepsilon_p \tag{169}$$

那么易得：

$$\frac{1}{13}<B<\frac{1}{12},\quad \frac{1}{12(n+1)}<r_n<\frac{1}{12n} \tag{170}$$

带入 S_n 的表达式：

$$S_n=(n+\frac{1}{2})\ln n-n+1-B+r_n \tag{171}$$

可得：

$$n!=e^{1-B}n^{n+1/2}e^{-n}e^{r_n} \tag{172}$$

令 $C=e^{1-B}$，我们可知常数 C 的取值范围为 $(e^{11/12},e^{12/13})$，此处我们取 $C=\sqrt{2\pi}$，该公式得证。$\square$

1.27 散度定理

散度定理（Divergence Theorem），也称为高斯定理（Gauss's Theorem），是向量分析中的重要定理，它将体积积分和曲面积分联系起来。

具体而言，如果考虑一个 n-维球体（n-ball）B^n 的体积为 V，其表面为 S^{n-1}，对于一个位于 n-维空间中的光滑向量场 $\mathbf{F}$，则有：

$$\int_{B^n} (\nabla \cdot \mathbf{F})\, dV = \oint_{S^{n-1}} \mathbf{F} \cdot \mathbf{n}\, dS$$

其中：

- $\nabla \cdot \mathbf{F}$ 是向量场 $\mathbf{F}$ 的散度。
- dV 是体积元素。
- dS 是边界表面的面积元素。
- $\mathbf{n}$ 是边界的单位外法向量。

体积积分计算的是在 n-球内的散度，而表面积分计算的是在 $n-1$ 维球面上的通量。 这种形式的散度定理在物理学和工程学中广泛应用，比如电磁学中的高斯定理、流体力学中的质量守恒等。

1.28 分离超平面定理

如果有两个不相交的非空凸集，则存在一个超平面能够将它们完全分隔开，这个超平面叫做分离超平面（Separating Hyperplane）。形式上，设 A 和 B 是 $\mathbb{R}^n$ 中的两个不相交的非空凸集，那么存在一个非零向量 v 和一个实数 c，使得：

$$\langle x, v\rangle \geq c \text{ 且 } \langle y, v\rangle \leq c \tag{173}$$

对所有 $x \in A$ 和 $y \in B$ 都成立。即超平面 $\langle \cdot, v\rangle = c$ 以 v 作为分离轴（Separating Axis），将 A 和 B 分开。

进一步，如果这两个集合都是闭集，并且至少其中一个是紧致的，那么这种分离可以是严格的，即存在 $c_1 > c_2$ 使得：

$$\langle x, v\rangle > c_1 \text{ 且 } \langle y, v\rangle < c_2 \tag{174}$$

在不同情况下，我们可以通过调整 v 和 c 来使得分离超平面的边界更加清晰。

A	B	$\langle x, v\rangle$	$\langle y, v\rangle$
闭紧集	闭集	$> c_1$	$< c_2$ 且 $c_2 < c_1$
闭集	闭紧集	$> c_1$	$< c_2$ 且 $c_2 < c_1$
开集	闭集	$> c$	$\leq c$
开集	开集	$> c$	$< c$

在支持向量机的背景下，最佳分离超平面（或最大边缘超平面）是分离两个点凸包并且与两者等距的超平面。

证明

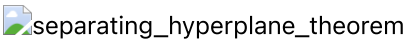
证明基于以下引理：

设 A 和 B 是 $\mathbb{R}^n$ 中两个不相交的闭集，且假设 A 是紧致的。则存在点 $a_0 \in A$ 和 $b_0 \in B$ 使得 $\|a - b\|$ 在 $a \in A$ 和 $b \in B$ 之间取最小值。

我们给出引理的证明：

令 $a \in A$ 和 $b \in B$ 是任意一对点，并令 $r_1 = \|b - a\|$。由于 A 是紧致的，它被包含在以 a 为中心的一些球中，设该球的半径为 r_2。令 $S = B \cap \overline{B_{r_1+r_2}(a)}$ 为 B 与以 a 为中心、半径为 $r_1 + r_2$ 的闭球的交集。那么 S 是紧致且非空的，因为它包含 b。由

于距离函数是连续的，存在点 a_0 和 b_0 使得 $\|a_0 - b_0\|$ 在所有 $A \times S$ 的点对中取最小值。现在要证明 a_0 和 b_0 实际上在所有 $A \times B$ 的点对中具有最小距离。假设存在点 a' 和 b' 使得 $\|a' - b'\| < \|a_0 - b_0\|$。则特别地，$\|a' - b'\| < r_1$，并且根据三角不等式，$\|a - b'\| \leq \|a - a'\| + \|a' - b'\| < r_1 + r_2$。因此 b' 包含在 S 中，这与 a_0 和 b_0 在 $A \times S$ 中的最小距离相矛盾。



不失一般性地，假设 A 是紧致的。根据引理，存在点 $a_0 \in A$ 和 $b_0 \in B$ 使得它们之间的距离最小。由于 A 和 B 是不相交的，我们有 $a_0 \neq b_0$。现在，构造两条与线段 $[a_0, b_0]$ 垂直的超平面 L_A, L_B，其中 L_A 穿过 a_0，L_B 穿过 b_0。我们声称 A 和 B 都没有进入 L_A, L_B 之间的空间，因此与 (a_0, b_0) 垂直的超平面满足定理的要求。

代数上，超平面 L_A, L_B 由向量 $v := b_0 - a_0$ 定义，并由两个常数 $c_A := \langle v, a_0 \rangle < c_B := \langle v, b_0 \rangle$ 确定，使得 $L_A = \{x : \langle v, x \rangle = c_A\}, L_B = \{x : \langle v, x \rangle = c_B\}$。我们的主张是 $\forall a \in A, \langle v, a \rangle \leq c_A$ 并且 $\forall b \in B, \langle v, b \rangle \geq c_B$。

假设存在某个 $a \in A$ 使得 $\langle v, a \rangle > c_A$，则令 a' 为从 b_0 到线段 $[a_0, a]$ 的垂足。由于 A 是凸集，a' 在 A 内部，并且根据平面几何，a' 比 a_0 更接近 b_0，这与 a_0 和 b_0 的最小距离相矛盾。类似的论证适用于 B。□

1.29 支撑超平面定理

对于一个凸集，支撑超平面（Supporting Hyperplane）是与凸集边界切线的超平面，即它“支撑”了凸集，使得所有的凸集内的点都位于支撑超平面的一侧。形式上，若 S 是非空凸集，且 x_0 是 S 的边界上的一点，那么存在一个包含 x_0 的支撑超平面。 如果 $x^* \in X^* \backslash \{0\}$（$X^*$ 是 X 的对偶空间，x^* 是一个非零的线性泛函），并且对于所有 $x \in S$ 都有 $x^*(x_0) \geq x^*(x)$，那么 $H = \{x \in X : x^*(x) = x^*(x_0)\}$ 定义了一个支撑超平面。

证明

定义 T 为所有支撑闭合半空间的交集，显然 $S \subset T$。现在令 $y \notin S$，证明 $y \notin T$。

设 $x \in \text{int}(S)$，并考虑线段 $[x, y]$。令 t 为最大的数，使得 $[x, t(y - x) + x]$ 被包含在 S 中。则 $t \in (0, 1)$。令 $b = t(y - x) + x$，那么 $b \in \partial S$。在 b 处画一条支撑超平面，令其表示为一个非零线性泛函 $f : \mathbb{R}^n \to \mathbb{R}$，使得 $\forall a \in T, f(a) \geq f(b)$。由于 $x \in \text{int}(S)$，我们有 $f(x) > f(b)$。因此，由 $\frac{f(y) - f(b)}{1 - t} = \frac{f(b) - f(x)}{t - 0} < 0$，我们得到 $f(y) < f(b)$，所以 $y \notin T$。□

第2章：可学性

编辑：赵志民，王茂霖，李一飞，詹好

本章前言

本章的内容围绕学习理论中的可学性理论展开，主要讨论「事件是否能够通过机器学习来解决」这一问题。通过学习理论事先辨别某个问题是否能够被学习，将节省大量的时间与资源。

在讨论学习算法的设计之前，首先要思考以下几个问题：这个问题是否能被解决（从模型的角度看是否可学习），哪些内容容易学习（如两个凸集），哪些内容难学习（如两个非凸集之间的划分），在可学习的情况下，所需的样本量以及通用的学习模型有哪些?

在本章中，我们将通过介绍 "概率近似正确的"（PAC）学习框架，开始正式讨论这些问题。PAC 框架有助于根据实现近似解所需的样本点数量、样本复杂度以及学习算法的时间/空间复杂度（取决于概念的计算表示成本）来定义可学习的概念。

我们首先会描述 PAC 框架并对其进行说明，然后针对所用假设集包含要学习的概念的一致情况和相反的不一致情况，介绍当所用假设集有限时该框架内的一些一般学习保证。

2.1 【概念解释】概念与假设空间

在具体介绍 PAC 模型之前，首先需要明确几个基础定义和符号，这些定义和符号将贯穿本书的大部分内容：

输入空间 X：表示所有可能的例子或实例的集合。X 有时也被称为输入空间。

输出空间 Y：表示所有可能的标签或目标值的集合。Y 有时也被称为输出空间。

在本介绍性章节中，我们将 Y 限制为只有两个标签的情况，即 $Y = \{0, 1\}$（或者 $Y = \{-1, 1\}$，两者仅是符号上的替代）。例如，Y 也可以是 $\{皮卡丘, 海绵宝宝\}$。这是二元分类问题的典型假设。虽然这种简化假设便于理解，但并不会影响后续推论的路径与思路，因为多分类问题只是二分类问题的扩展，虽然从证明和论证上更为复杂。后续章节将扩展这些结果以涵盖更一般的情况。

从数学角度看待机器学习中的概念，机器学习可以定义为学习一个映射函数：

概念(concept) $c : X \to Y$ 是一个从 X 到 Y 的映射。由于 $Y = \{0, 1\}$，我们可以将 c 视为从 X 中得到其取值为 1 的部分，即 X 的子集。

在学习理论中，学习的概念可以等同于从 X 到 $\{0, 1\}$ 的映射，或 X 的子集。概念类是我们希望学习的概念的集合，用 $\mathcal{C}$ 表示。例如，它可以是平面中所有三角形的集合。

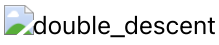
假设空间（hypothesis space）$\mathcal{H}$ 是指所有可能假设的集合，每个假设 $h \in \mathcal{H}$ 是一个从输入空间 X 到输出空间 Y 的映射函数，形式化定义为：

$$\mathcal{H} = \{h : X \to Y\} \tag{175}$$

假设空间的大小和复杂性决定了算法能够学习到的解决方案的类型。如果假设空间太小或太简单，它可能无法捕捉到数据中的复杂模式，导致欠拟合（Underfitting）。相反，如果假设空间过大或太复杂，它可能包含过于复杂的模型，这些模型可能会过度拟合（Overfitting）训练数据，从而在新的、未见过的数据上表现不佳。

例如，在一个简单的线性分类器中，假设空间可能包括所有可能的线性边界，每个线性边界都是一个假设。在更复杂的模型中，如神经网络，假设空间可能包括所有可能的网络结构和权重配置，这些构成了网络的能力来学习数据的非线性和复杂模式。

虽然这种理解适用于机器学习，但我们必须注意，对于深度学习，需要进一步的考虑。例如双下降现象与传统机器学习理论相矛盾，后者认为增加模型大小和数据量通常会提高模型的泛化性能。



双下降现象中描绘模型泛化性能的曲线图由三个阶段组成：

1. 第一阶段：当模型规模小且数据量不足时，模型泛化性能较差。
2. 第二阶段：随着模型规模和数据量的增加，模型泛化性能最初出现下降。
3. 第三阶段：随着模型规模和数据量的进一步增加，模型泛化性能再次下降，但最终达到更好的水平。 双下降现象的出现表明，对于深度神经网络来说，增加模型规模和数据量并不总是有益的。因此，应该采用诸如正则化和增广策略等技术，有效地控制模型规模和数据量，以实现最佳的泛化性能。更多实验细节参考文献：Deep Double Descent: Where Bigger Models and More Data Hurt）。

2.2 【概念解释】经验误差与泛化误差

为了衡量学习到的概念 h 与目标概念 c 之间的差异，定义了以下的度量方式：

泛化误差：

$$R(h) = \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[h(x) \neq c(x)] = \mathop{\mathbb{E}}_{x \sim \mathcal{D}}\left[1_{h(x) \neq c(x)}\right], \tag{1}$$

其中，1_ω 是事件 ω 的指示函数。

由于泛化误差无法直接求得（其原因在于 $\mathcal{D}$ 的未知性），我们需要利用能够获取的信息来近似泛化误差，因此定义了经验误差：

经验误差：

$$\widehat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} 1_{h(x_i) \neq c(x_i)}. \tag{2}$$

经验误差的期望等于其泛化误差：

$$\mathrm{E}[\widehat{R}(h;D)] = R(h;\mathcal{D}) \tag{176}$$

证明过程分为两步，首先考察等式右边，泛化误差可表示为：

$$R(h;\mathcal{D}) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}(h(x) \neq y)] \tag{177}$$

然后考察等式左边，经验误差可表示为：

$$\widehat{R}(h;D) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(h\left(\boldsymbol{x}_i\right) \neq y_i\right) \tag{178}$$

经验误差的期望为：

$$\mathrm{E}[\widehat{R}(h;D)] = \mathop{\mathrm{E}}_{D\sim D^m}[\widehat{R}(h)] = \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathrm{E}}_{(x,y)\sim D}[\mathbb{I}\left(h\left(\boldsymbol{x}_i\right) \neq y_i\right)] \tag{179}$$

由于样本服从独立同分布，所有样本的期望值相同，期望的平均值就等于样本的期望，因此：

$$\mathrm{E}[\widehat{R}(h;D)] = R(h;\mathcal{D}) \tag{180}$$

证毕。

2.3 【概念解释】假设空间的可分性与学习的复杂度

假设空间的可分性决定了学习算法能否有效地找到正确的假设。我们讨论假设空间的可分性与不可分性，并探讨可分性对于学习算法性能的影响。

假设空间：可分性是一个针对假设空间的概念，即考察对于给定学习算法，是否存在能够完全区分所有样本的映射。如果存在，则该学习算法对于此假设空间可分；如果不存在，则不可分。

可分性的严格性指的是其要求所有样本都可分。有时，由于噪声或异常值的影响，数据并非完全可区分，算法只能区分绝大多数样本。因此，可分性并未完全定义学习算法的有效性。

此外，可分性仅表示了学习算法的能力上限。例如，当我们在线性模型中使用高斯核技巧时，能够对任意二分类样本进行区分（维度为无穷）。但从如此庞大的假设空间中找到正确映射函数却非常困难，这在深度学习中尤为明显。在这个意义上，可分性仅表示了学习算法的能力上限。

时间复杂度与样本复杂度

时间复杂度和样本复杂度是评估学习算法效率的两个重要指标。我们讨论这两个概念的等价性，以及它们对学习算法选择的影响。

由于不同的机器、操作系统会带来完全不同的运行时间，因此在考察时间复杂度时通常会使用抽象机。抽象机通常是抽象意义上的图灵机或实体意义上的图灵机。在该抽象机中，时间复杂度被定义为「需要执行的“操作”数量」。

一般而言，学习问题是否可以有效解决，取决于如何将其分解为一系列特定的学习问题。考虑学习有限假设类的问题，例如训练样本的数量为 $m_H(\epsilon,\ \delta) = \log(|\mathcal{H}|/\delta)/\epsilon^2$ 的情况。如果对一个 h 的评估花费固定的时间，那么可以通过对 $\mathcal{H}$ 进行详尽搜索，在时间 $O(|H|m_H(\epsilon,\ \delta))$ 内完成这项任务。对于任何固定的有限假设类 $\mathcal{H}$，穷举搜索算法都可以在多项式时间内运行。如果问题序列 $|H_n| = n$，那么穷举搜索被认为是高效的；如果 $|H_n| = 2^n$，则样本复杂度为 n 的多项式，而穷举搜索算法的计算复杂度随 n 呈指数增长。此时，穷举搜索被认为是低效的。

2.4 【概念解释】PAC-Bayes理论与样本复杂度

PAC学习理论主要研究如何在有限的样本和计算资源下，从给定的假设空间中找到一个近似正确的假设。PAC-Bayes理论结合了PAC学习和贝叶斯方法的优点，其核心思想是通过考虑假设空间中的概率分布来描述学习算法的行为，并给出关于学习算法在有限数据情况下泛化误差的界限。

PAC-Bayes不等式是PAC-Bayes理论的核心结果之一，它为后验分布下的泛化误差提供了一个上界。典型的PAC-Bayes不等式形式如下（详细证明参考：PAC-Bayesian Stochastic Model Selection）：

$$\mathbb{E}_Q[L(h)] \leq \mathbb{E}_Q[\hat{L}(h)] + \sqrt{\frac{KL(Q\|P) + \ln\frac{1}{\delta} + \ln m + \ln 2}{2m-1}} \tag{181}$$

其中：

- $L(h)$ 是假设 h 的真实误差（泛化误差）。
- $\hat{L}(h)$ 是假设 h 在训练集上的经验误差。
- Q 是假设的后验分布。
- P 是假设的先验分布。
- $KL(Q\|P)$ 是后验分布 Q 和先验分布 P 之间的 KL 散度。
- δ 是置信参数，表示上界成立的概率。
- m 是样本数量。

2.5 【定理证明】3项析取范式的不可PAC学习性

32页中有提到，3项析取范式(3-term Disjunctive Normal Form, 3-DNF)概念类并不是高效PAC可学的，除非 $RP = NP$，我们这里给出完整的证明过程。

3项DNF的定义

- **3项DNF公式**: 由三个子句（项）组成，每个子句是布尔变量的合取（AND）。整个公式是这三个子句的析取（OR）。
- **公式的大小**: 由所有子句中的文字（变量或其否定）数量之和决定。对于n个布尔变量，这个大小最多为$6n$。

$RP \neq NP$

在计算复杂性理论中，RP 类包含那些可以通过随机算法在多项式时间内解决的问题，其中算法在给定一个“是”的实例时有很高的概率（至少 $1/2$）返回“是”，而在给定一个“否”的实例时总是返回“否”。NP 类包含那些在多项式时间内可以被验证而不一定是被解决的问题。$RP \neq NP$ 这个表达的意思是假设 RP 类和 NP 类是不相同的。即，存在一些问题在 NP 中，但不在 RP 中。

证明策略

我们通过将一个NP完全问题（在这里选择图的3-着色问题）化简为学习3项DNF公式的问题来进行证明。关键是构造一个样本集 S_G，使得如果图G是3-可着色的，那么存在一个3项DNF公式与S_G一致；反之，如果G不可3-着色，那么不存在这样的公式与S_G一致。

图的3-着色问题

- **图的3-着色**: 给定一个无向图$G = (V, E)$，判断是否可以用三种颜色对顶点进行着色，使得任意一条边的两个端点颜色不同。

构造样本集 S_G

- **正例** S_G^+: 对于每个顶点i，构造向量$v(i)$，该向量在第i位为0，其他位为1，并标记为正例$(v(i), 1)$。
- **反例** S_G^-: 对于每条边(i, j)，构造向量$e(i, j)$，该向量在第i和第j位为0，其他位为1，并标记为反例$(e(i, j), 0)$。

一致性和3-可着色性的等价性

- **一致性**: 如果一个3项DNF公式对样本集S_G中的所有样本都给出正确的分类结果，我们说这个公式与S_G一致。
- **等价性**: 图G是3-可着色的，当且仅当存在一个3项DNF公式与S_G一致。

3项DNF公式的构造

详细说明如何根据图G的3-可着色性构造3项DNF公式，并解释为什么这种构造与样本集S_G一致。

- **颜色划分与子句构造**：
  - 假设图G是3-可着色的，意味着我们可以将所有顶点分成三组，分别着红、蓝、黄三种颜色。
  - 对于每种颜色，我们构造一个合取项。例如，假设T_R表示红色顶点的集合，那么T_R由所有不着红色的顶点的变量的否定组成。
  - 例如，如果顶点j和k不着红色，则$T_R = \neg x_j \wedge \neg x_k$。这里的$\neg x_j$表示顶点$j$没有被着红色。
- **正例与一致性**：
  - 对于每个正例$v(i)$，我们需要这个向量能满足某个子句T_c，即$v(i)$输入到T_c中时，T_c应该为真。
  - 假设顶点i被着成红色，那么$v(i)$中在第i位是0，其他位置是1。此时，$v(i)$会使T_R为真，因为T_R的合取项中的所有文字都与$v(i)$一致——即$v(i)$中对应于非红色顶点的位置都是1，这些位置的$\neg x_j$为真。
- **反例与一致性**：
  - 对于每条边(i,j)的反例$e(i,j)$，我们需要这个向量不能满足整个DNF公式$T_R \vee T_B \vee T_Y$，即$e(i,j)$输入到该公式中时，公式应为假。
  - 假设顶点i着红色，顶点j着蓝色。则T_R由不着红色的顶点变量的否定组成，T_B则由不着蓝色的顶点变量的否定组成。
  - 因为$e(i,j)$在第i和j位都是0，T_R的合取项需要这些位是1才能为真，因此$e(i,j)$不能满足T_R。同理，由于j着蓝色，$e(i,j)$也不能满足T_B，同样它也不能满足T_Y。

因此，对于每个反例$e(i,j)$，公式$T_R \vee T_B \vee T_Y$都不会为真，这就保证了公式与样本集S_G一致。

如果我们可以有效地学习3项DNF公式，那么就可以用它来解决 NP 完全问题（如图的3-着色），这意味着$RP = NP$。由于普遍认为$RP \neq NP$，所以3项DNF类在PAC学习下是不可有效学习的。

第3章：复杂性分析

Edit: 王茂霖，李一飞，詹好，赵志民

本章前言

在机器学习理论中，复杂性分析与计算理论中的算法复杂度类似，是衡量模型和假设空间能力的关键指标。复杂性越高，模型的表达能力越强，但同时也意味着过拟合的风险增加。因此，研究假设空间的复杂性有助于理解模型的泛化能力。

3.1 【概念解释】VC维

VC维（Vapnik-Chervonenkis 维度）是衡量假设空间$\mathcal{H}$复杂性的重要工具。它表示假设空间能够打散的最大样本集的大小，是描述二元分类问题下假设空间复杂度的核心指标。

VC维的定义如下：

$$VC(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\} \tag{182}$$

其中，$\Pi_{\mathcal{H}}(m)$是假设空间$\mathcal{H}$对大小为m的样本集的增长函数。VC维可以理解为模型在二元分类问题中有效的自由度。

例子：对于假设空间$sign(wx + b)$（即线性分类器），其在二维空间R^2中的VC维为3。这意味着，线性分类器能够打散最多三个点，但无法打散四个点。

3.2 【概念解释】Natarajan维

在多分类问题中，我们使用Natarajan维来描述假设空间的复杂性。Natarajan维是能被假设空间$\mathcal{H}$打散的最大样本集的大小。

当类别数$K = 2$时，Natarajan维与VC维相同：

$$VC(\mathcal{H}) = Natarajan(\mathcal{H}) \tag{183}$$

对于更一般的K分类问题，Natarajan维的增长函数上界为：

$$\Pi_{\mathcal{H}}(m) \leqslant m^d K^{2d} \tag{184}$$

随着样本数m和分类数K的增加，Natarajan维的复杂度呈指数级增长。

3.3 【概念解释】Rademacher复杂度

VC维和Natarajan维均未考虑数据分布的影响，而Rademacher复杂度则引入了数据分布因素。它通过考察数据的几何结构和信噪比等特性，提供了更紧的泛化误差界。

函数空间$\mathcal{F}$关于$\mathcal{Z}$在分布$\mathcal{D}$上的Rademacher复杂度定义如下：

$$\Re_{\mathcal{Z}}(\mathcal{F}) = E_{Z \subset \mathcal{Z}:|Z|=m}\left[E_{\sigma}\left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i)\right]\right] \tag{185}$$

其中σ_i是服从均匀分布的随机变量。

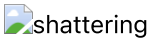
假设空间$\mathcal{H}$的Rademacher复杂度上界为：

$$\Re_m(\mathcal{H}) \leqslant \sqrt{\frac{2 \ln \Pi_{\mathcal{H}}(m)}{m}} \tag{186}$$

3.4 【概念解释】shattering 概念的可视化

Shattering是指假设空间能够实现样本集上所有对分的能力。以下通过二维空间R^2中的线性分类器示例来说明。

示例：对于二维空间R^2中的三个点，线性分类器$sign(wx+b)$可以实现三点的所有对分，但无法实现四点的所有对分，如下图所示：



因此，线性分类器在R^2中的VC维为3。

第4章：泛化界

Edit: 赵志民，李一飞，王茂霖，詹好

本章前言

在机器学习中，泛化能力是衡量模型性能的核心标准之一。如何从有限的训练数据中获得能够在未见数据上表现良好的模型，始终是研究者关注的重要问题。本章将深入探讨与泛化界相关的理论基础和定理，通过对关键概念的补充说明和定理的详细推导，帮助读者更好地理解泛化误差的收敛性质以及不同假设空间下的泛化能力。本章还将介绍与泛化界密切相关的Rademacher复杂度及其在实际应用中的意义，为进一步的研究提供理论支持。

4.1 【概念解释】可分情形中的“等效”假设

61页中的「可分情形」部分提到了“等效假设”的概念。这其实是我们在面对模型选择时需要处理的问题。机器学习的任务实际上是从样本空间或属性空间中选择一个最符合实际的模型假设。在理想状态下，我们希望能排除不可能的情况，直接选择唯一可能的模型。然而，这是不现实的，因为训练数据无法覆盖所有可能的情况，这些数据仅是部分经验片段的记录。因此，机器学习成为了一个不适定问题（ill-posed problem）。

通常而言，不适定问题是指不满足以下任一条件的问题：

1. **存在解**：对于给定的问题，至少存在一个解，即这个问题是可以解决的。
2. **唯一解**：对于给定的问题，解是唯一的，没有其他可能的解。
3. **解连续依赖于定解条件**：解会随着初始条件或参数的变化而连续变化，不会出现突然跳跃或不连续的情况

在这里，由于我们无法仅依靠输入数据找到唯一解，这使得学习问题成为一个不适定问题，主要违反了条件2。而在更多时候，我们说机器学习是不适定的，主要是指其违反了条件3，在那种情况下，我们通常会用正则化等方式来解决。

4.2 【概念解释】定理4.1与定理2.1、定理2.2的关系

61页中的**定理4.1**与**定理2.1**和**定理2.2**之间存在密切联系。

定理2.1指出一个学习算法 $\mathfrak{L}$ 能从假设空间 $\mathcal{H}$ 中PAC辨识概念类 $\mathcal{C}$ ，需要满足：

$$P(\mathbb{E}(h) \leqslant \epsilon) \geqslant 1-\delta \tag{187}$$

其中， $0<\epsilon, \delta<1$，所有 $c \in \mathcal{C}$， $h \in \mathcal{H}$ 。

定理2.2指出，所谓PAC可学，是指对于任何 $m \geqslant \operatorname{poly}(1/\epsilon, 1/\delta, \operatorname{size}(\boldsymbol{x}), \operatorname{size}(c))$ ，学习算法 $\mathfrak{L}$ 能从假设空间 $\mathcal{H}$ 中PAC辨识概念类 $\mathcal{C}$ 。

在**定理4.1**中，假设学习算法 $\mathfrak{L}$ 能从假设空间 $\mathcal{H}$ 中 PAC 辨识概念类 $\mathcal{C}$，且这一过程依赖于大小为 m 的训练集 D ，其中 $m \geqslant \frac{1}{\epsilon}\left(\ln |\mathcal{H}|+\ln \frac{1}{\delta}\right)$，满足

$$m \geqslant \operatorname{poly}(1/\epsilon, 1/\delta, \operatorname{size}(\boldsymbol{x}), \operatorname{size}(c)) \tag{188}$$

的条件，从而得到

$$P(\mathbb{E}(h) \leqslant \epsilon) \geqslant 1-\delta \tag{189}$$

因此，**定理4.1**实际上就是逆向使用了**定理2.1**和**定理2.2**。

4.3 【证明补充】定理4.2补充

63页中，在证明定理4.2时，省略了从式4.6到式4.7的推导过程。在这一过程中，主要用到了**28页**中式2.7的内容。

根据式4.6，有

$$\begin{aligned} & P(\exists h \in \mathcal{H}:|\widehat{E}(h)-\mathbb{E}(h)|>\epsilon) \\ =& P\left(\left(\left|\widehat{E}\left(h_1\right)-E\left(h_1\right)\right|>\epsilon\right) \vee \cdots \vee\left(\left|\widehat{E}\left(h_{|\mathcal{H}|}\right)-E\left(h_{|\mathcal{H}|}\right)\right|>\epsilon\right)\right) \\ \leqslant & \sum_{h \in \mathcal{H}} P(|\widehat{E}(h)-\mathbb{E}(h)|>\epsilon) \end{aligned} \tag{190}$$

引理2.1提出，若训练集 D 包含 m 个从分布 D 上独立同分布采样而得的样本, $0<\epsilon<1$ 则对任意 $h \in \mathcal{H}$, 有

$$\begin{aligned} P(\widehat{E}(h)-\mathbb{E}(h) \geqslant \epsilon) & \leqslant \exp\left(-2 m \epsilon^2\right) \\ P(\mathbb{E}(h)-\widehat{E}(h) \geqslant \epsilon) & \leqslant \exp\left(-2 m \epsilon^2\right) \\ P(|\mathbb{E}(h)-\widehat{E}(h)| \geqslant \epsilon) & \leqslant 2 \exp\left(-2 m \epsilon^2\right) \end{aligned} \tag{191}$$

使用第三个式子，即，

$$P(|\mathbb{E}(h)-\widehat{E}(h)| \geqslant \epsilon) \leqslant 2 \exp\left(-2 m \epsilon^2\right) \tag{192}$$

将其带入式4.6，则有，

$$\sum_{h \in \mathcal{H}} P(|\widehat{E}(h)-\mathbb{E}(h)|>\epsilon) \leqslant \sum_{h \in \mathcal{H}} 2 \exp\left(-2 m \epsilon^2\right) \tag{193}$$

令 $2 \exp\left(-2 m \epsilon^2\right)=\delta /|\mathcal{H}|$，则有，

$$\sum_{h\in\mathcal{H}} P(|\widehat{E}(h)-\mathbb{E}(h)|>\epsilon) \leqslant \sum_{h\in\mathcal{H}} \delta/|\mathcal{H}| \leqslant |\mathcal{H}|\cdot\delta/|\mathcal{H}| = \delta \tag{194}$$

从而得到式4.7。

4.4 【证明补充】引理4.1的证明思路

63页中，引入了引理4.1及其相关的证明。由于证明过程较长，这里对其思路进行梳理和分析。

对于假设空间 $\mathcal{H}, h\in\mathcal{H}, m\in\mathbb{N}, \epsilon\in(0,1)$，当 $m\geq 2/\epsilon^2$ 时有：

$$P(|\mathbb{E}(h)-\hat{E}|>\epsilon) \leq 4\Pi_{\mathcal{H}}(2m)\exp(-\frac{m\epsilon^2}{8}) \tag{195}$$

证明简述

当我们要证明这个定理时，需要首先回忆增长函数的定义：对于 $m\in\mathbb{N}$, 假设空间 $\mathcal{H}$ 的**增长函数** (growth function) $\Pi_{\mathcal{H}}(m)$ 表示为

$$\Pi_{\mathcal{H}}(m) = \max_{\{\mathbf{x}_1,...,\mathbf{x}_m\}\subset\mathcal{X}} |\{(h(\mathbf{x}_1),...,h(\mathbf{x}_m))|h\subset\mathcal{H}\}| \tag{196}$$

由于泛化误差在实际过程中难以评估，证明中首先将泛化误差和经验误差的差距缩放为经验误差之间的差距。通过概率与期望之间的转化，我们将问题进一步转化，并通过上确界的定义给出一个具体的概念 h_0 ，用三角不等式将经验误差与泛化误差之间的差距缩放至经验误差之间。再使用 Chebyshev 不等式中的概率与分布函数积分关系，拆分三角不等式，得出前一半概率（即经验误差与泛化误差之间的差距）与经验误差之间的不等式。

第二步则是将经验误差之间的差距进一步转化为增长函数的差距，即证明了第二个公式：

$$P(\sup_{h\in\mathcal{H}}|\hat{E}_D(h)-\hat{E}_{D'}(h)|\geq\frac{1}{2}\epsilon) \leq 2|\mathcal{H}_{|D+D'|}|\exp(-\frac{\epsilon^2 m}{8})| \tag{197}$$

在这个过程中，使用了式 4.16，通过给出任意置换下的情况，将期望问题转化为级数求和，进一步缩放成有关指数函数的公式：

$$\frac{1}{2m}\sum_{i=1}^{(2m)!}\mathbb{I}(|\hat{E}_{T_iD}(h)-\hat{E}_{T_iD'}(h)\|) = \sum_{k\in[l]s.t.|2k/m-l/m|\geq\epsilon/2}\frac{\binom{l}{k}\binom{2m-l}{m-k}}{\binom{2m}{m}} \tag{198}$$

注意，原不等式中的上界 $2\exp(-\frac{\epsilon^2 l}{8})$ 可以通过 Hoeffding 不等式推导出。

再通过进一步缩放，得到最后的缩放公式（4.19）。此时，结合前述推导可证明引理。

即使将原不等式中的 $2\exp(-\frac{\varepsilon^2 l}{8})$ 替换为 $2\exp(-\frac{\varepsilon^2 l}{4})$，原不等关系依然成立。此结论亦可推广到定理4.3的结论，但即便如此，泛化误差的收敛率依旧为 $O(\sqrt{\frac{ln(m/d)}{m/d}})$。

4.5 【证明补充】定理4.3补充

67页中提到将式（4.24）带入引理4.1，即可证明定理4.3，具体推导如下：

定理4.3 表示为：

$$P(|\mathbb{E}(h)-\widehat{E}(h)|\leqslant\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}})\geqslant 1-\delta \tag{199}$$

可以将其等价转化为：

$$P(|\mathbb{E}(h)-\widehat{E}(h)|>\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}})\leqslant\delta \tag{200}$$

将（4.24）带入引理4.1可得：

$$P(|\mathbb{E}(h)-\widehat{E}(h)|>\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}})\leqslant 4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{m\epsilon^2}{8}\right) \tag{201}$$

根据 3.1 可得：

$$4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{m\epsilon^2}{8}\right)\leqslant 4\left(\frac{2em}{d}\right)^d\exp\left(-\frac{m\epsilon^2}{8}\right) \tag{202}$$

所以引理4.1可以转化为：

$$P(|\mathbb{E}(h)-\widehat{E}(h)|>\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}})\leqslant 4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{m\epsilon^2}{8}\right)\leqslant 4\left(\frac{2em}{d}\right)^d\exp\left(-\frac{m\epsilon^2}{8}\right) \tag{203}$$

令 $4\left(\frac{2em}{d}\right)^d\exp\left(-\frac{m\epsilon^2}{8}\right)=\delta$，由此可得：

$$P(|\mathbb{E}(h)-\widehat{E}(h)|>\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}})\leqslant \delta \tag{204}$$

从而得到了定理4.3的结论。

定理4.3 说明了期望误差和经验误差之间的差异程度，以概率形式限定在一定的区域范围内，虽然这并不完全代表误差一定会在 $\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}}$ 这个范围内，但在此范围内的概率达到了 $1-\delta$。我们可以发现其差异程度的控制范围和样本量及维度之间的关系。当 $\frac{m}{d}$ 较大时（即样本量大，而 VC 维较低），由于 $ln(x)$ 相对于 x 增加较慢，所以其差异可以控制得越小，反之亦然。

4.6 【概念解释】回顾 Rademacher 复杂度

68页谈论了基于 Rademacher 的泛化误差界，这里对 Rademacher 复杂度进行回顾。

由于 VC 维和数据分布无关，未考虑数据的特定分布情况，其得到的结论往往是“松”的。Rademacher 复杂度则是基于数据分布的考虑，在牺牲了一定“普适性”的情况下，得到更为“紧”的结论。

复杂度是人为定义的一套量化复杂度程度的概念。对应 Rademacher 复杂度，假设空间中表示能力越强的函数，其复杂度越高。回到**46-47页**，如果 $\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m\sigma_i h\left(\boldsymbol{x}_i\right)\right]=1$，即对于 x 的任意标签分布情况都能打散（特别注意这里针对的是这个特定的 x 数据，这也是 Rademacher 复杂度和数据分布相关的原因，我们只有知道数据的具体分布情况，才能求解其 Rademacher 复杂度）。由 3.27 可等价得到 3.29 的经验 Rademacher 复杂度。

对于 Rademacher 复杂度的定义，我们进一步将具体的数据样本点转化为数据的样本空间分布，在经验误差的形式外面套一层期望，从而得到了一般化的 Rademacher 复杂度的定义。经验 Rademacher 复杂度和 Rademacher 复杂度的关系就如同概率论中掷硬币的观测序列和将其视为一个先验分布的随机变量序列一样。

4.7 【证明补充】引理4.6的证明解析

71页的定理4.6给出了泛化误差下界的形式化表述：

$$P\left(\mathbb{E}(h_D,c)>\frac{d-1}{32m}\right)\geq\frac{1}{100} \tag{205}$$

虽然不等式右边的常数 $\frac{1}{100}$ 看似有些随意，但作者意在表明：对于任意学习算法，总是存在某种分布和目标概念，使得学习算法输出的假设在较高概率下产生显著错误。

事实上，根据公式（4.50）的推导，只要选择一个小于 $\frac{1-e^{-\frac{d-1}{12}}}{7}$ 的常数，原不等式仍然成立。以 $d=2$ 为例，此时该常数约为 0.0114，因此取 $\frac{1}{100}$ 是较为合理的选择。

进一步分析发现，随着维度 d 的增加，这个常数会逐渐增大，最终逼近 $\frac{1}{7}$。然而，这并不意味着在任何数据分布和目标概念下，泛化误差下界都不会超过 $\frac{1}{7}$。这一限制是由定理证明过程中所假设的数据分布（公式4.42）导致的。

至于常数 32，则是证明过程中产生的结果。通过公式（4.50）的推导，可以看到为了套用公式（4.49）的结论，需要将 ϵ 设为 $\frac{d-1}{16(1+r)}$。在取 $r=1$ 的情况下，分母部分自然得到 32。

4.8 【证明补充】引理4.2补充

74页提出了引理4.2，这里给出完整的证明过程。

令 σ 为服从 $\{-1,+1\}$上均匀分布的随机变量，对于 $0<\alpha<1$构造随机变量 $\alpha_\sigma=1/2-\alpha\sigma/2$，基于 σ 构造 $X\sim D_\sigma$，其中 D_σ 为伯努利分布 $Bernoulli(\alpha_\sigma)$，即 $P(X=1)=\alpha_\sigma$。令 $S=\{X_1,\cdots,X_m\}$ 表示从分布 D_α^m 独立同分布采样得到的大小为 m 的集合，即 $S\sim D_\alpha^m$，这对于函数 $f:X^m\to\{-1,+1\}$ 有：

$$\mathbb{E}_\sigma[P_{S\sim D_\alpha^m}(f(S)\neq\sigma)]\geq\Phi(2\lceil m/2\rceil,\alpha) \tag{206}$$

其中 $\Phi(m,\alpha)=\frac{1}{4}(1-\sqrt{1-\exp(-\frac{m\alpha^2}{1-\alpha^2})})$

证明

我们设想两枚硬币 x_A 和 x_B。两枚硬币都稍有不均匀，即 $P[x_A=0]=1/2-\alpha/2$ 和 $P[x_B=0]=1/2+\alpha/2$，其中 $0<\alpha<1$。0 表示正面，1 表示反面。假设我们随机从口袋里拿出一枚硬币 $x\in\{x_A,x_B\}$，抛 m 次，得到的 0 和 1 的序列即为引理中构造的随机变量 α_σ。如果我们想通过序列推测是哪一枚硬币被抛出，即选取并求得最佳决策函数 $f:\{0,1\}^m\to\{x_A,x_B\}$，则该实验假设的泛化误差可表示为 $error(f)=\mathbb{E}[\mathbb{P}_{S\sim\mathcal{D}_\alpha^m}(f(S)\neq x)]$。

用 f 代表任意决策函数，用 F_A 代表满足 $f(S)=x_A$ 的样本集合，用 F_B 代表满足 $f(S)=x_B$ 的样本集合，用 $N(S)$ 表示样本 S 中出现 0 的个数，根据泛化误差的定义，有：

$$\begin{aligned}
error(f)&=\sum_{S\in F_A}\mathbb{P}[S\wedge x_B]+\sum_{S\in F_B}\mathbb{P}[S\wedge x_A]\\
&=\frac{1}{2}\sum_{S\in F_A}\mathbb{P}[S|x_B]+\frac{1}{2}\sum_{S\in F_B}\mathbb{P}[S|x_A]\\
&=\frac{1}{2}\sum_{\substack{S\in F_A\\N(S)<\lceil m/2\rceil}}\mathbb{P}[S|x_B]+\frac{1}{2}\sum_{\substack{S\in F_A\\N(S)\geq\lceil m/2\rceil}}\mathbb{P}[S|x_B]+\frac{1}{2}\sum_{\substack{S\in F_B\\N(S)<\lceil m/2\rceil}}\mathbb{P}[S|x_A]+\frac{1}{2}\sum_{\substack{S\in F_B\\N(S)\geq\lceil m/2\rceil}}\mathbb{P}[S|x_A]
\end{aligned} \tag{207}$$

如果 $N(S)\geq\lceil m/2\rceil$，易证 $\mathbb{P}[S|x_B]\geq\mathbb{P}[S|x_A]$。类似地，如果 $N(S)<\lceil m/2\rceil$，易证 $\mathbb{P}[S|x_A]\geq\mathbb{P}[S|x_B]$。因此，我们可以得到：

$$\begin{aligned}
error(f)&\geq\frac{1}{2}\sum_{\substack{S\in F_A\\N(S)<\lceil m/2\rceil}}\mathbb{P}[S|x_B]+\frac{1}{2}\sum_{\substack{S\in F_A\\N(S)\geq\lceil m/2\rceil}}\mathbb{P}[S|x_A]+\frac{1}{2}\sum_{\substack{S\in F_B\\N(S)<\lceil m/2\rceil}}\mathbb{P}[S|x_B]+\frac{1}{2}\sum_{\substack{S\in F_B\\N(S)\geq\lceil m/2\rceil}}\mathbb{P}[S|x_A]\\
&=\frac{1}{2}\sum_{S:N(S)<\lceil m/2\rceil}\mathbb{P}[S|x_B]+\frac{1}{2}\sum_{S:N(S)\geq\lceil m/2\rceil}\mathbb{P}[S|x_A]\\
&=error(f_o)
\end{aligned} \tag{208}$$

因此，当我们选取 f_o 为决策函数时，泛化误差取得最小值，即当且仅当 $N(S)<\lceil m/2\rceil$ 时，我们认为被抛的硬币是 $f_o(S)=x_A$。

注意到 $\mathbb{P}[N(S)\geq\lceil m/2\rceil|x=x_A]=\mathbb{P}[B(2\lceil m/2\rceil,p)\geq k]$，且 $p=1/2-\alpha/2,k=\lceil m/2\rceil$，因此 $2\lceil m/2\rceil p\leq k\leq 2\lceil m/2\rceil(1-p)$。

根据 Slud 不等式，我们有：

$$error(f_o) \geq \frac{1}{2}\mathbb{P}[N \geq \frac{\lceil m/2 \rceil \alpha}{\sqrt{1/2(1-\alpha^2)\lceil m/2 \rceil}}] = \frac{1}{2}\mathbb{P}[N \geq \sqrt{\frac{2\lceil m/2 \rceil}{1-\alpha^2}}\alpha] \tag{209}$$

根据第一章补充内容中的正态分布不等式推论，我们有：

$$error(f_o) \geq \frac{1}{4}(1-\sqrt{1-e^{-\frac{2}{\pi}u^2}}) \geq \frac{1}{4}(1-\sqrt{1-e^{-u^2}}) \tag{210}$$

此处 $u = \sqrt{\frac{2\lceil m/2 \rceil}{1-\alpha^2}}\alpha$

事实上，根据上面的推导，我们可以进一步提升泛化误差的下界，即：

$$\mathbb{E}[\mathbb{P}_{S\sim\mathcal{D}_\alpha^m}(f(S) \neq x)] \geq \frac{1}{4}(1-\sqrt{1-e^{-\frac{2}{\pi}u^2}}) \tag{211}$$

在引理末尾处，提到了至少需要 $\Omega(\frac{1}{\alpha^2})$ 次采样才能准确估计 σ_i 的取值，其推理过程如下： 令泛化误差下界至多为 $error(f_o) = \delta > 0$，则有：

$$\frac{1}{4}(1-\sqrt{1-e^{-u^2}}) \leq \delta \Leftrightarrow m \geq 2\lceil \frac{1-\epsilon^2}{2\epsilon^2}\ln\frac{1}{8\delta(1-2\delta)} \rceil \tag{212}$$

此时，我们发现 m 至少为 $\Omega(\frac{1}{\alpha^2})$ 时，才能以 $1-\delta$ 的概率确定 σ 的取值。

4.9 【证明补充】引理4.7的补充

75页的定理4.7主要表达的是：无论算法有多强，在不可分的情况下，总会有某种"坏"分布使得输出假设的泛化误差以常数概率为 $O(\sqrt{\frac{d}{m}})$。其中（4.61）中第二步变形用到了以下等式：

$$\sum_{x_i \in S}(\mathbb{I}(h(x_i) \neq h_{\mathcal{D}_\sigma^*}(x_i)) + \mathbb{I}(h(x_i) = h_{\mathcal{D}_\sigma^*}(x_i))) = d \tag{213}$$

另外，（4.63）的第三步为何不直接利用引理4.2进行推导呢？这是考虑到函数$\Phi(\cdot,\alpha)$为减函数，即由$m/d+1 \leq 2\lceil m/2 \rceil$可知$\Phi(m/d+1,\alpha) \geq \Phi(2\lceil m/2 \rceil,\alpha)$。可见后者并不是一个特别紧致的下界，因此我们转而考虑按照$|Z|_x$的取值进行拆分。

在**76页**左下角的最后一个脚注中，提到了m/d为变量$|Z|_x$的期望值，如何得到这个结论呢？根据（4.58）和（4.59）以及$\mathcal{U}$为$\{-1,+1\}^d$均匀分布的性质，我们可以得到从分布中抽取给定点x的期望概率为$1/d$。 当我们从 D_σ 中独立抽取 m 个样本的情况下，S 中点 x 出现的次数的期望值为 m/d。

此外，（4.65）中用到了引理4.3。令 $Z' = \frac{1}{\alpha}(\mathbb{E}(h_Z) - \mathbb{E}(h^*_{\mathcal{D}^m_{\sigma^*}}))$，根据（4.62）可知 $0 \leq Z' \leq 1$。令 $\gamma' = \gamma u$，因为 $\Phi(\cdot,\alpha)$为减函数，易知其最大值为$1/4$，因此有$\gamma' \in [0,1/4) \subseteq [0,1)$。此时带入引理4.3可得：

$$P(Z' > \gamma') \geq \mathbb{E}[Z'] - \gamma' \geq u - u\gamma = (1-\gamma)u \tag{214}$$

同时，（4.69）到（4.70）的推导中体现了充分条件的思想。由（4.69）可知：

$$\frac{m}{d} \leq \frac{A}{\epsilon^2} + B \tag{215}$$

其中 $A = (\frac{7}{64})^2\ln\frac{4}{3}$，$B = -\ln\frac{4}{3} - 1$。

我们希望能推导出更为简洁的 $\frac{m}{d}$ 与 $\frac{1}{\epsilon^2}$ 之间的关系，因此考虑寻找充分条件使以下不等式成立：

$$\frac{m}{d} \leq \frac{A}{\epsilon^2} + B \leq \frac{\omega}{\epsilon^2} \tag{216}$$

即使得 $\omega \geq B\epsilon^2 + A$ 成立。当 $\epsilon \leq 1/64$ 时，很容易得到 ω 的最小值（4.70）。

值得注意的是，整个证明过程共进行了四次启发式限制，分别为 $\gamma = 1-8\delta$，$\alpha = 8\epsilon/(1-8\epsilon)$，$\delta \leq 1/64$ 和 $\epsilon \leq 1/64$。这些启发式限制构造出来都是为了使得最终的不等式成立，实际上我们亦可根据实际需要进行调整，继而得到该定理的不同变种。

4.10 【概念解释】ρ-间隔损失函数的 Lipschitz 性

79页提到，由经验损失（公式4.72）可知 Φ_ρ 最多是 $\frac{1}{\rho}$-Lipschitz。对此进行详细解读如下：

根据Lipschitz连续性的定义，我们可以通过拉格朗日中值定理来证明这一点。具体来说，由拉格朗日中值定理可得：

$$|\Phi_\rho(x_1) - \Phi_\rho(x_2)| \leq |\Phi'_\rho(\xi)||x_1 - x_2| \tag{217}$$

其中 ξ 是 x_1 和 x_2 之间的某一点。

已知 Φ_ρ 的具体表达式，因此可以直接计算其导数 $\Phi'_\rho(\xi)$。通过计算，我们可以得到：

$$|\Phi'_\rho(\xi)| \leq \frac{1}{\rho} \tag{218}$$

因此，根据Lipschitz条件的定义，ρ-间隔损失函数是 $\frac{1}{\rho}$-Lipschitz 函数。

4.11 【证明补充】定理4.8补充

79页的定理4.8给出了关于间隔损失函数的分类问题SVM的泛化误差界。

此处存在一个小的错误：公式4.80前的 **“代入 (4.96)”** 应为 **“代入 (4.76)”**。

观察要证明的公式，我们发现这是关于 Rademacher 复杂度的泛化上界推理，自然地回顾一下 Rademacher 复杂度。

现实任务中样本标记有时会受到噪声影响，因此我们与其在假设空间 $\mathcal{H}$ 中选择训练集上表现最好的假设，不如选择 $\mathcal{H}$ 中事先已经考虑了随机噪声影响的假设。

在此直接考虑利用前面讲到的关于实值假设空间中的期望与 Rademacher 复杂度的不等式。通过前面 4.73 讲到的关于间隔函数的经验间隔损失的式子，可以带入得到大体形式。

由于前面引理提到的关于 Lipschitz 函数的性质，结合 ρ-间隔损失函数的 Lipschitz 性，在简单改写复杂度之后便能得到要证明的定理。

第5章：稳定性

编辑：赵志民，李一飞，王茂霖，詹好

本章前言

本章将探讨学习理论中的稳定性。在前一章中，我们介绍了不同的复杂度度量方法，并给出了与特定算法无关的泛化界限。 然而，这些泛化界限是否能通过分析特定算法的性质得到更好的学习保障？这些分析是否能够扩展到具有相似性质的其他学习算法上？ 本章旨在回答这些问题，通过算法稳定性的应用推导出依赖于算法的学习保证。

5.1 【概念解释】留一交叉验证的风险

90页中提到的留一风险（leave-one-out risk）是指依次从数据集中移除某一数据后，利用剩余数据训练的模型与被移除数据之间的风险。本质上，这保证了用于风险测试的数据不会包含在训练集中，类似于模型选择时的留一验证。

5.2 【证明补充】均匀稳定性与泛化误差上界

92页中，定理5.1讨论了**均匀稳定性**与泛化性的关系。以下是该证明过程中均匀稳定性与泛化性之间联系的分析。

证明简述

对于读者来说，前几章的阅读应使大家对涉及 ln 和根号的不等式已经有所了解，并能意识到这与指数函数的不等式有关，并反解风险 ϵ。这里我们希望通过样本的稳定性推导出关于风险的泛化性。因此，在证明时必须将风险之间的差距转化为损失函数之间的风险。

由于定理中提到的替换样本 β-均匀稳定性和移除样本 γ-均匀稳定性是非常强的条件，适用于任意的数据集 D 和任意的样本 **z**，因此我们可以得到关于*经验风险与泛化风险差距*（即 $\Phi(D)$） 的估计式。

通过对损失函数的差值求和平均可以得到风险 (Risk) 的差距。由于替换样本的 β-均匀稳定性适用于任意 **z**，因此我们可以推导出 (5.22) 和 (5.23) 式，并使用 McDiarmid 不等式推导出*经验风险与泛化风险的差距*（即 $\Phi(D)$） 超过其平均值至少 ϵ 的概率。即：

$$P(\phi(D) \geq \mathbb{E}[\Phi(D)] + \epsilon) \leq exp(\frac{-2m\epsilon^2}{(2m\beta + M)^2}) \tag{219}$$

之后进行简单的放缩估计即可得到最终的结果：

$$P(R(\mathcal{L}_{\mathcal{D}}) - \hat{R}(\mathcal{L}_{\mathcal{D}}) \geq \beta + \epsilon) \leq exp(\frac{-2m\epsilon^2}{(2m\beta + M)^2}) \tag{220}$$

值得注意的是，（5.22）中的最后一步不等式推导其实省略了一步：

$$\begin{aligned} &\frac{|\ell(\mathfrak{L}_D, z_i) - \ell(\mathfrak{L}_{D^{i,z_i'}}, z_i')|}{m} + \sum_{j \neq i} \frac{|\ell(\mathfrak{L}_D, z_j) - \ell(\mathfrak{L}_{D^{i,z_i'}}, z_j)|}{m} \\ &\leq \frac{M}{m} + \frac{m-1}{m}\beta \\ &\leq \frac{M}{m} + \beta \end{aligned} \tag{221}$$

之所以这么做，是因为当样本量 m 较大时，$\frac{\beta}{m}$ 的大小可以忽略不计，因此在结论中并未出现这一项。

另外，（5.23）式也省略了一步：

$$|E_{z\sim\mathcal{D}}[\ell(\mathfrak{L}_D, z) - \ell(\mathfrak{L}_{D^{i,z_i'}}, z)]| \leq E_{z\sim\mathcal{D}}[|\ell(\mathfrak{L}_D, z) - \ell(\mathfrak{L}_{D^{i,z_i'}}, z)|] \leq E_{z\sim\mathcal{D}}[\beta] = \beta \tag{222}$$

关于移除样本 γ-均匀稳定性（5.18）的证明用到了（5.14）的结论，因此在不等式中构造出了类似于 $2m\beta$ 的 $4m\gamma$ 形式，其他推理步骤与（5.17）基本一致。

均匀稳定性与泛化性的关系

在证明过程中，多处涉及了损失函数作差的放缩，即替换样本的 β-均匀稳定性，但实际上大多数情况下使用该稳定性只是为了简化式子，只有在 (5.24) 与 (5.25) 中体现了稳定性与泛化性的关系。

在 (5.24) 中，通过替换样本的稳定性，我们可以得到*经验风险与泛化风险的差距*（即 $\Phi(D)$) 在替换样本前后的风险可以被上界 $2\beta + M/m$ 控制住。根据 McDiarmid 不等式的描述，如果实值函数关于变量的替换具有较好的稳定性，那么该实值函数与期望的差距也将受到上界控制。简言之，如果实值函数替换一个变量后变化不大，那么无论如何替换，变化都不会过大，因此该实值函数的取值总会在一定范围内，与其均值（即期望）相差不大。

因此在 (5.25) 中，我们可以得到*经验风险与泛化风险的差距*（即 $\Phi(D)$) 也有了上界。通过简单的放缩可以得到一个常数上界，从而得出泛化风险的上界。

5.3 【证明补充】假设稳定性与泛化误差上界

94页中，定理5.2讨论了**假设稳定性**与泛化性的关系。以下是该证明过程中假设稳定性与泛化性之间联系的分析。

证明简述

证明涉及 $R(\mathcal{L}_{\mathcal{D}}) - \hat{R}(\mathcal{L}_{\mathcal{D}})$ 的平方平均，因为假设稳定性是较弱的条件，只能保证风险的期望被上界控制，因此只能得到关于期望的不等式。由于不涉及概率与置信度，因此不需要复杂的不等式，简单的放缩即可得到答案。

证明中的一处难点在于（5.30）至（5.33）中关于变量 z 之间的替换。根据独立同分布假设，即 $\forall i, j \in \mathbb{N}^+, z, z', z_i, z_j \sim \mathcal{D}$，可以任意交换 z, z', z_i, z_j 的顺序而期望值不变。

例如，在（5.30）中的第一步推导中，不失一般性地用 z_1, z_2 替代 z_i, z_j，因此原期望值之和得以简化为只与 z_1, z_2 相关的期望值。

理解这一点后，任何关于变量 z 之间的替换都不会令人感到困惑，其中也包括了定理5.3证明中（5.35）的第二步推导。

另外，在（5.32）的第一步推导中，使用了绝对值不等式 $\mathbb{E}(X + Y) \leq \mathbb{E}(|X|) + \mathbb{E}(|Y|)$。这种在期望放缩中运用绝对值不等式的处理方式在全书中非常实用，值得读者留意。

假设稳定性与泛化性的关系

该定理实际上给出了经验风险与泛化风险的差距的平方平均的界，这是因为假设稳定性并不是非常强的条件，而是为了放松均匀稳定性这一较强的条件而引入的。

5.4 【概念解释】过拟合与欠拟合的关系

过拟合和欠拟合是泛化性研究中的重要概念。当经验风险与泛化风险的差距较大时，会发生过拟合；相反，当泛化风险与经验风险的差距较大时，则发生欠拟合。因此，我们在算法设计时，希望尽可能缩小泛化风险与经验风险的差距。

96页中，定理5.3从算法稳定性的角度提出了防止过拟合的方案：当替换训练集的单个样本时，算法的输出函数变化不大，我们认为学习算法 $\mathfrak{L}$ 是稳定的，否则就需要重新进行训练。该方法同样适用于欠拟合的情况，但在实际应用中，算法欠拟合的情况较少，因此我们更多地关注过拟合的预防。

5.5 【概念解释】稳定性与可学习性

97页中，定理5.4讨论了稳定性与可学性之间的关系。以下是定理5.4的梳理分析，探讨稳定性与可学性在证明中的关联。

证明简述

首先，我们回顾不可知 PAC 可学的概念：对于所有分布 $\mathcal{D}$，若存在学习算法 $\mathfrak{L}$ 与多项式函数 $poly(\cdot, \cdot, \cdot, \cdot)$，使得对于任意 $m \geq poly(1/\epsilon, 1/\delta, size(\mathbf{x}), size(c))$，$\mathfrak{L}$ 输出的假设能够满足：

$$P\big(\mathbb{E}(h) - \min_{h' \in \mathcal{H}} \mathbb{E}(h') \leq \epsilon\big) \geq 1 - \delta \tag{223}$$

该证明利用了经验风险与泛化风险之间的联系，构造出（5.39），然后分而治之地讨论不同情况下的稳定性关系。

其中，泛化风险与经验风险之差（5.40）可以根据定理5.1改写为：对于任意的 $\delta \in (0, 1)$，以至少 $1 - \delta$ 的概率有：

$$R(\mathfrak{L}_D) - \hat{R}(\mathfrak{L}_D) \leq \frac{1}{m} + (2m\beta + M)\sqrt{\frac{ln(1/\delta)}{2m}} \tag{224}$$

参考**95页**中对 $lim_{m \to +\infty} m\beta$ 的讨论，发现只要满足 $lim_{m \to +\infty} m\beta < \infty$ 的条件，算法的泛化性能便可得到保障，因此应确保 β 的取值不要太大。在此定理中，我们选取 $\beta = 1/m$，此时（5.40）简化为：

$$R(\mathfrak{L}_D) - \hat{R}(\mathfrak{L}_D) \leq \frac{1}{m} + (2 + M)\sqrt{\frac{ln(1/\delta)}{2m}} \tag{225}$$

在处理 ERM 算法情况下泛化风险与经验风险之差（5.42）时，原书中有一处小错误，但对最终结论影响不大。以下是正确的推导过程：

根据 $\ell(\mathfrak{L}_D, z) \in [0, M]$，可以得到 $\hat{R} \in [0, M]$，又因为 $R(h^*) = E_{\mathcal{D}}(\hat{R}(h^*))$，此时根据 Hoeffding 不等式（1.30），可知至少以 $1 - \delta$ 的概率有：

$$\hat{R}(h^*) - R(h^*) \leq M\sqrt{\frac{ln(1/\delta)}{2m}} \tag{226}$$

结合（5.39）至（5.42）可知，至少以 $1-\delta$ 的概率有：

$$R(\mathfrak{L}_D) - R(h^*) \leq \frac{1}{m} + (2+M)\sqrt{\frac{ln(1/\delta)}{2m}} + M\sqrt{\frac{ln(1/\delta)}{2m}} \tag{227}$$

此时（5.44）变为：

$$\epsilon = \frac{1}{m} + (1+M)\sqrt{\frac{2ln(1/\delta)}{m}} \tag{228}$$

令 $m' = \sqrt{m}$，则可以将上式转化为关于 m' 的一元二次方程：

$$\epsilon m'^2 - Am' - 1 = 0 \tag{229}$$

其中 $A = (1+M)\sqrt{2ln(1/\delta)}$，根据求根公式可得：

$$m' = \frac{A \pm \sqrt{A^2 + 4\epsilon}}{2\epsilon} = O(\frac{1}{\epsilon}\sqrt{ln(\frac{1}{\delta})}) \tag{230}$$

至此，我们得到了 m 的渐近复杂度：

$$m = m'^2 = O(\frac{1}{\epsilon^2}ln(\frac{1}{\delta})) \tag{231}$$

接下来的推导便水到渠成。

稳定性与可学性

这里只能达到不可知 PAC 可学的原因是泛化界只能以概率达到，无法保证在任何函数空间内都能达到上界以下。因此，这里只能讨论稳定性与不可知 PAC 可学性的关系。

事实上，稳定性与可学性的关系类似于第四章中讲到的泛化界与可学性的关系。通过 ERM 算法得到最小经验风险函数后，结合均匀稳定性带来的**泛化上界**，我们可以获得可学性。

5.6 【证明补充】二次分布下的 k-近邻算法稳定性

105页中，引理5.2讨论了二次分布 $X \sim B(k, 1/2)$ 的 k-近邻的稳定性。这里我们给出详细的证明过程。

给定整数 $k > 0$，若随机变量 X 满足：

$$P(X = i) = \frac{1}{2^k}\binom{k}{i}, i \in [k] \tag{232}$$

则对任意正整数 a 有：

$$P(|X - \frac{k}{2}| \leq \frac{a}{2}) < \frac{2\sqrt{2}a}{\sqrt{\pi k}} \tag{233}$$

首先，我们根据 k 的取值将情况分为两类讨论。

当 k 为偶数时，二项式展开的最大项为：

$$\frac{1}{2^k}\binom{k}{k/2} \leq \frac{2}{\sqrt{2\pi k}}\exp(\frac{1}{12k} - \frac{2}{6k+1}) < \frac{2}{\sqrt{2\pi k}} \tag{234}$$

第二步推导利用了 Stirling 公式，最后一步推导则利用了函数 $l(x) = \exp(\frac{1}{12x} - \frac{2}{6x+1})$ 在 $[1, \infty)$ 区间单调递增且取值在 $(0, 1)$ 之间的特性。

因此，我们有：

$$P(|X-\frac{k}{2}|\leq\frac{a}{2})=(a+1)\frac{2}{\sqrt{2\pi k}}<\frac{4a}{\sqrt{2\pi k}} \tag{235}$$

当 k 为奇数且 $k>1$ 时，二项式展开的最大项为：

$$\frac{1}{2^k}\binom{k}{(k-1)/2}<\frac{1}{2^{k-1}}\binom{k-1}{(k-1)/2}<\frac{1}{\sqrt{2\pi(k-1)}}<\frac{2}{\sqrt{\pi k}} \tag{236}$$

当 $k=1$ 时，二项式展开的最大项为 $\frac{1}{2}<\frac{2}{\sqrt{\pi}}$ 因此，我们有：

$$P(|X-\frac{k}{2}|\leq\frac{a}{2})=a\frac{2}{\sqrt{\pi k}}<\frac{4a}{\sqrt{2\pi k}} \tag{237}$$

综上，引理5.2得证。

5.7 【概念解释】稳定性理论的适用范围

细心的读者可能已经注意到，这里的稳定性仅在某些条件下才能适用，以下是对这些条件的总结。

首先，本章的分析假设输出函数 $\mathfrak{L}_D$ 与训练集 D 的顺序无关，但这在实际应用中并不一定成立。例如，在随机梯度下降算法中，训练集的顺序会影响最终的输出函数，因此这里的稳定性并不适用于随机梯度下降算法。

另外，在样本扰动分析中，我们几乎没有单独讨论新增样本的情况。这是因为在数据或概念发生漂移的情况下，稳定性的要求不一定成立，因为此时训练集的分布与真实分布已不再一致。而在研究训练集 D 的扰动对算法 $\mathfrak{L}_D$ 输出函数的影响时，我们希望经验风险的变化尽可能小，这恰好与在线学习（Online Learning）的目标相抵触。

具体而言，在线学习指的是在数据不断到来的过程中，动态地更新模型，因此该训练方式更关注模型的可塑性，即在旧场景中训练的模型是否能通过优化在新场景中表现优异。因此，在实际应用中，我们需要平衡学习算法的可塑性与稳定性。

为了更好地评估在线学习的性能，本书引入了遗憾界的概念，即在线学习与离线学习算法之间最小损失的差值，具体分析请参见第八章。

第6章：一致性

编辑：赵志民，王茂霖，詹好

本章前言

本章内容主要探讨学习理论中的一致性（consistency），研究随着训练数据的增加，通过学习算法所获得的分类器是否逐渐逼近贝叶斯最优分类器。具体内容包括一致性的定义、参数方法下的一致性分析、非参数方法下的一致性分析，以及随机森林一致性分析的案例。

6.1 【证明补充】泛化风险的无偏估计

117页中，公式（6.25）给出了分类器的经验风险 $\hat{R}$，并指出其为泛化风险 R 的无偏估计。以下对这一概念进行详细说明。

首先，需要理解经验风险 $\hat{R}$ 和泛化风险 R 的概念。经验风险是基于模型的预测结果与真实结果的比较计算出的量化风险指标。泛化风险则是基于数据-标签联合分布的样本（视为随机变量）的预测结果与真实值的比较的期望值。由于实际情况下数据-标签联合分布通常未知，泛化风险 R 更多是一个理论化的概念。

其次，当我们说 y 是 x 的无偏估计时，意味着 $\mathbb{E}[x]=y$。根据这一概念，我们可以证明经验风险是泛化风险的无偏估计。

泛化风险定义为：

$$
\begin{aligned}
R(f) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}(yf(x)\leqslant 0)] \\
&= \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\eta(x)\mathbb{I}(f(x)\leqslant 0)+(1-\eta(x))\mathbb{I}(f(x)\geqslant 0)]
\end{aligned} \tag{238}
$$

经验风险定义为：

$$
\hat{R}(f)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(y_i f\left(x_i\right)\leqslant 0\right) \tag{239}
$$

现在我们证明经验风险是泛化风险的无偏估计：

假设所有样本都是从一个未知的样本-标签空间 D 中独立同分布采样的，对经验风险求期望：

$$
\begin{aligned}
\mathbb{E}(\hat{R}(f)) &= \mathbb{E}_{(\boldsymbol{x_i},y_i)\sim\mathcal{D}}[\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(y_i f\left(x_i\right)\leqslant 0\right)] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{(\boldsymbol{x_i},y_i)\sim\mathcal{D}}[\mathbb{I}\left(y_i f\left(x_i\right)\leqslant 0\right)] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}\left(yf\left(x\right)\leqslant 0\right)] \\
&= \frac{1}{m}\sum_{i=1}^{m}R(f) \\
&= R(f)
\end{aligned} \tag{240}
$$

6.2 【证明补充】替代函数一致性

120页的定理6.1给出了替代一致性的充分条件。首先，我们推导了泛化风险与贝叶斯风险之间的差异不等式。根据一致性的定义，我们需要证明，当 $R_\phi\left(\hat{f}_m\right)\to R_\phi^*$ 时，$R\left(\hat{f}_m\right)\to R^*$。

为此，我们进一步构造了关于 $R_\phi\left(\hat{f}_m\right)-R_\phi^*$ 的不等式。通过分析两个不等式之间的关联性，最终得出结论：

$$
R\left(\hat{f}_m\right)-R^*\leqslant 2c\sqrt[s]{R_\phi\left(\hat{f}_m\right)-R_\phi^*} \tag{241}
$$

因此，当 $R_\phi\left(\hat{f}_m\right)\to R_\phi^*$ 时，$R\left(\hat{f}_m\right)$ 也会收敛于 R^*。

其中，不等式（6.40）的推导涉及一定的构造技巧，接着通过定理中的条件推导出不等式（6.43）。利用所构造的凸函数的性质，最终完成了这一证明。

6.3 【概念解释】划分机制方法

122页介绍了一种将样本空间划分成多个互不相容区域的方法，然后对各区域内的正例和反例分别计数，并以多数类别作为区域中样本的标记。这种方法本质上不同于参数方法，它并不是在参数空间中进行搜索构建划分超平面，而是在泛函空间上直接进行搜索。一个典型的例子是我们熟悉的决策树模型：

每当构造一个决策树的节点时，相当于在样本空间上进行了一次划分（即划分机制）。这种洞察方式同样适用于解释剪枝操作，即通过减少不必要的节点来简化树结构，同时保持或提高模型的性能。

6.4 【概念解释】依概率成立

124页的定理6.2提到一个定义——依概率成立（almost sure）。这是概率论与数理统计中的一个概念，表达如下：

$$
\lim_{n\to\infty}P((Diam(\Omega)-0)\geq\epsilon)=0 \tag{242}
$$

和对于所有 $N>0$：

$$\lim_{n\to\infty} P(N(x)>N)=1 \tag{243}$$

它意味着当 n 趋于无穷时，几乎处处（almost everywhere）的 $Diam(\Omega)$ 都处于 0 的 ϵ 邻域内。而 $N(x)$ 的极限几乎处处为无穷大。依概率成立是一种比极限更弱的情况，即可以忽略概率趋于 0 的情形。

6.5 【证明补充】划分机制一致性

124页的定理6.2给出了划分一致性的充分条件。首先我们定义了 $\Omega(x)$ 作为划分区域的条件概率极大似然估计量：

$$\hat{\eta}(x)=\sum_{x_i\in\Omega(x)}\frac{\mathbb{I}\left(y_i=+1\right)}{N(x)} \tag{244}$$

再根据划分机制构造分类器（输出函数）$h_m(x)=2\mathbb{I}\left(\hat{\eta}(x)\geqslant\frac{1}{2}\right)-1$。为了证明划分机制的一致性，我们需要证明其输出函数的泛化风险在 m 趋于无穷时，趋于贝叶斯风险。

在此，我们利用了基于条件概率估计的插值法，并借助引理6.2得到了输出函数的泛化风险与贝叶斯风险之间的差值不等式。对于不等式右侧的期望，利用三角不等式进行放缩，可得（6.62）。

根据假设条件：

$$\lim_{m\to\infty}P((Diam(\Omega)-0)\geq\epsilon)=\lim_{m\to\infty}P((\sup_{x,x'\in\Omega}\|x-x'\|-0)\geq\epsilon)=0 \tag{245}$$

由于 $\eta(x)$ 在样本空间中具有连续性，因此在任意邻域内我们都可以用 $\hat{\eta}(x)$ 的期望值来近似 $\eta(x)$。当邻域趋于 0 时，可得：

$$\mathbb{E}[|\bar{\eta}(x)-\eta(x)|]\to 0 \tag{246}$$

这是由于 x' 被依概率限制在一个 ϵ 邻域内，且期望可以忽略概率趋于 0 的点，因此 $\bar{\eta}(x)$ 由于 $\eta(x)$ 的连续性也被限制在一个 $\eta(x)$ 的 ϵ 邻域内，从而期望的极限得证。

接下来，针对三角不等式右式的前半部分，将其拆分为 $N(x)=0$ 和 $N(x)>0$ 两部分：

$$\begin{aligned}
&\mathbb{E}\left[|\hat{\eta}(x)-\bar{\eta}(x)|\mid x,x_1,\ldots,x_m\right]=\mathbb{E}\left[|\hat{\eta}(x)-\bar{\eta}(x)|\mid N(x)=0,x,x_1,\ldots,x_m\right]\\
&\qquad+\mathbb{E}\left[\left|\textstyle\sum_{x_i\in\Omega(x)}\frac{\mathbb{I}(y_i=+1)-\bar{\eta}(x)}{N(x)}\right|N(x)>0,x,x_1,\ldots,x_m\right]\\
&\leqslant P\left(N(x)=0\mid x,x_1,\ldots,x_m\right)+\mathbb{E}\left[\left|\textstyle\sum_{x_i\in\Omega(x)}\frac{\mathbb{I}(y_i=+1)-\bar{\eta}(x)}{N(x)}\right|N(x)>0,x,x_1,\ldots,x_m\right]
\end{aligned} \tag{247}$$

然后，对于不等式右侧的第二部分，利用引理6.3的不等式，可以得到：

$$\begin{aligned}
&\mathbb{E}\left[\left|\textstyle\sum_{x_i\in\Omega(x)}\frac{\mathbb{I}(y_i=+1)-\bar{\eta}(x)}{N(x)}\right|N(x)>0,x,x_1,\ldots,x_m\right]\\
&\leqslant\mathbb{E}\left[\sqrt{\tfrac{\bar{\eta}(x)(1-\bar{\eta}(x))}{N(x)}}\mathbb{I}(N(x)>0)\mid x,x_1,\ldots,x_m\right]
\end{aligned} \tag{248}$$

对于此不等式的右侧，再进行放缩。对于任意 $k\geq 3$，当 $N(x)\leqslant k$ 时，$\sqrt{\frac{\bar{\eta}(x)(1-\bar{\eta}(x))}{N(x)}}\leqslant\frac{1}{2}$，当 $N(x)>k$ 时，$\sqrt{\frac{\bar{\eta}(x)(1-\bar{\eta}(x))}{N(x)}}\leqslant\frac{1}{2\sqrt{k}}$，从而得到不等式右侧的进一步放缩：

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\frac{\bar{\eta}(x)(1-\bar{\eta}(x))}{N(x)}}\mathbb{I}(N(x)>0)\mid x,x_1,\ldots,x_m\right]&\leqslant\frac{1}{2}P\left(N(x)\leqslant k\mid x,x_1,\ldots,x_m\right)+\frac{1}{2\sqrt{k}}P\left(N(x)>k\mid x,x_1,\ldots,x_m\right)\\
&\leqslant\frac{1}{2}P\left(N(x)\leqslant k\mid x,x_1,\ldots,x_m\right)+\frac{1}{2\sqrt{k}}
\end{aligned} \tag{249}$$

结合前面的结果，我们可以得出：

$$\mathbb{E}[|\hat{\eta}(x)-\bar{\eta}(x)|]\leqslant\frac{1}{2}P(N(x)\leqslant k)+\frac{1}{2\sqrt{k}}+P(N(x)=0) \tag{250}$$

根据 $N(x) \to \infty$ 依概率成立，当 $m \to \infty$ 时，$P(N(x) \leqslant k) \to 0$，$P(N(x)=0) \to 0$。并且当取 $k=\sqrt{N(x)}$ 时，$\frac{1}{2\sqrt{k}} \to 0$ 依概率成立，从而得出结论：

$$\mathbb{E}[|\hat{\eta}(x)-\bar{\eta}(x)|] \to 0 \tag{251}$$

最终证明了其输出函数的泛化风险在 m 趋于无穷时，趋于贝叶斯风险：

$$R\left(h_m\right)-R^* \leqslant 2\mathbb{E}[|\hat{\eta}(x)-\eta(x)|] \to 0 \tag{252}$$

6.6 【证明补充】随机森林的划分一致性

130页中的定理6.5提到了一种简化版本的随机森林，即每次划分都是均匀随机的，并不依赖于训练集的标签。以下对证明直径 $Diam(\Omega(x,Z)) \to 0$ 的步骤进行补充说明。

首先，令 L_j 表示区域 $\Omega(x,Z)$ 中第 j 个属性的边长，我们可以得到 $Diam(\Omega(x,Z))$ 与 L_j 的关系：

$$Diam(\Omega(x,Z)) = sup_{x,x' \in \Omega}\|x-x'\| = \sqrt{\sum_{j=1}^{d} L_j^2} \tag{253}$$

对于 $Diam(\Omega(x,Z))$ 求期望时，我们得到：

$$\mathbb{E}(Diam(\Omega(x,Z))) = \mathbb{E}(\sqrt{\sum_{j=1}^{d} L_j^2}) \tag{254}$$

令 $L=\sum_{j=1}^{d} L_j^2$，因为 $\sqrt{L}$ 是关于 L 的凸函数，根据 Jensen 不等式（1.11），我们可以得到：

$$\mathbb{E}(\sqrt{\sum_{j=1}^{d} L_j^2}) \leq \sqrt{\sum_{j=1}^{d} \mathbb{E}(L_j^2)} \tag{255}$$

由于每个属性的边长 L_j 在随机决策树构造中都是独立同分布的，因此可以得到：

$$\sqrt{\sum_{j=1}^{d} \mathbb{E}(L_j^2)} = \sqrt{d\mathbb{E}(L_1^2)} = \sqrt{d}\mathbb{E}(L_1) \tag{256}$$

综合以上各式，我们只需证明当 $k \to \infty$ 时有 $\mathbb{E}(L_1) \to 0$，便可证明 $Diam(\Omega(x,Z)) \to 0$。

令随机变量 $U_i \sim \mathcal{U}(0,1)$ 表示第 j 个属性在第 i 次划分中的位置，因此 $max(U_i, 1-U_i)$ 表示第 j 个属性在第 i 次划分中的最大长度。令 $K_j \sim \mathcal{B}(T_m, 1/d)$ 表示第 j 个属性被选用划分的次数。此时，第 j 个属性的边长的 K_j 次划分中最大长度的期望值为 $\mathbb{E}_{K_j}[\prod_{i=1}^{K_j} max(U_i, 1-U_i)]$，于是我们可以得到属性边长的期望满足（6.97）。

令 T_m 表示区域 $\Omega(x,Z)$ 被划分的次数，结合（6.98）及划分点的独立性，我们可以得到：

$$
\begin{aligned}
\mathbb{E}(L_j) &\leq \mathbb{E}[\mathbb{E}_{K_j}[\prod_{i=1}^{K_j} max(U_i, 1-U_i)]] \\
&= \mathbb{E}[(\mathbb{E}[max(U_1, 1-U_1)])^{K_j}] \\
&= \mathbb{E}[(\frac{3}{4})^{K_j}] \\
&= \sum_{K_j=0}^{T_m} P(K_j) \cdot (\frac{3}{4})^{K_j} \\
&= \sum_{K_j=0}^{T_m} \binom{T_m}{K_j} \cdot (\frac{1}{d})^{K_j} \cdot (1-\frac{1}{d})^{T_m-K_j} \cdot (\frac{3}{4})^{K_j} \\
&= \sum_{K_j=0}^{T_m} \binom{T_m}{K_j} \cdot (\frac{3}{4d})^{K_j} \cdot (1-\frac{1}{d})^{T_m} \\
&= (1-\frac{1}{d}+\frac{3}{4d})^{T_m} \\
&= (1-\frac{1}{4d})^{T_m}
\end{aligned}
\tag{257}
$$

此时，只需证明当 $k \to \infty$ 时 $T_m \to \infty$，便可证明 $\mathbb{E}(L_j) \to 0$。

每次划分节点都会增加一个新节点，且每次选择节点进行划分的概率均为 $p = 1/i$，其中 i 为当前的节点数目。因此，区域 $\Omega(x, Z)$ 在节点数为 i 时被选中进行划分的概率分布满足 $\xi_i \sim Bernoulli(p)$。此时，划分次数 ξ_i 之和表示 $T_m = \sum_{i=1}^{k} \xi_i$。

由于 T_m 的期望为 $\mathbb{E}[T_m] = \sum_{i=1}^{k} \frac{1}{i}$，根据调和级数的发散性，当 $k \to \infty$ 时 $\mathbb{E}[T_m] \to \infty$。因此，$T_m \to \infty$ 必然依概率成立，从而证明了 $Diam(\Omega(x, Z)) \to 0$。

第7章：收敛率

编辑：赵志民

本章前言

本章的内容围绕学习理论中的算法收敛率（convergence rate）展开。具体来说，我们将探讨在确定性优化和随机优化中的收敛率问题，并在最后分析支持向量机的实例。

7.1 【概念解释】算法收敛率

在算法分析中，收敛率是指迭代算法逼近解或收敛到最优或期望结果的速度，它衡量算法在减少当前解与最优解之间差异的快慢。

设 $\{x_k\}$ 是算法生成的迭代序列，我们可以根据以下公式来衡量算法的收敛率：

$$
\lim_{t \to +\infty} \frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|^p} = C \tag{258}
$$

其中，C为收敛因子，p为收敛阶数，x^* 表示最优解，$\|.\|$ 表示适当的范数。

根据收敛率的不同情况，我们可以将其分类如下：

1. **超线性收敛**：$p \geq 1$，$C = 0$，表明每次迭代都会使得误差减小，且减小的速度越来越快。特别地，当$p > 1$时，称为p阶收敛。例如，$p = 2$时称为平方收敛，$p = 3$时称为立方收敛。
2. **线性收敛**：$p = 1$，$C > 0$，表明每次迭代都会使得误差减小（误差呈几何级数下降），但减小的速度是一定的。
3. **次线性收敛**：$p = 1$，$C = 1$，表明每次迭代都会使得误差减小，但减小的速度越来越慢。

7.2 【证明补充】凸函数的确定性优化

书中给出的梯度下降算法中，输出的是 T 轮迭代的均值 ω，而不是最后一次迭代的结果 ω_T。这是因为在凸函数的梯度下降过程中，所设定的步长 η 是启发式的，因此每次迭代产生的 ω' 无法保证是局部最优解。

根据定理7.1，T 轮迭代的 ω 均值具有次线性收敛率，而无法证明最后一次迭代值 ω_T 也具有相同的收敛率。因此，返回 ω 的均值虽然会增加计算代价，但可以确保稳定的收敛率。这一思想在7.3.1和7.3.2中梯度下降算法中也有体现。

作为对比，在7.2.2中的强凸函数梯度下降算法中，我们只输出了最后一次迭代值 ω_T。这是因为在强凸函数的条件下，每次迭代的梯度更新都有闭式解 $\omega_{t+1} = \omega_t - \frac{1}{\gamma}\nabla f(\omega_t)$。这种情况下，每次迭代无需启发式算法便可得到该临域的全局最优解，这也是此算法拥有更快收敛率（线性收敛率）的原因。因此，无需返回历史 ω 的均值。

另外，在 **139页** 定理7.1的（7.12）推导中，利用了第一章补充内容 AM-GM 不等式 $n = 2$ 的结论，即对于任意非负实数 x, y，有：

$$\sqrt{xy} \leq \frac{x+y}{2} \tag{259}$$

当且仅当 $x = y$ 时取等号。

因此，只有当 $\frac{\Gamma^2}{2\eta T} = \frac{\eta l^2}{2}$ 时，公式（7.12）中 $\frac{\Gamma^2}{2\eta T} + \frac{\eta l^2}{2}$ 才能取得最小值 $\frac{l\Gamma}{\sqrt{T}}$，此时步长 η 应设置为 $\frac{\Gamma}{l\sqrt{T}}$。类似的推导可以在（7.35）和（7.39）中找到。

7.3 【证明补充】强凸函数的确定性优化

142页 中，在证明定理7.3时，对于（7.19）的推导补充如下。

首先，如果目标函数满足 λ-强凸且 γ-光滑，那么根据第一章补充内容中的结论，我们有 $\gamma \geq \lambda$。这是因为对于任意 ω, ω'，光滑系数 γ 被定义为：

$$f(\omega) \leq f(\omega') + \nabla f(\omega')^T(\omega - \omega') + \frac{\gamma}{2}\|\omega - \omega'\|^2 \tag{260}$$

而强凸系数 λ 被定义为：

$$f(\omega) \geq f(\omega') + \nabla f(\omega')^T(\omega - \omega') + \frac{\lambda}{2}\|\omega - \omega'\|^2 \tag{261}$$

光滑系数 γ 决定了 $f(\omega)$ 的上界，而强凸系数 λ 决定了 $f(\omega)$ 的下界，因此光滑系数 γ 不小于强凸系数 λ。

接着，令 $f(\alpha) = \frac{\gamma-\lambda}{\lambda}\alpha^2 - \alpha$，由于 $\frac{\gamma-\lambda}{\lambda} \geq 0$，我们可以分成以下两种情况讨论：

1. 当 $\frac{\gamma-\lambda}{\lambda} = 0$ 时，（7.19）转化为：

$$\begin{aligned} f(\omega_{t+1}) &\leq \min_{\alpha\in[0,1]}\{f(\omega_t) - \alpha(f(\omega_t) - f(\omega^*))\} \\ \Rightarrow f(\omega_{t+1}) - f(\omega^*) &\leq \min_{\alpha\in[0,1]}\{1-\alpha\}(f(\omega_t) - f(\omega^*)) \end{aligned} \tag{262}$$

因为 $f(\omega_t) - f(\omega^*) \geq 0$，所以当且仅当 $\alpha = 1$ 时，不等式右侧取得最小值 0，此时易知 $f(\omega_{t+1}) = f(\omega^*)$。根据凸函数局部最优解等于全局最优解的结论，我们可以得到 $\omega_{t+1} = \omega^*$，即算法在第 $t+1$ 轮迭代中收敛到最优解。

2. 当 $\frac{\gamma-\lambda}{\lambda} > 0$ 时，$f(\alpha)$ 为开口向上的二次函数。令 $f'(\alpha) = 2\frac{\gamma-\lambda}{\lambda}\alpha - 1 = 0$，得到 $f(\alpha)$ 的对称轴为 $\alpha = \frac{\lambda}{2(\gamma-\lambda)}$。我们可以分成以下两种情况讨论：
   - 当 $\frac{\lambda}{2(\gamma-\lambda)} \geq 1$ 时，$f(\alpha)$ 取得最小值只能在 $\alpha = 1$ 处，故而得到（7.20）。
   - 当 $0 < \frac{\lambda}{2(\gamma-\lambda)} < 1$ 时，$f(\alpha)$ 取得最小值在 $\alpha = \frac{\lambda}{2(\gamma-\lambda)}$ 处，故而得到（7.21）。

余下的推导部分与书中相同，此处不再赘述。

7.4 【定理证明】鞅差序列的 Bernstein 不等式

149页 定理7.6 给出了鞅差序列的 Bernstein 不等式，我们在这里给出完整的证明过程。

假设 $X_1, \cdots, X_n$ 是定义在 $f = (f_i)_{1 \le i \le n}$ 上的有界鞅差序列且 $|X_i| \le K$，令：

$$S_i = \sum_{j=1}^{i} X_j \tag{263}$$

将X_n 的条件方差定义为：

$$V_n^2 = \sum_{k=1}^{n} \mathbb{E}[X_k^2 | F_{k-1}] \tag{264}$$

那么对于任意正数 t 和 v，有：

$$P(\max_{i=1,\cdots,k} S_i > t, V_k^2 \le v) \le \exp\left(-\frac{t^2}{2(v + Kt/3)}\right) \tag{265}$$

证明

考虑函数 $f(x) = (e^{\theta x} - \theta x - 1)/x^2$，且 $f(0) = \theta^2/2$。通过对该函数求导，我们知道该函数是非减的。即 $f(x) \le f(1)$，当 $x \le 1$ 时：

$$e^{\theta x} = 1 + \theta x + x^2 f(x) \le 1 + \theta x + x^2 f(1) = 1 + \theta x + g(\theta) x^2, \quad x \le 1 \tag{266}$$

将其用于随机变量 X_k/K 的期望，可得：

$$\mathbb{E}\left[\exp\left(\frac{\theta X_k}{K}\right) \middle| \mathcal{F}_{k-1}\right] \le 1 + \frac{\theta}{K}\mathbb{E}\left[X_k | \mathcal{F}_{k-1}\right] + \frac{g(\theta)}{K^2}\mathbb{E}\left[X_k^2 | \mathcal{F}_{k-1}\right] \tag{267}$$

由于 $\{X_k\}$ 是一个鞅差序列，我们有 $\mathbb{E}\left[X_k | \mathcal{F}_{k-1}\right] = 0$，结合 $1 + x \le e^x, x \ge 0$，我们得到：

$$\mathbb{E}\left[\exp\left(\frac{\theta X_k}{K}\right) \middle| \mathcal{F}_{k-1}\right] = 1 + \frac{g(\theta)}{K^2}\mathbb{E}\left[X_k^2 | \mathcal{F}_{k-1}\right] \le \exp\left(g(\theta)\frac{\mathbb{E}[X_k^2 | \mathcal{F}_{k-1}]}{K^2}\right) \tag{268}$$

考虑一个随机过程：

$$Q_k = \exp\left(\theta \frac{S_k}{K} - g(\theta)\frac{V_k^2}{K^2}\right), \quad Q_0 = 1 \tag{269}$$

我们证明这个过程对于滤波 $\mathcal{F}_n$ 是一个超鞅，即 $\mathbb{E}[Q_k | \mathcal{F}_{k-1}] \le Q_{k-1}$。

证明如下：

$$\begin{aligned}
\mathbb{E}[Q_k | \mathcal{F}_{k-1}] &= \mathbb{E}\left[\exp\left(\theta \frac{S_k}{K} - g(\theta)\frac{V_k^2}{K^2}\right) \middle| \mathcal{F}_{k-1}\right] \\
&= \mathbb{E}\left[\exp\left(\theta \frac{S_{k-1}}{K} - g(\theta)\frac{V_{k-1}^2}{K^2} - g(\theta)\frac{\mathbb{E}[X_k^2 | \mathcal{F}_{k-1}]}{K^2} + \theta\frac{X_k}{K}\right) \middle| \mathcal{F}_{k-1}\right] \\
&= \exp\left(\theta \frac{S_{k-1}}{K} - g(\theta)\frac{V_{k-1}^2}{K^2} - g(\theta)\frac{\mathbb{E}[X_k^2 | \mathcal{F}_{k-1}]}{K^2}\right) \mathbb{E}\left[\exp\left(\theta\frac{X_k}{K}\right) \middle| \mathcal{F}_{k-1}\right]
\end{aligned} \tag{270}$$

应用之前证明的不等式，我们得到：

$$\mathbb{E}[Q_k | \mathcal{F}_{k-1}] \le \exp\left(\theta \frac{S_{k-1}}{K} - g(\theta)\frac{V_{k-1}^2}{K^2}\right) = Q_{k-1} \tag{271}$$

我们定义 $A = \{k \ge 0 : \max_{i=1,\cdots,k} S_i > t, V_k^2 \le v\}$，则有：

$$Q_k \geq \exp\left(\theta\frac{t}{K} - g(\theta)\frac{v}{K^2}\right), k \in A \tag{272}$$

由于 $\{Q_k\}$ 是超鞅，我们有：

$$\mathbb{E}[Q_k|\mathcal{F}_{k-1}] \leq \mathbb{E}[Q_{k-1}|\mathcal{F}_{k-2}] \leq \cdots \leq Q_0 = 1 \tag{273}$$

考虑到 $1 \geq \mathbb{P}(A)$，我们有：

$$1 \geq \mathbb{E}[Q_k|\mathcal{F}_{k-1}] \geq \exp\left(\theta\frac{t}{K} - g(\theta)\frac{v}{K^2}\right)\mathbb{P}(A), k \in A \tag{274}$$

因此：

$$\mathbb{P}(A) \leq \exp\left(g(\theta)\frac{v}{K^2} - \theta\frac{t}{K}\right) \tag{275}$$

由于上述不等式对任何 $\theta > 0$ 都成立，我们可以写为：

$$P(A) \leq \inf_{\theta>0} \exp\left(g(\theta)\frac{v}{K^2} - \theta\frac{t}{K}\right) \tag{276}$$

检查不等式右边的一阶导数，我们知道该下确界在 $\theta = \log(1 + Kt/v)$ 处取得。

对于指数内部的表达式，我们进行如下变换：

$$\begin{aligned}
\theta\frac{t}{K} - g(\theta)\frac{v}{K^2} &= \log\left(1 + \frac{Kt}{v}\right)\frac{t}{K} - \frac{v}{K^2}\left(\frac{Kt}{v} - \log\left(1 + \frac{Kt}{v}\right)\right) \\
&= \frac{v}{K^2}\left(\left(1 + \frac{Kt}{v}\right)\log\left(1 + \frac{Kt}{v}\right) - \frac{Kt}{v}\right) \\
&= \frac{v}{K^2}h\left(\frac{Kt}{v}\right)
\end{aligned} \tag{277}$$

其中 $h(u) = (1+u)\log(1+u) - u$。

通过对表达式求二阶导数的方法，我们也可以证明：

$$h(u) \geq \frac{u^2}{2(1 + u/3)}, \quad u \geq 0 \tag{278}$$

综上所述，我们有：

$$P(A) \leq \exp\left(-\frac{v}{K^2}h\left(\frac{Kt}{v}\right)\right) \leq \exp\left(-\frac{v}{K^2}\frac{K^2t^2}{2v(v + Kt/3)}\right) = \exp\left(-\frac{t^2}{2(v + Kt/3)}\right) \tag{279}$$

7.5 【证明补充】Epoch-GD 的收敛率

150页 引理7.2给出了Epoch-GD外层循环收敛率的泛化上界，我们对其中部分推导进行必要补充。

首先，（7.60）中第二个不等式的推导利用了Cauchy-Schwarz不等式（1.14），即 $\|x^Ty\| \leq \|x\|\|y\|$。这里，我们令 $x = \underbrace{[1, \cdots, 1]}_{T}$，$y = \underbrace{[\|\omega_1 - w^*\|, \cdots, \|\omega_T - w^*\|]}_{T}$，则有：

$$|x^Ty| = \sum_{t=1}^{T}\|\omega_t - w^*\| \leq \sqrt{T}\sqrt{\sum_{t=1}^{T}\|\omega_t - w^*\|^2} = |x\|y| \tag{280}$$

其次，（7.62）中最后两个不等式的推导利用了一些常见的缩放技巧，我们在这里给出完整形式：

$$
\begin{aligned}
&\sum_{i=1}^{m} P\left(\sum_{t=1}^{T}\delta_t \geq 2\sqrt{4l^2A_T\tau}+\frac{2}{3}\frac{4l^2}{\lambda}\tau+\frac{4l^2}{\lambda}, V_T^2\leq 4l^2A_T, A_T\in\left(\frac{4l^2}{\lambda^2T}2^{i-1},\frac{4l^2}{\lambda^2T}2^{i}\right)\right)\\
\leq&\sum_{i=1}^{m} P\left(\sum_{t=1}^{T}\delta_t \geq 2\sqrt{4l^2A_T\tau}+\frac{2}{3}\frac{4l^2}{\lambda}\tau, V_T^2\leq 4l^2A_T, A_T\in\left(\frac{4l^2}{\lambda^2T}2^{i-1},\frac{4l^2}{\lambda^2T}2^{i}\right)\right)\\
\leq&\sum_{i=1}^{m} P\left(\sum_{t=1}^{T}\delta_t \geq \sqrt{2\frac{16l^4 2^i}{\lambda^2T}\tau}+\frac{2}{3}\frac{4l^2}{\lambda}\tau, V_T^2\leq \frac{16l^4 2^i}{\lambda^2T}\right)\\
\leq&\sum_{i=1}^{m} P\left(\max_{j=1,\cdots,T}\underbrace{\sum_{t=1}^{j}\delta_t}_{S_j} \geq \sqrt{2\underbrace{\frac{16l^4 2^i}{\lambda^2T}}_{\nu}\tau}+\frac{2}{3}\underbrace{\frac{4l^2}{\lambda}}_{K}\tau, V_T^2\leq \underbrace{\frac{16l^4 2^i}{\lambda^2T}}_{\nu}\right)\\
\leq&\sum_{i=1}^{m} e^{-\tau}\\
=&me^{-\tau}
\end{aligned}
\tag{281}
$$

这里，第一个不等式利用了 $\frac{4l^2}{\lambda}>0$ 的事实对 $\sum_{t=1}^{T}\delta_t$ 的范围进行概率缩放； 第二个不等式利用了 A_T 的下界和上界分别对 $\sum_{t=1}^{T}\delta_t$ 和 V_T^2 的范围进行概率缩放； 第三个不等式利用了 $\max_{j=1,\cdots,T}\sum_{t=1}^{j}\delta_t$ 比 $\sum_{t=1}^{T}\delta_t$ 更为宽松的事实对 V_T^2 进行概率缩放； 第四个不等式利用了定理7.6的结论。

最后，（7.64）中第二个不等式的推导利用了开口向下的二次函数 $f(x)=ax^2+bx+c, a<0$ 拥有最大值点 $x_0=-\frac{b}{2a}$ 的事实。我们令 $x=\sqrt{A_T}$，然后取 $a=-\frac{\lambda}{2}, b=2\sqrt{4l^2\ln\frac{m}{\delta}}, c=0$，则易知 $f(x)$ 的最大值为 $\frac{8l^2}{\lambda}\ln\frac{m}{\delta}$，于是得到了（7.64）中的结论。

进一步地，**152页**引理7.3利用数学归纳法给出了特定步长和迭代次数下Epoch-GD外层循环收敛率的泛化上界，这为**154页**定理7.7中Epoch-GD的收敛率奠定了基础。我们对后者的部分推导进行必要补充。

首先，观察（7.75）可以发现，Epoch-GD外层的迭代次数 k 需要满足 $\frac{\alpha}{2}(2^k-1)\leq T$，即 $k=\lfloor\log_2(\frac{2T}{\alpha}+1)\rfloor$，因此构造了（7.66）中的 $k^\dagger$。

其次，（7.77）的推导利用了函数 $f(x)=(1-\frac{1}{x})^x$ 在 $x=\frac{k^\dagger}{\delta}>1$ 时单调递增的事实，以下是更严格的证明。

对函数 $f(x)$ 两边取对数，得到：

$$
\ln f(x)=x\ln(1-\frac{1}{x}) \tag{282}
$$

接着对两边分别求导，可得：

$$
\frac{f'(x)}{f(x)}=\ln(1-\frac{1}{x})+\frac{1}{x-1} \tag{283}
$$

易知当 $x>1$ 时，$f(x)>0$，因此我们只需要关注等式右边在 $x>1$ 时的符号。令 $g(x)=\ln(1-\frac{1}{x})+\frac{1}{x-1}$，则有：

$$
g'(x)=\frac{1}{x(x-1)^2} \tag{284}
$$

易知当 $x>1$ 时，$g'(x)<0$，因此：

$$
g(x)>\lim_{x\to+\infty}g(x)=\lim_{x\to+\infty}\ln(1-\frac{1}{x})+\lim_{x\to+\infty}\frac{1}{x-1}=0 \tag{285}
$$

综上，当 $x>1$ 时，$\frac{f'(x)}{f(x)}=g(x)>0$，即 $f'(x)>0$，因此 $f(x)$ 在 $x>1$ 时单调递增。

第8章：遗憾界

编辑：赵志民，詹好

本章前言

本章的内容围绕学习理论中的遗憾（regret）概念展开（有的教材里也翻译为“悔”）。通常，我们使用超额风险（excess risk）来评估批量学习的分类器性能，而用遗憾来评估在线学习的分类器性能。二者的不同在于，前者衡量的是整个学习过程结束后所得到的分类器性能，可以理解为学习算法**最终输出的模型**与假设空间内**最优模型**的风险之差；而后者衡量的是算法运行过程中，所产生的**模型**与假设空间内**最优模型**的损失之差的**和**。

8.1 【概念解释】超额风险与遗憾的区别

8.1介绍了遗憾这一评估指标的基本概念，我们在此基础上梳理一下其与超额风险这一评估指标的区别。

超额风险这一评估指标被定义为：

$$ER = \mathbb{E}_{(x,y)\sim D}[l(w_{T+1},(x,y))] - \min_{w\in W}\mathbb{E}_{(x,y)\sim D}[l(w,(x,y))] \tag{286}$$

其中，ER 指的是excess risk，等式右边的前半部分 $\mathbb{E}_{(x,y)\sim D}[l(w_{T+1},(x,y))]$ 指的是模型 w_{T+1} 的风险，等式右边的后半部分 $\min_{w\in W}\mathbb{E}_{(x,y)\sim D}[l(w,(x,y))]$ 指的是假设空间内的最优模型的风险。值得注意的是，这里的评估是在整个数据集上进行的，也正是因为如此，我们必须要引入期望的操作。

而遗憾这一评估指标，被定义为：

$$regret = \sum_{t=1}^{T} f_t(w_t) - \min_{w\in W}\sum_{t=1}^{T} f_t(w) \tag{287}$$

其中，$f_t(w_t)$ 指的是：

$$\sum_{t=1}^{T} l(w_t,(x_t,y_t)) - \min_{w\in W}\sum_{t=1}^{T} l(w,(x_t,y_t)) \tag{288}$$

由于w_t的计算过程与样本(x_t,y_t) 无关，而是与$(x_1,y_1),...,(x_{t-1},y_{t-1})$ 有关，因此可以直接使用 $l(w,(x_t,y_t))$ 来衡量性能。

由此，我们可以总结出二者之间的两个主要区别：首先，超额风险引入了**期望**，而遗憾没有；其次，超额风险是在所有数据上进行的一次性计算，而遗憾是对多次损失的一个**求和**。同时，由于在线学习不依赖于任何分布假设，因此适用于非独立同分布样本或固定分布的情形。

8.2 【案例分享】Maler 算法

在8.2.3节的**170页**末尾，作者提到了Maler算法（multiple sub-algorithms and learning rates）（详细证明参考：Adaptivity and Optimality: A Universal Algorithm for Online Convex Optimization），这是一个能够自适应选择最优专家的在线学习算法，并在不同类型的损失函数上实现最优的遗憾界限：

- **一般凸函数**：$R(T) \leq O\sqrt{T})$
- **指数凹函数**：$R(T) \leq O(d\log T)$
- **强凸函数**：$R(T) \leq O(\log T)$ 这里T表示时间总步数，d表示特征空间的维度。

下面，我们简要补充Maler算法的原理和实现。

假设和定义

1. **假设 1（梯度有界性）**：所有损失函数 $f_t(x)$ 的梯度被 G 所有界：

$$\forall t>0,\quad \max_{x\in D}\|\nabla f_t(x)\| \leq G \tag{289}$$

2. **假设 2（行动集的直径有界性）**：行动集 D 的直径被 D 所有界：

$$\max_{x_1,x_2\in D}\|x_1-x_2\|\le D \tag{290}$$

3. **定义 1（凸函数）**：函数 $f:D\to\mathbb{R}$ 是凸的，如果：

$$f(x_1)\ge f(x_2)+\nabla f(x_2)^\top(x_1-x_2),\quad \forall x_1,x_2\in D \tag{291}$$

4. **定义 2（强凸函数）**：函数 $f:D\to\mathbb{R}$ 是 λ-强凸的，如果：

$$f(x_1)\ge f(x_2)+\nabla f(x_2)^\top(x_1-x_2)+\frac{\lambda}{2}\|x_1-x_2\|^2,\quad \forall x_1,x_2\in D \tag{292}$$

5. **定义 3（指数凹函数）**：函数 $f:D\to\mathbb{R}$ 是 α-指数凹的（简称 α-exp-concave），如果：

$$\exp(-\alpha f(x))\text{是凹的} \tag{293}$$

元算法（Maler）

输入：学习率 $\eta^c,\eta_1,\eta_2,\dots$，专家的先验权重 $\pi_1^c,\pi_1^{\eta_1,s},\pi_1^{\eta_2,s}\dots$，以及 $\pi_1^{\eta_1,l},\pi_1^{\eta_2,l},\dots$。

1. **对于每个回合** $t=1,\dots,T$：
   - 从凸专家算法（专家 1）获取预测 x_t^c，从指数凹专家算法（专家 2）和强凸专家算法（专家 3）分别获取 $x_t^{\eta,l}$ 和 $x_t^{\eta,s}$。
   - 执行：

$$x_t=\frac{\pi_t^c\eta^c x_t^c+\sum_\eta(\pi_t^{\eta,s}\eta x_t^{\eta,s}+\pi_t^{\eta,l}\eta x_t^{\eta,l})}{\pi_t^c\eta^c+\sum_\eta(\pi_t^{\eta,s}\eta+\pi_t^{\eta,l}\eta)} \tag{294}$$

   - 观察梯度 g_t 并发送给所有专家算法。
   - 对所有的 η 更新权重：

$$\pi_{t+1}^c=\frac{\pi_t^c e^{-c_t(x_t^c)}}{\Phi_t},\quad \pi_{t+1}^{\eta,s}=\frac{\pi_t^{\eta,s}e^{-s_t^\eta(x_t^{\eta,s})}}{\Phi_t},\quad \pi_{t+1}^{\eta,l}=\frac{\pi_t^{\eta,l}e^{-l_t^\eta(x_t^{\eta,l})}}{\Phi_t} \tag{295}$$

   其中：

$$\Phi_t=\sum_\eta(\pi_t^{\eta,s}e^{-s_t^\eta(x_t^{\eta,s})}+\pi_t^{\eta,l}e^{-l_t^\eta(x_t^{\eta,l})})+\pi_t^c e^{-c_t(x_t^c)} \tag{296}$$

凸专家算法（专家 1）

1. $x_1^c=0$
2. **对于每个回合** $t=1,\dots,T$：
   - 将 x_t^c 发送给元算法
   - 从元算法接收梯度 g_t
   - 更新：

$$x_{t+1}^c=\Pi_D^{I_d}(x_t^c-\frac{D}{\eta^c G\sqrt{t}}\nabla c_t(x_t^c)) \tag{297}$$

   其中 $\nabla c_t(x_t^c)=\eta^c g_t$

指数凹专家算法（专家 2）

1. **输入**：学习率 η
2. $x_1^{\eta,l}=0,\beta=\frac{1}{2}\min\{\frac{1}{4G^lD},1\},G^l=\frac{7}{25D},\Sigma_1=\frac{1}{\beta^2D^2}I_d$
3. **对于每个回合** $t=1,\dots,T$：

- 将 $x_t^{\eta,l}$ 发送给元算法
- 从元算法接收梯度 g_t
- 更新：

$$\begin{aligned}\Sigma_{t+1} &= \Sigma_t + \nabla l_t^\eta(x_t^{\eta,l})\nabla l_t^\eta(x_t^{\eta,l})^\top \\ x_{t+1}^{\eta,l} &= \Pi_D^{\Sigma_{t+1}}(x_t^{\eta,l} - \frac{1}{\beta}\Sigma_{t+1}^{-1}\nabla l_t^\eta(x_t^{\eta,l}))\end{aligned} \tag{298}$$

其中 $\nabla l_t^\eta(x_t^{\eta,l}) = \eta g_t + 2\eta^2 g_t g_t^\top(x_t^{\eta,l} - x_t)$

强凸专家算法（专家 3）

1. **输入**：学习率 η
2. $x_1^{\eta,s} = 0$
3. **对于每个回合** $t = 1, \dots, T$：
   - 将 $x_t^{\eta,s}$ 发送给元算法
   - 从元算法接收梯度 g_t
   - 更新：

$$x_{t+1}^{\eta,s} = \Pi_D^{I_d}(x_t^{\eta,s} - \frac{1}{2\eta^2 G^2 t}\nabla s_t^\eta(x_t^{\eta,s})) \tag{299}$$

其中 $\nabla s_t^\eta(x_t^{\eta,s}) = \eta g_t + 2\eta^2 G^2(x_t^{\eta,s} - x_t)$

8.3 【证明补充】随机多臂赌博机的遗憾界

172页中定理8.3给出了随机多臂赌博机的遗憾界，我们在此基础上对公式（8.42）至（8.47）证明过程进行补充。

首先，（8.42）给出当$\overline{\mu}_*(p) + \sqrt{\frac{2\ln t}{p}} \le \overline{\mu}_i(q) + \sqrt{\frac{2\ln t}{q}}$成立时，必然有三种可能情况中的一种成立。但这三种情况并不是互斥的，因此显得不直观，这里将第二种情况做了细微调整，即：

$$\overline{\mu}_*(p) + \sqrt{\frac{2\ln t}{p}} \le \mu_*, \mu_* \le \overline{\mu}_i(q) + \sqrt{\frac{2\ln t}{q}}, \overline{\mu}_i(q) + \sqrt{\frac{2\ln t}{q}} \le \overline{\mu}_i(p) \tag{300}$$

此时，构造（8.44）和（8.45）的逻辑更加顺畅。我们令$l = \lceil (2\ln T)/\Delta_i^2 \rceil$，则（8.45）转化为：

$$P(\mu_* \le \mu_i + \sqrt{\frac{2\ln t}{q}}) = 0, q \ge l \tag{301}$$

代入（8.44），可得：

$$\begin{aligned}\mathbb{E}[n_i^T] &\le \lceil \frac{2\ln T}{\Delta_i^2} \rceil + 2\sum_{t=1}^{T-1}\sum_{p=1}^{t-1}\sum_{q=l}^{t-1} t^{-4} \\ &\le \frac{2\ln T}{\Delta_i^2} + 1 + 2\sum_{t=1}^{T-1}\sum_{p=1}^{t}\sum_{q=1}^{t} t^{-4} \\ &\le \frac{2\ln T}{\Delta_i^2} + 1 + 2\lim_{T\to+\infty}\sum_{t=1}^{T-1} t^{-2}\end{aligned} \tag{302}$$

根据p-级数判别法，当$p = 2 > 1$时，级数收敛，因此$\lim_{T\to+\infty}\sum_{t=1}^{T-1} t^{-2}$是有界的。至于该级数的具体值，对定理的结论没有影响，因此我们可以将其视为一个常数，然后带入后续推导中。为了证明的完整性，我们对此进行简要说明。

$\lim_{T\to+\infty}\sum_{t=1}^{T-1} t^{-2}$的取值在数学界被称为Basel问题，推导过程涉及诸多前置定理，感兴趣的读者可以查看这个讲义：The Basel Problem - Numerous Proofs。此处提供另一种在微积分变换中常见的缩放方法：

$$\begin{aligned}\sum_{t=1}^{T-1} t^{-2} &\leq 1+\int_{1}^{T-1} \frac{1}{x^2} dx \\ &= 1+(-\frac{1}{x})|_1^{T-1} \\ &= 2-\frac{1}{T}\end{aligned} \tag{303}$$

对不等式两边同时取极限，可得：

$$\lim_{T\to+\infty} \sum_{t=1}^{T-1} t^{-2} \leq 2 \tag{304}$$

代入（8.46），同样可得类似（8.47）的结论。

这里继续沿用书中给出的$\lim_{T\to+\infty} \sum_{t=1}^{T} t^{-2} = \frac{\pi^2}{6}$，代入（8.46）得到遗憾界（8.47）：

$$\mathbb{E}[regret] \leq \sum_{i=1}^{K} \frac{2\ln T}{\Delta_i^2} + O(1) \tag{305}$$

此时（8.46）变为：

$$\mathbb{E}[n_i^T] \leq \sum_{i\neq *}^{K} \frac{2\ln T}{\Delta_i} + (1+\frac{\pi^2}{3})\Delta_i = O(K\log T) \tag{306}$$

观察（8.47）可知，求和公式中的每一项符合对钩函数的构造，即：

$$f(x) = Ax + \frac{B}{x}, x>0, A>0, B>0 \tag{307}$$

这里$x=\Delta_i, A=1+\frac{\pi^2}{3}, B=2\ln T$，因此无论$\Delta_i$过大或过小时，都会导致遗憾界的上界变大。另外，遗憾界跟摇臂的个数K呈线性关系，当K越大时，遗憾界也越大。

8.4 【概念解释】线性赌博机

176页的8.3.2节介绍了线性赌博机的概念，我们在此基础上对参数估计部分进行补充。

为了估计线性赌博机的参数，我们将原问题转化为岭回归问题，即（8.52）：

$$f(w) = (Y-w^TX)^T(Y-w^TX) + \lambda w^Tw \tag{308}$$

为了求得最优解w^*，我们令$f'(w)=0$，可推导出（8.53）：

$$\begin{aligned}\frac{\partial f(w)}{\partial w} = -2X^T(Y-w^TX) + 2\lambda w = 0 \\ \rightarrow X^TY = (X^TX+\lambda I)w \\ \rightarrow w^* = (X^TX+\lambda I)^{-1}X^TY\end{aligned} \tag{309}$$

相比于每次传入新数据(x_t, y_t)时从头计算w_t，这里巧妙地利用了 Sherman-Morrison-Woodbury 公式将任何形如$(A+uv^T)^{-1}$的矩阵逆转化为可逆矩阵A和列向量u, v之间的运算，在$O(d^2)$的时间复杂度内完成参数的更新。

8.5 【证明补充】Sherman-Morrison-Woodbury (或 Woodbury) 公式

177页的 Sherman-Morrison-Woodbury 公式变种是矩阵求逆中的一个重要工具，它可以通过已知矩阵的逆来快速计算被低秩修正的矩阵的逆。

该公式如下所示：

$$(A+UCV)^{-1}=A^{-1}-A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1} \tag{310}$$

其中，A 是一个 $n\times n$ 的矩阵，C 是 $k\times k$ 的矩阵，U 和 V 是 $n\times k$ 的矩阵，（8.54）中C为单位矩阵。

证明

该公式可以通过验证 $A+UCV$ 与其假设的逆（公式右侧）的乘积是否为单位矩阵来证明。我们对以下乘积进行计算：

$$(A+UCV)[A^{-1}-A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}] \tag{311}$$

逐步推导如下：

$$\begin{aligned}
&=\{I+UCVA^{-1}\}-\{U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}+UCVA^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}\}\\
&=I+UCVA^{-1}-(U+UCVA^{-1}U)(C^{-1}+VA^{-1}U)^{-1}VA^{-1}\\
&=I+UCVA^{-1}-UC(C^{-1}+VA^{-1}U)(C^{-1}+VA^{-1}U)^{-1}VA^{-1}\\
&=I+UCVA^{-1}-UCVA^{-1}\\
&=I
\end{aligned} \tag{312}$$

8.6 【证明补充】单样本的近似梯度

第181页的引理8.2给出了单样本条件下的梯度近似公式，本节将提供该引理的完整证明过程。

$$\mathbb{E}_{u\in\mathbb{S}}[f(x+\delta u)u]=\frac{\delta}{d}\nabla\mathbb{E}_{v\in\mathbb{B}}[f(x+\delta v)] \tag{313}$$

其中：

- d 为空间的维数；
- δ 为任意正数；
- $\mathbb{B}$ 为单位球的空间，即 $\mathbb{B}=\{v\in\mathbb{R}^d\mid\|v\|\le 1\}$；
- $\mathbb{S}$ 为单位球的表面，即 $\mathbb{S}=\{u\in\mathbb{R}^d\mid\|u\|=1\}$。

证明

为了证明上述等式，我们将分三个步骤进行推导。

1. 表达左边的期望

首先，考虑左边的期望：

$$\mathbb{E}_{u\in\mathbb{S}}[f(x+\delta u)u]=\frac{1}{\mathrm{Vol}_{d-1}(\mathbb{S})}\int_{\mathbb{S}}f(x+\delta u)u\,dS(u) \tag{314}$$

其中，$\mathrm{Vol}_{d-1}(\mathbb{S})$ 表示 $(d-1)$ 维单位球面的体积，$dS(u)$ 为球面上的微分面积元素。

进行变量替换，令 $w=\delta u$。此时：

- 当 $u\in\mathbb{S}$ 时，$w\in\delta\mathbb{S}$；
- 球面上的微分面积元素变化为 $dS(u)=\frac{dS(w)}{\delta^{d-1}}$，因为每个维度按 δ 缩放，$(d-1)$ 维体积按 δ^{d-1} 缩放。

将变量替换代入期望的表达式：

$$\mathbb{E}_{u\in\mathbb{S}}[f(x+\delta u)u]=\frac{1}{\mathrm{Vol}_{d-1}(\mathbb{S})}\int_{\mathbb{S}}f(x+\delta u)u\,dS(u)=\frac{1}{\mathrm{Vol}_{d-1}(\mathbb{S})\cdot\delta^{d-1}}\int_{\delta\mathbb{S}}f(x+w)\frac{w}{\delta}\,dS(w) \tag{315}$$

简化后得到：

$$\mathbb{E}_{u\in\mathbb{S}}[f(x+\delta u)u]=\frac{1}{\mathrm{Vol}_{d-1}(\delta\mathbb{S})}\int_{\delta\mathbb{S}}f(x+w)\frac{w}{\|w\|}\,dS(w) \tag{316}$$

2. 表达右边的期望及其梯度

接下来，考虑右边的期望：

$$\mathbb{E}_{v\in\mathbb{B}}[f(x+\delta v)] = \frac{1}{\mathrm{Vol}_d(\mathbb{B})}\int_{\mathbb{B}} f(x+\delta v)\,dv \tag{317}$$

其中，$\mathrm{Vol}_d(\mathbb{B})$ 表示 d 维单位球的体积，dv 为体积上的微分元素。

同样进行变量替换，令 $w=\delta v$。则：

- 当 $v\in\mathbb{B}$ 时，$w\in\delta\mathbb{B}$；
- 微分体积元素变化为 $dv=\frac{dw}{\delta^d}$，因为每个维度按 δ 缩放，体积按 δ^d 缩放。

代入后得到：

$$\mathbb{E}_{v\in\mathbb{B}}[f(x+\delta v)] = \frac{1}{\mathrm{Vol}_d(\mathbb{B})\cdot\delta^d}\int_{\delta\mathbb{B}} f(x+w)\,dw = \frac{1}{\mathrm{Vol}_d(\delta\mathbb{B})}\int_{\delta\mathbb{B}} f(x+w)\,dw \tag{318}$$

为了计算 $\nabla\mathbb{E}_{v\in\mathbb{B}}[f(x+\delta v)]$，令：

$$F(x) = \mathbb{E}_{v\in\mathbb{B}}[f(x+\delta v)] = \frac{1}{\mathrm{Vol}_d(\delta\mathbb{B})}\int_{\delta\mathbb{B}} f(x+w)\,dw \tag{319}$$

梯度作用在积分上，由于 x 和 w 是独立变量，可以将梯度算子移入积分内部：

$$\nabla F(x) = \frac{1}{\mathrm{Vol}_d(\delta\mathbb{B})}\int_{\delta\mathbb{B}} \nabla_x f(x+w)\,dw \tag{320}$$

注意到：

$$\nabla_x f(x+w) = \nabla_w f(x+w) \tag{321}$$

这是因为 x 和 w 的关系是通过相加连接的，故梯度对 x 的作用等同于对 w 的作用。

根据散度定理，有：

$$\int_{\delta\mathbb{B}} \nabla_w f(x+w)\,dw = \int_{\delta\mathbb{S}} f(x+w)n(w)\,dS(w) \tag{322}$$

其中，$\delta\mathbb{S}$ 是半径为 δ 的球面，$n(w)$ 为点 w 处的单位外法向量。因此：

$$\nabla F(x) = \frac{1}{\mathrm{Vol}_d(\delta\mathbb{B})}\int_{\delta\mathbb{S}} f(x+w)\frac{w}{\|w\|}\,dS(w) \tag{323}$$

3. 关联两边的表达式

将步骤 1 和步骤 2 的结果进行对比，可以得到：

$$\mathbb{E}_{u\in\mathbb{S}}[f(x+\delta u)u] = \frac{\mathrm{Vol}_d(\delta\mathbb{B})}{\mathrm{Vol}_{d-1}(\delta\mathbb{S})}\nabla\mathbb{E}_{v\in\mathbb{B}}[f(x+\delta v)] \tag{324}$$

为了确定系数，我们需要利用 d 维球的体积与表面积之间的关系。

d 维球的体积与半径 δ 的关系为：

$$\mathrm{Vol}_d(\delta\mathbb{B}) = \delta^d\cdot\mathrm{Vol}_d(\mathbb{B}) \tag{325}$$

而球面的表面积与半径 δ 的关系为：

$$\mathrm{Vol}_{d-1}(\delta\mathbb{S}) = \delta^{d-1}\cdot\mathrm{Vol}_{d-1}(\mathbb{S}) \tag{326}$$

结合这两个关系，可以得到：

$$\mathrm{Vol}_d(\delta\mathbb{B}) = \int_0^\delta \mathrm{Vol}_{d-1}(r\mathbb{S})\,dr = \int_0^\delta \mathrm{Vol}_{d-1}(\mathbb{S})\,r^{d-1}\,dr = \frac{\mathrm{Vol}_{d-1}(\mathbb{S})\cdot\delta^d}{d} = \frac{\delta}{d}\cdot\mathrm{Vol}_{d-1}(\delta\mathbb{S}) \tag{327}$$

带入上述等式中，得证：

$$\mathbb{E}_{u\in\mathbb{S}}[f(x+\delta u)u] = \frac{\delta}{d}\nabla\mathbb{E}_{v\in\mathbb{B}}[f(x+\delta v)] \tag{328}$$

8.7 【证明补充】凸赌博机的在线梯度下降

182页中引理8.3给出了凸赌博机的随机版本在线梯度下降，我们在此给出完整的证明过程。

设 $f_1, f_2, \ldots, f_T : W \to \mathbb{R}$ 为一列凸且可微的函数，$\omega_1, \omega_2, \ldots, \omega_T \in W$ 的定义满足 ω_1 为任意选取的点，且 $\omega_{t+1} = \Pi_W(\omega_t - \eta g_t)$，其中 $\eta > 0$，且 $g_1, \ldots, g_T$ 是满足 $\mathbb{E}[g_t|\omega_t] = \nabla f_t(\omega_t)$ 的随机向量变量，且 $\|g_t\| \le l$，其中 $l > 0$。则当 $\eta = \frac{\Lambda}{l\sqrt{T}}$ 时，有：

$$\sum_{t=1}^T \mathbb{E}[f_t(\omega_t)] - \min_{\omega\in W}\sum_{t=1}^T f_t(\omega) \le l\Lambda\sqrt{T} \tag{329}$$

证明:
设 $\omega^\star$ 为在 W 中使 $\sum_{t=1}^T f_t(\omega)$ 最小化的点。由于 f_t 是凸且可微的，我们可以使用梯度界定 $f_t(\omega_t)$ 和 $f_t(\omega^\star)$ 之间的差异：

$$f_t(\omega^\star) - f_t(\omega_t) \ge \nabla f_t(\omega_t)^\top(\omega^\star - \omega_t) = \mathbb{E}[g_t|\omega_t]^\top(\omega^\star - \omega_t) \tag{330}$$

对该不等式取期望，得到：

$$\mathbb{E}[f_t(\omega_t) - f_t(\omega^\star)] \le \mathbb{E}[g_t^\top(\omega_t - \omega^\star)] \tag{331}$$

我们使用 $\|\omega_t - \omega^\star\|^2$ 作为潜在函数。注意到 $\|\Pi_W(\omega) - \omega^\star\| \le \|\omega - \omega^\star\|$，因此：

$$\begin{aligned}
\|\omega_{t+1} - \omega^\star\|^2 &= \|\Pi_W(\omega_t - \eta g_t) - \omega^\star\|^2 \\
&\le \|\omega_t - \eta g_t - \omega^\star\|^2 \\
&= \|\omega_t - \omega^\star\|^2 + \eta^2\|g_t\|^2 - 2\eta(\omega_t - \omega^\star)^\top g_t \\
&\le \|\omega_t - \omega^\star\|^2 + \eta^2 l^2 - 2\eta(\omega_t - \omega^\star)^\top g_t
\end{aligned} \tag{332}$$

整理后得到：

$$g_t^\top(\omega_t - \omega^\star) \le \frac{\|\omega_t - \omega^\star\|^2 - \|\omega_{t+1} - \omega^\star\|^2 + \eta^2 l^2}{2\eta} \tag{333}$$

因此，我们有：

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}[f_t(\omega_t)] - \sum_{t=1}^T f_t(\omega^\star) &= \sum_{t=1}^T \mathbb{E}[f_t(\omega_t) - f_t(\omega^\star)] \\
&\le \sum_{t=1}^T \mathbb{E}[g_t^\top(\omega_t - \omega^\star)] \\
&\le \sum_{t=1}^T \mathbb{E}\left[\frac{\|\omega_t - \omega^\star\|^2 - \|\omega_{t+1} - \omega^\star\|^2 + \eta^2 l^2}{2\eta}\right] \\
&= \frac{\mathbb{E}[\|\omega_1 - \omega^\star\|^2] - \mathbb{E}[\|\omega_{T+1} - \omega^\star\|^2]}{2\eta} + \frac{T\eta l^2}{2} \\
&\le \frac{\mathbb{E}[\|\omega_1 - \omega^\star\|^2]}{2\eta} + \frac{T\eta l^2}{2} \\
&\le \frac{\Lambda^2}{2\eta} + \frac{T\eta l^2}{2}
\end{aligned} \tag{334}$$

代入 $\eta = \frac{\Lambda}{l\sqrt{T}}$ 可得最终结果。

8.8 【证明补充】凸赌博机的缩减投影误差

182页中引理8.4给出了凸赌博机的缩减投影误差，我们在此给出完整的证明过程。

设 $f_1, f_2, \ldots, f_T : W \to \mathbb{R}$ 为一列凸且可微的函数且 $\forall \omega \in W, i \in [T]$ 满足 $|f_i(\omega)| \leq c$，有：

$$\min_{\omega \in (1-\alpha)W} \sum_{t=1}^{T} f_t(\omega) - \min_{\omega \in W} \sum_{t=1}^{T} f_t(\omega) \leq 2\alpha c T \tag{335}$$

证明

显然，$(1-\alpha)W \subseteq W$。因此，有：

$$\min_{\omega \in (1-\alpha)W} \sum_{t=1}^{T} f_t(\omega) = \min_{\omega \in W} \sum_{t=1}^{T} f_t((1-\alpha)\omega) \tag{336}$$

由于每个f_t是凸函数，且$0 \in W$，则我们有：

$$\begin{aligned} \min_{\omega \in W} \sum_{t=1}^{T} f_t((1-\alpha)\omega) &\leq \min_{\omega \in W} \sum_{t=1}^{T} \alpha f_t(0) + (1-\alpha) f_t(\omega) \\ &= \min_{\omega \in W} \sum_{t=1}^{T} \alpha (f_t(0) - f_t(\omega)) + f_t(\omega) \end{aligned} \tag{337}$$

最后，由于对于任意$\omega \in W$和$t \in \{1, \ldots, T\}$，我们有$|f_t(\omega)| \leq c$，因此可以得出：

$$\begin{aligned} \sum_{t=1}^{T} \min_{\omega \in W} \alpha (f_t(0) - f_t(\omega)) + f_t(\omega) &\leq \min_{\omega \in W} \sum_{t=1}^{T} 2\alpha c + f_t(\omega) \\ &= 2\alpha c T + \min_{\omega \in W} \sum_{t=1}^{T} f_t(\omega) \end{aligned} \tag{338}$$

进行适当移项即可得原不等式。

8.9 【证明补充】凸赌博机的遗憾界

182页中定理8.5给出了凸赌博机的遗憾界，在证明开始时，作者对η, α, δ的取值进行了限定。我们可以发现这些取值不是很直观，证明给出的解释也较为分散，部分取值与证明略有出入，因此我们在此进行补充。

对于步长η，在缩放（8.87）中 $\mathbb{E}[\sum_{t=1}^{T} \hat{f}_t(z_t)] - \min_{w \in (1-\alpha)\mathcal{W}} \sum_{t=1}^{T} \hat{f}_t(w)$ 时，为使用引理8.3创造条件，因此采用步长$\eta = \frac{\Lambda}{l'\sqrt{T}}$。根据（8.89）的推导，我们可令$\Lambda = \Lambda_2$且$l' = \frac{dc}{\delta}$，此时，将$\eta = \frac{\Lambda_2}{(dc/\delta)\sqrt{T}}$带入到更新公式（8.76）中即可得到（8.88）。

对于缩减系数α与扰动系数δ，可以一同考虑这两个系数的取值。观察（8.91）第一个不等式的形式，我们发现这是一个关于δ的对钩函数：

$$f(\delta) = A\delta + \frac{B}{\delta} + C \tag{339}$$

假设α的取值与δ无关，那么：

$$A = 3lT, B = dc\Lambda_2\sqrt{T}, C = 2\alpha c T \tag{340}$$

令$f'(\delta) = 0$，可得：

$$\delta^* = T^{-1/4}\sqrt{\frac{dc\Lambda_2}{3l}} \tag{341}$$

此时，$f(\delta)$的最小值为：

$$f(\delta^*) = O(T^{3/4}) \tag{342}$$

如果我们想加速收敛，则可将α的取值与δ相关联。根据上面的结论，当迭代次数T足够大时，必然有$\delta \to 0$。因此，不妨取$\alpha = \frac{\delta}{\Lambda_1}$，代入（8.91）中并利用对钩函数$f(\delta)$的性质，得到：

$$\begin{aligned} \delta^* &= T^{-1/4}\sqrt{\frac{dc\Lambda_1\Lambda_2}{3(l\Lambda_1 + c)}} \\ f(\delta^*) &= O(T^{3/4}) \end{aligned} \tag{343}$$

进一步地，可以发现，δ^*的取值并不唯一，这是因为（8.91）的第二个不等式缩放并非必需。如果取$\delta^* = T^{-1/4}\sqrt{\frac{dc\Lambda_1\Lambda_2}{3l\Lambda_1 + 2c}}$，同样可以得到更紧致的遗憾界，并保证定理的结论不变。

附录

编辑：赵志民, 李一飞

范数

范数（norm）是数学中用于为向量空间中的每个非零向量分配严格正长度或大小的函数。几何上，范数可理解为向量的长度或大小。例如，绝对值是实数集上的一种范数。与之相对的是半范数（seminorm），它可以将非零向量赋予零长度。

向量空间上的半范数需满足以下条件：

1. **半正定性（非负性）**：任何向量的范数总是非负的，对于任意向量 v，$\|v\| \geq 0$。
2. **可伸缩性（齐次性）**：对于任意标量 a 和任何向量 v，标量乘法 av 的范数等于标量的绝对值乘以向量的范数，即 $\|av\| = |a|\|v\|$。
3. **次可加性（三角不等式）**：对于任何向量 v 和 w，向量和 $u = v + w$ 的范数小于或等于向量 v 和 w 的范数之和，即 $\|v + w\| \leq \|v\| + \|w\|$。

范数在具备上述半范数特性的基础上，还要求：对于任意向量 v，若 $\|v\| = 0$，则 v 必须为零向量。换句话说，所有范数都是半范数，但它们可以将非零向量与零向量区分开来。

常用的向量范数包括：

1. ℓ_0 **范数**：向量 x 中非零元素的个数，表示为 $\|x\|_0 = \sum_{i=1}^{n} \mathbb{I}(x_i \neq 0)$。
2. ℓ_1 **范数**：向量 x 中各元素绝对值之和，表示为 $\|x\|_1 = \sum_{i=1}^{n} |x_i|$。
3. ℓ_2 **范数（欧几里得范数）**：向量 x 各元素绝对值的平方和再开平方，表示为 $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$。
4. ℓ_p **范数**：向量 x 各元素绝对值的 p 次方和再开 p 次方，表示为 $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$。
5. ℓ_∞ **范数（极大范数）**：向量 x 中各元素绝对值的最大值，表示为 $\|x\|_\infty = \max_{i=1,\cdots,n} |x_i|$。
6. **加权范数**：设 A 为 n 阶 Hermite 正定矩阵，则向量 x 的加权范数定义为 $\|x\|_A = \sqrt{x^T A x}$。此类范数在本书第 8.3.2 和 8.4.2 节中经常使用。

凸集合

凸集合（convex set）是向量空间（如欧几里得空间）中的一个子集，对于集合中的任意两点，连接它们的线段完全位于该集合内。换句话说，若一个集合包含了连接集合内任意两点的线段上的所有点，则该集合是凸集合。

形式化地说，考虑向量空间 $\mathcal{V}$。若对于该空间中的任意两点 x 和 y，以及满足 $\alpha \in [0,1]$ 的任意标量 α，点 $\alpha x + (1-\alpha)y$ 也属于 $\mathcal{D}$，那么集合 $\mathcal{D} \subseteq \mathcal{V}$ 是凸集合。

凸集合具有非扩张性（non-expansiveness），即对于集合内的任意两点，连接这两点的线段完全包含在集合内。这种性质使得凸集合在许多数学环境中易于处理，特别是在优化问题中：在凸集合中找到的最小值或最大值必为全局值，没有局部最小值或最大值，从而简化了搜索过程。

不仅凸集合具有非扩张性，映射到凸集合的投影操作也是非扩张的，即两点在凸集合上的投影之间的距离不大于两点本身之间的距离。形式上，对于闭合凸集合 $K \subseteq \mathbb{R}^D$，投影算子 $\Pi : \mathbb{R}^D \to K$ 定义为：

$$\Pi(x) = \arg\min_{y \in K} \|x - y\|_2 \tag{344}$$

即将一个向量映射到最接近它的凸集合中的点。投影算子 Π 在 ℓ_2 范数下是非扩张的，即对于任意 $x, x' \in \mathbb{R}^D$，有：

$$\|\Pi(x) - \Pi(x')\|_2 \leq \|x - x'\|_2 \tag{345}$$

该性质证明如下：
令 $y = \Pi(x)$，易知 x 和 K 分处于通过 y 的超平面 $H = \{z \in \mathbb{R}^D : \langle z - y, x - y \rangle = 0\}$ 的两侧。因此，对于 K 中的任意 u，有以下不等式成立：

$$\langle x - y, u - y \rangle \leq 0 \tag{346}$$

同理，令 $y' = \Pi(x')$，对于 K 中的任意 u'，有以下不等式成立：

$$\langle x' - y', u' - y' \rangle \leq 0 \tag{347}$$

此时，令 $u = y'$ 且 $u' = y$，则有：

$$\langle x - y, y' - y \rangle \leq 0 \langle x' - y', y - y' \rangle \leq 0 \tag{348}$$

将两个不等式相加可得：

$$\langle (x - x') + (y' - y), y' - y \rangle \leq 0 \tag{349}$$

根据 Cauchy-Schwarz 不等式，有：

$$\begin{aligned} &\|y - y'\|_2^2 \leq \langle x - x', y - y' \rangle \leq \|x - x'\|_2 \|y - y'\|_2 \\ \Rightarrow &\|y - y'\|_2 \leq \|x - x'\|_2 \\ \Rightarrow &\|\Pi(x) - \Pi(x')\|_2 \leq \|x - x'\|_2 \end{aligned} \tag{350}$$

这种投影映射经常用于凸优化中，因为它能将问题简化为凸优化问题，从而提高算法效率，并在许多情况下保证全局最优解。

Hessian 矩阵

Hessian 矩阵 H_f 是由函数 $f(x)$ 的二阶偏导数组成的方阵，即：

$$\mathbf{H}_f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \, \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \, \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \, \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \, \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \, \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \, \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}. \tag{351}$$

其中，$x = [x_1, x_2, \cdots, x_n]$。

凸函数

凸函数（convex function）是定义在凸集上的实值函数，满足以下性质：对于定义域内的任意两个点 x 和 y 以及满足 $\alpha \in [0,1]$ 的任意标量 α，函数图像上这两点之间的线段位于或位于函数图像上方，即：

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \tag{352}$$

该不等式被称为凸性条件。

除了上述定义，凸函数还有以下几种等价的定义方式：

1. **一阶条件**：若一个定义在凸集上的函数 $f(x)$ 满足下述条件：

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) \tag{353}$$

其中，$\nabla f(x)$ 表示函数 $f(x)$ 在点 x 处的梯度。几何上，这意味着函数的图像位于任意一点处的切线之上。

2. **二阶条件**：若函数 $f(x)$ 是二次可微的，则它是凸函数当且仅当其 Hessian 矩阵 H_f 在其定义域内的所有点 x 上都是半正定的（即矩阵的所有特征值均为非负）。

3. **Jensen 不等式**：若 $f(x)$ 是凸函数，则对于定义域内的任意一组点 $x_1, x_2, \cdots, x_n$ 和归一化的非负权重 $w_1, w_2, \cdots, w_n$，即 $\sum_{i=1}^n w_i = 1$，有：

$$f(\sum_{i=1}^n w_i x_i) \leq \sum_{i=1}^n w_i f(x_i) \tag{354}$$

4. **上图集定义**：凸函数与凸集合的概念密切相关。函数 f 是凸函数，当且仅当其上图集（epigraph）是一个凸集。上图集是位于函数图像上方的点的集合，定义为：

$$epi(f) = \{(x,y) | x \in dom(f),\ y \geq f(x)\} \tag{355}$$

其中，$dom(f)$ 是函数 f 的定义域。

凸函数的一些特性包括：

1. **正比例性质**：若函数 $f(x)$ 是凸函数，则对于任意常数 $\alpha > 0$，函数 $\alpha f(x)$ 也是凸函数。
2. **正移位性质**：若函数 $f(x)$ 是凸函数，则对于任意常数 $c > 0$，函数 $f(x) - c$ 也是凸函数。
3. **加法性质**：若 $f(x)$ 和 $g(x)$ 均为凸函数，则它们的和 $f(x) + g(x)$ 也是凸函数。

凹函数

凹函数（concave function）的定义与凸函数相反。对于其定义域内的任意两个点 x 和 y 以及满足 $\alpha \in [0,1]$ 的任意标量 α，满足以下不等式：

$$f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y) \tag{356}$$

此不等式被称为凹性条件。

其他定义与凸函数类似，这里不再赘述。值得注意的是，若函数 $f(x)$ 为凹函数，则 $-f(x)$ 为凸函数。因此，可以将凹函数问题转化为凸函数问题，从而利用凸函数的性质来求解凹函数问题。

强凸函数

若$f(x)$为定义在凸集上的强凸函数，则对于任意$x, y \in dom(f)$，$\alpha \in [0,1]$，存在$\lambda > 0$，使得：

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\lambda}{2}\alpha(1-\alpha)\|x-y\|_2^2 \tag{357}$$

此时，称 $f(x)$ 为 λ-强凸（strongly convex）函数，其中 λ 为强凸系数。

强凸函数的其他等价定义包括：

1. **Hessian 矩阵条件**：若一个两次可微的函数 $f(x)$ 的 Hessian 矩阵 H_f 在凸集中的所有 x 处均为正定的（即矩阵的所有特征值为正），则该函数是强凸的。
2. **梯度条件**：若一个可微函数 $f(x)$ 是强凸的，则存在一个常数 m，使得对于凸集中的任意 x, y，有 $\|\nabla f(x) - \nabla f(y)\|_2 \geq m\|x - y\|_2$。其中，$\nabla f(x)$ 表示 $f(x)$ 在点 x 处的梯度。

直观上，对于强凸函数 $f(x)$，可以在任意一点处构造一个二次函数作为其下界。这一性质使得优化算法更加高效，并具有类似于 **90页** 中定理 7.2 的良好性质。

以下给出定理 7.2 的证明：

根据强凸函数的定义，取 $x = w$，$y = w^*$，然后两边除以 α，可得：

$$\begin{aligned} &\frac{f(\alpha w + (1-\alpha)w^*)}{\alpha} \leq f(w) + \frac{1-\alpha}{\alpha}f(w^*) - \frac{\lambda}{2}(1-\alpha)\|w - w^*\|_2^2 \\ \Rightarrow &\frac{\lambda}{2}(1-\alpha)\|w - w^*\|_2^2 \leq f(w) - f(w^*) - \frac{f(w^* + (w - w^*)\alpha) - f(w^*)}{\alpha} \end{aligned} \tag{358}$$

令 $\alpha \to 0^+$，则有：

$$\begin{aligned} &\lim_{\alpha \to 0^+} \frac{\lambda}{2}(1-\alpha)\|w - w^*\|_2^2 \leq f(w) - f(w^*) + \lim_{\alpha \to 0^+} \frac{f(w^* + (w - w^*)\alpha) - f(w^*)}{\alpha} \\ \Rightarrow &\frac{\lambda}{2}\|w - w^*\|_2^2 \leq f(w) - f(w^*) + \nabla f(w^*)^T(w - w^*) \end{aligned} \tag{359}$$

其中 $\Delta = (w - w^*)\alpha$。

由于 w^* 为最优解，因此 $\nabla f(w^*) = 0$，则有：

$$f(w) - f(w^*) \geq \frac{\lambda}{2}\|w - w^*\|_2^2 \tag{360}$$

指数凹函数

若函数 $f(x)$ 的指数 $\exp(f(x))$ 为凹函数，则称 $f(x)$ 为指数凹（exponentially concave）函数。注意，当 $\exp(f(x))$ 是凹函数时，$f(x)$ 本身不一定是凹函数。 若 $f(x)$ 为指数凹函数，则 $\exp(-f(x))$ 必为凸函数。因此，指数凹是一种弱于强凸但强于凸的性质。

指数凹函数的一些特性包括：

1. **正比例性质**：若函数 $f(x)$ 为指数凹函数，则对于任意常数 α，函数 $\alpha f(x)$ 也是指数凹函数。
2. **负移位性质**：若函数 $f(x)$ 为指数凹函数，且 c 为常数，则函数 $f(x) - c$ 也是指数凹函数。

指数凹函数提供了一种灵活且富有表现力的方式来建模各种现象。它能捕捉广泛的形状和行为。例如，在凸优化中使用指数凹函数可以加快迭代优化算法（如梯度下降或牛顿法）的收敛速度。因此，指数凹函数在处理概率模型或存在不确定性的场景中具有重要意义，特别是在限制或量化不确定性方面。

凸优化

凸优化（convex optimization）是优化理论的一个分支，研究的是在凸函数的凸集上进行优化的问题。凸优化的目标是在满足一组凸约束条件的情况下，找到凸目标函数的最小值。

一般形式的凸优化问题可以表示为：

$$
\begin{aligned}
\min \quad & f_0(x) \\
s.t. \quad & f_i(x) \leq 0, i \in [m] \\
& g_j(x) = 0, j \in [n]
\end{aligned}
\tag{361}
$$

其中，$f_0(x)$ 是凸目标函数，$f_i(x)$ 是凸不等式约束条件，$g_j(x)$ 是仿射等式约束条件。

凸优化具有以下有利特性，使其成为一个被广泛研究和应用的领域：

1. **全局最优性**：凸优化问题的一个关键性质是，任何局部最小值也是全局最小值。此性质确保凸优化算法找到的解是给定凸集中的最优解。
2. **高效算法**：凸优化拥有多项式时间内找到最优解的高效算法。这些算法基于凸目标函数和约束条件的凸性，能够有效解决复杂的优化问题。
3. **广泛应用**：凸优化在工程学、金融学、机器学习、运筹学和信号处理等领域有着广泛的应用。它被用于解决如投资组合优化、信号重构、资源分配和机器学习模型训练等问题。凸优化技术，如线性规划、二次规划和半定规划，构成了许多优化算法的基础，为高效解决复杂优化问题提供了强大工具。

以下证明凸函数任何局部最优解均为全局最优解的性质。

假设 $f(x)$ 是凸函数，x^* 是 f 在凸集合 $\mathcal{D}$ 中的局部最优解。由于凸集的性质，对于任意 y，$y - x^*$ 是一个可行方向。因此，总可以选择足够小的 $\alpha > 0$，使得：

$$
f(x^*) \leq f(x^* + \alpha(y - x^*))
\tag{362}
$$

由 f 的凸性可得:

$$
f(x^* + \alpha(y - x^*)) = f((1 - \alpha)x^* + \alpha y) \leq (1 - \alpha)f(x^*) + \alpha f(y)
\tag{363}
$$

结合以上两式，可得：

$$
\begin{aligned}
f(x^*) &\leq (1 - \alpha)f(x^*) + \alpha f(y) \\
rightarrow f(x^*) &\leq f(y)
\end{aligned}
\tag{364}
$$

由于 y 是凸集合 $\mathcal{D}$ 中的任意点，故 x^* 为全局最优解。对于 $f(x)$ 的全局最大解，可以通过考虑函数 $-f(x)$ 的局部最优解得到类似的结论。

仿射

仿射变换（Affine transformation），又称仿射映射，是指在几何中，对一个向量空间进行一次线性变换并加上一个平移，变换为另一个向量空间。若该线性映射被表示为矩阵 A，平移被表示为向量 $\vec{b}$，则仿射映射 f 可表示为：

$$
\vec{y} = f(\vec{x}) = A\vec{x} + \vec{b}
\tag{365}
$$

其中，A 被称为仿射变换矩阵或投射变换矩阵。

仿射变换具有以下性质：

1. **点之间的共线性**：在同一条直线上的三个或更多的点（即共线点）在变换后依然位于同一条直线上（共线）。
2. **直线的平行性**：两条或以上的平行直线在变换后仍保持平行。
3. **集合的凸性**：凸集合在变换后依然是凸集合，且最初的极值点被映射到变换后的极值点集。
4. **平行线段的长度比例恒定**：两条由点 p_1, p_2, p_3, p_4 定义的平行线段，其长度比例在变换后保持不变，即 $\frac{\overrightarrow{p_1p_2}}{\overrightarrow{p_3p_4}} = \frac{\overrightarrow{f(p_1)f(p_2)}}{\overrightarrow{f(p_3)f(p_4)}}$。
5. **质心位置恒定**：不同质量的点组成集合的质心位置在仿射变换后保持不变。

仿射集（affine set）是指欧氏空间 R^n 中具有以下性质的点集 S：对于任意 $x, y \in S$，以及 $\forall \lambda \in [0, 1]$，有 $(1 - \lambda)x + \lambda y \in S$。容易证明，包含原点的仿射集 S 是 R^n 的子空间。

仿射包（affine hull/span）是包含集合 S 的所有仿射集的交集，也是集合 S 中元素通过不断连接直线所形成的所有元素的集合。仿射包是包含集合 S 的最小仿射集，记为 $aff(S)$，即：

$$aff(S) = \{\sum_{i=1}^{k} \alpha_i x_i \mid k > 0, x_i \in S, \alpha_i \in R, \sum_{i=1}^{k} \alpha_i = 1\} \tag{366}$$

仿射包具有以下性质：

1. $aff(aff(S)) = aff(S)$
2. $aff(S + T) = aff(S) + aff(T)$
3. 若 S 为有限维度，则 $aff(S)$ 为闭集合。

Slater条件/定理

关于强对偶性的讨论，**11页** 已给出了详细说明，此处不再赘述。这里着重讨论 **11页** 左下角附注提到的 Slater 条件，即：

存在一点 $x \in \text{relint}(D)$，该点称为 Slater 向量，有：

$$f_i(x) < 0, \quad i \in [m] \tag{367}$$

其中，$D = \bigcap_0^m \text{dom}(f_i)$，$relint(D)$ 为 D 的相对内部，即其仿射包的内部所有点，即 $relint(D) = \text{int}(aff(D))$。

当满足 Slater 条件且原始问题为凸优化问题时：

1. 强对偶性成立。
2. 对偶最优解集合非空且有界。

这就是 Slater 定理。

证明

首先证明对偶间隙（Duality Gap）为零，即原始问题与对偶问题的目标函数值之差 $p^* - d^* = 0$。考虑集合 $\mathcal{V} \subset \mathbb{R}^m \times \mathbb{R}$，满足：

$$\mathcal{V} := \{(u, w) \in \mathbb{R}^m \times \mathbb{R} \mid f_0(x) \leq w, f_i(x) \leq u_i, \forall i \in [m], \forall x\} \tag{368}$$

集合 $\mathcal{V}$ 具有以下性质：

1. 它是凸集合，可由 f_i，$i \in \{0\} \cup [m]$ 的凸性质得出。
2. 若 $(u, w) \in \mathcal{V}$，且 $(u', w') \succeq (u, w)$，则 $(u', w') \in \mathcal{V}$。

易证向量 $(0, p^*) \notin int(\mathcal{V})$，否则一定存在 $\varepsilon > 0$，使得 $(0, p^* - \varepsilon) \in int(\mathcal{V})$，这明显与 p^* 为最优解矛盾。因此，必有 $(0, p^*) \in \partial\mathcal{V}$ 或 $(0, p^*) \notin \mathcal{V}$。应用支撑超平面定理（定理 23），可知存在一个非零点 $(\lambda, \lambda_0) \in \mathbb{R}^m \times \mathbb{R}$，满足以下条件：

$$(\lambda, \lambda_0)^T (u, w) = \lambda^T u + \lambda_0 w \geq \lambda_0 p^*, \forall (u, w) \in \mathcal{V} \tag{369}$$

在此情况下，必然有 $\lambda \succeq 0$ 和 $\lambda_0 \geq 0$。这是因为，若 λ 和 λ_0 中的分量出现任何负数，根据集合 $\mathcal{V}$ 的性质二，(u, w) 的分量可以在集合 $\mathcal{V}$ 内取得任意大的值，从而导致上式不一定成立。

因此，只需考虑两种情况：

1. $\lambda_0 = 0$：此时根据上式，可得

$$\inf_{(u,w) \in \mathcal{V}} \lambda^T u = 0 \tag{370}$$

另一方面，根据 $\mathcal{V}$ 的定义，$\lambda \succeq 0$ 且 $\lambda \neq 0$，可得：

$$\inf_{(u,w)\in\mathcal{V}} \lambda^T u = \inf_x \sum_{i=1}^m \lambda_i f_i(x) \leq \sum_{i=1}^m \lambda_i f_i(\bar{x}) < 0 \tag{371}$$

其中，$\bar{x}$ 是 Slater 向量，而最后一个不等式依据 Slater 条件得出。此时，两个结论互相矛盾，因此 $\lambda_0 \neq 0$。

2. $\lambda_0 > 0$：对上式左右两边除以 λ_0，得：

$$\inf_{(u,w)\in\mathcal{V}} \{\tilde{\lambda}^T u + w\} \geq p^* \tag{372}$$

其中，$\tilde{\lambda} := \frac{\lambda}{\lambda_0} \succeq 0$。

考虑拉格朗日函数 $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$：

$$L(x, \tilde{\lambda}) := f_0(x) + \sum_{i=1}^m \tilde{\lambda}_i f_i(x) \tag{373}$$

其对偶函数为：

$$g(\tilde{\lambda}) := \inf_x L(x, \tilde{\lambda}) \geq p^* \tag{374}$$

其对偶问题为：

$$\max_{\lambda} g(\lambda), \lambda \succeq 0 \tag{375}$$

因此，可得 $d^* \geq p^*$。根据弱对偶性，$d^* \leq p^*$，从而推断出 $d^* = p^*$。

接着证明对偶问题最优解集合非空且有界。对于任意对偶最优解 $\tilde{\lambda} \succeq 0$，有：

$$\begin{aligned} d^* = g(\tilde{\lambda}) &= \inf_x \{f_0(x) + \sum_{i=1}^m \tilde{\lambda}_i f_i(x)\} \\ &\leq f_0(\bar{x}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\bar{x}) \\ &\leq f_0(\bar{x}) + \max_{i\in[m]}\{f_i(\bar{x})\}[\sum_{i=1}^m \tilde{\lambda}_i] \end{aligned} \tag{376}$$

因此，有：

$$\min_{i\in[m]}\{-f_i(\bar{x})\}[\sum_{i=1}^m \tilde{\lambda}_i] \leq f_0(\bar{x}) - d^* \tag{377}$$

进而得出：

$$\|\tilde{\lambda}\| \leq \sum_{i=1}^m \tilde{\lambda}_i \leq \frac{f_0(\bar{x}) - d^*}{\min_{i\in[m]}\{-f_i(\bar{x})\}} < \infty \tag{378}$$

其中，最后一个不等式依据 Slater 条件得出。□

KKT条件

KKT条件（Karush-Kuhn-Tucker条件）在凸优化领域具有至关重要的地位。虽然在**12-13页** 中对其进行了基本解释，此处将进行更为深入的分析。KKT条件中的符号 λ_i，$i \in [m]$ 和 μ_i，$i \in [n]$ 被视为 KKT 乘子。特别地，当 $m = 0$ 时，即不存在不等式约束条件时，KKT条件退化为拉格朗日条件，此时 KKT 乘子也被称为拉格朗日乘子。

证明

首先，对于 $x^*, (\mu^*, \lambda^*)$ 满足 KKT 条件等价于它们构成一个纳什均衡。

固定 (μ^*, λ^*)，并变化 x，均衡等价于拉格朗日函数在 x^* 处的梯度为零，即主问题的稳定性（stationarity）。

固定 x，并变化 (μ^*, λ^*)，均衡等价于主问题的约束（feasibility）和互补松弛条件。

充分性：若解对 $x^*, (\mu^*, \lambda^*)$ 满足 KKT 条件，则它们构成一个纳什均衡，从而消除对偶间隙。

必要性：任意解对 $x^*, (\mu^*, \lambda^*)$ 必然消除对偶间隙，因此它们必须构成一个纳什均衡，从而满足 KKT 条件。□

在此对 KKT 和 Slater 条件进行区分：

1. **KKT条件** 是一组用于确定约束优化问题中解的最优性的条件。它们通过将约束纳入条件，扩展了无约束优化中设定目标函数梯度为零的思路到约束优化问题中。
   Slater条件 是凸优化中确保强对偶性的特定约束条件，即主问题和对偶问题最优解的等价性。
2. KKT条件包括对偶问题的约束、互补松弛条件、主问题约束和稳定性。它们整合了目标和约束函数的梯度以及 KKT 乘子，以形成最优性条件。
   Slater 条件要求存在一个严格可行点，即严格满足所有不等式约束的点。
3. 当点满足 KKT 条件时，表明问题的局部最优解已找到。这些条件弥合了主问题和对偶问题之间的差距，对于分析和解决约束优化问题至关重要。
   满足 Slater 条件时，确保凸优化问题的强对偶性，对于简化和解决这些问题至关重要。Slater 条件并不直接提供最优性条件，但为强对偶性铺平了道路，之后可以利用强对偶性寻找最优解。
4. **KKT条件** 较为通用，适用于更广泛的优化问题类别，包括非凸问题。
   Slater条件 则特定于凸优化问题，用于确保这些问题中的强对偶性。
5. 对于凸且可微的问题，满足 KKT 条件意味着最优性和强对偶性。相反，最优性和强对偶性意味着所有问题的 KKT 条件得到满足。
   当 Slater 条件成立时，KKT 条件是最优解的充要条件，此时强对偶性成立。

KKT条件和 Slater 条件通常被归类为“正则条件”（regularity condition）或“约束资格”（constraint qualification）。这些条件为优化问题提供了一个结构化的框架，以便在约束情况下分析和确定解的最优性。更多的正则条件详见参考文献：On regularity conditions in mathematical programming。

偏序集

序理论（Order Theory）是数学的一个分支，它的核心思想是通过定义某种“序”来描述元素之间的相对关系。在序理论中，一个偏序集（partial order set，简称 poset）包含一个非空集合 P 和一个满足特定条件的二元关系 $\leq$。这个二元关系称为偏序关系，它必须满足以下三个条件：

1. **自反性（Reflexivity）**：对于 P 中的任意元素 a，都有 $a \leq a$。
2. **反对称性（Antisymmetry）**：对于 P 中的任意元素 a 和 b，如果 $a \leq b$ 且 $b \leq a$，那么 $a = b$。
3. **传递性（Transitivity）**：对于 P 中的任意元素 a、b 和 c，如果 $a \leq b$ 且 $b \leq c$，那么 $a \leq c$。

这些条件定义了偏序关系，使其与全序（total order）关系不同。在偏序集中，可能存在某些元素是不可比较的，即对于 P 中的某些 a 和 b，既不满足 $a \leq b$，也不满足 $b \leq a$。

上下界

上界（upper bound 或 majorant）是与偏序集有关的特殊元素，指偏序集中大于或等于其子集中一切元素的元素。若数集 S 为实数集 R 的子集且有上界，则显然有无穷多个上界，其中最小的上界常常具有重要作用，称为数集 S 的上确界（tight upper bound 或 supremum）。同理，可以定义下界（lower bound 或 minorant）和下确界（tight lower bound 或 infimum）。

尾界

尾界（tail bound）是指给定一个随机变量，其概率分布尾部部分的界限。上尾界（upper tail bound）描述随机变量在其分布上尾处的概率上限，而下尾界（lower tail bound）描述随机变量在其分布下尾处的概率上限。Chebyshev 不等式、Hoeffding 不等式和 Bernstein 不等式都是尾界的例子，它们提供了随机变量偏离其期望值的概率界限。

置信界

置信界（confidence bound）是在估计一个未知参数时，给出一个包含该参数的区间，并且这个区间具有特定的置信水平。例如，一个95%的置信区间意味着我们有95%的信心该区间包含真实的参数值。置信界可以是上置信界（upper confidence bound），下置信界（lower confidence bound），或同时包含上下界的置信区间（confidence interval）。上置信界提供对参数估计的可能最大值的上限，下置信界提供对参数估计的可能最小值的下限。

连续性

连续性（continuity）表示函数在某处的变化不会突然中断或跳跃。形式上，如果函数 $f(x)$ 在 $x = a$ 处满足以下条件，则称其在该点连续：

1. 函数 $f(x)$ 在 $x = a$ 处有定义。
2. 当 x 趋近于 a 时，$f(x)$ 的极限存在且等于 $f(a)$。

连续性意味着输入的微小变化导致输出的微小变化。如果一个函数在其定义域的每个点上都是连续的，则称其为连续函数。

Lipschitz 连续性是连续性的更强形式，它要求函数在变化速度方面有界。具体而言，如果存在一个常数 L，使得函数在任意两点的函数值之间的绝对差小于等于 L 乘以两点之间的距离，则称该函数为 $L - Lipschitz$ 连续，即：

$$\forall x, y \in \mathrm{dom}(f),\ \exists L > 0 \text{ 使得 } \|f(x) - f(y)\|_2 \leq L\|x - y\|_2 \tag{379}$$

其中，L 称为 Lipschitz 常数，表示函数的最大变化率。若 L 较大，函数可以快速变化；若 L 较小，函数变化更渐进。

事实上，如果一个函数的导数有界，那么它一定是 Lipschitz 连续的；反之，如果一个可微函数是 Lipschitz 连续的，那么它的导数一定有界。

证明如下：

1. 若函数 $f(x)$ 的导数有界，即存在常数 $L \geq 0$，使得对于任意 x，有 $|f'(x)| \leq L$。根据微分中值定理，对于任意 $x \leq y$，存在 $c \in [x, y]$，使得：

$$\begin{aligned} \|f(x) - f(y)\|_2 &= \|f'(c)\|_2\|x - y\|_2 \\ \Rightarrow \|f(x) - f(y)\|_2 &\leq L\|x - y\|_2 \end{aligned} \tag{380}$$

此时，函数是 $L - Lipschitz$ 连续的。

2. 若函数 $f(x)$ 是 $L - Lipschitz$ 连续的，即对于任意 x, y，有

$$\|f(x) - f(y)\|_2 \leq L\|x - y\|_2 \tag{381}$$

根据微分中值定理，对于任意 $x \leq y$，存在 $c \in [x, y]$，使得：

$$\|f(x) - f(y)\|_2 = \|f'(c)\|_2\|x - y\|_2 \tag{382}$$

不妨令 $x \to y$，则 $c \to y$。因为 $f(y)$ 可微，可得：

$$\|f'(y)\|_2 = \|\lim_{x \to y} \frac{f(x) - f(y)}{x - y}\|_2 = \lim_{x \to y} \frac{\|f(x) - f(y)\|_2}{\|x - y\|_2} \leq \lim_{x \to y} L = L \tag{383}$$

因为 y 的任意性，所以函数的导数有界。

连续性关注函数图像中跳跃或中断的缺失，而 Lipschitz 连续性关注函数的变化速度。因此，Lipschitz 连续性是比连续性更严格的条件。一个连续函数不一定是 Lipschitz 连续的，因为连续性不要求函数变化速度有界。然而，一个 Lipschitz 连续的函数必然是

连续的，因为 Lipschitz 连续性蕴含连续性。

Lipschitz 连续性的性质在数学的各个领域中广泛应用，如分析、优化和微分方程研究。它在保证某些数学问题的解的存在性、唯一性和稳定性方面起着关键作用。

光滑性

在数学分析中，函数的光滑性（smoothness）通过函数在某个域（称为可微性类）上的连续导数的数量来衡量。最基本的情况下，如果一个函数在每个点上都可导（因此连续），则可以认为它是光滑的。 一方面，光滑性确保了梯度下降等优化算法能够更快收敛，并减少可能遇到的梯度震荡或发散的情况。 另一方面，光滑性提供了函数曲率的信息，从而帮助设计更有效的优化算法，如加速梯度下降法或牛顿法。

在优化理论中，L-光滑函数是指它的梯度具有 L-Lipschitz 连续性，这意味着函数的梯度在其定义域中的变化速率被 L 所限制。形式上，对于任意 $x, y \in \mathbb{R}^n$，存在 $L > 0$，使得：

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \tag{384}$$

或者等价地，

$$\|\nabla^2 f(x)\|_2 \leq L \tag{385}$$

或者等价地，

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2 \tag{386}$$

以上三种定义方式是等价的，且 L 被称为光滑系数。 由定义3，我们可以看出，在光滑函数的任意一点处都可以构造一个二次函数作为其上界。

接下来我们证明这些定义的等价性。首先，我们证明定义1可以推导出定义2。

考虑函数 f 的梯度 $\nabla f(x)$ 的二阶泰勒展开：

$$\nabla f(y) = \nabla f(x) + \nabla^2 f(\xi)(y - x) \tag{387}$$

其中 ξ 是 x 和 y 之间的一点，$\nabla^2 f(\xi)$ 表示在点 ξ 处的 Hessian 矩阵。

根据 L-光滑性的定义1，我们有：

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2 \tag{388}$$

将二阶泰勒展开的结果代入其中：

$$\|\nabla^2 f(\xi)(y - x)\|_2 \leq L\|y - x\|_2 \tag{389}$$

对于任意的非零向量 $v = y - x$，定义：

$$v' = \frac{v}{\|v\|_2} \tag{390}$$

我们得到：

$$\|\nabla^2 f(\xi)v'\|_2 \leq L \tag{391}$$

由于 v' 是一个单位向量，这意味着 Hessian 矩阵 $\nabla^2 f(\xi)$ 作用在任意单位向量上时的范数不超过 L，因此 Hessian 矩阵的谱范数（即最大特征值的绝对值）满足：

$$\|\nabla^2 f(\xi)\|_2 \leq L \tag{392}$$

其中，由于 ξ 是 x 和 y 之间的一点，因此我们可以将上述结论推广到整个定义域。

接下来我们证明定义2可以推导出定义3。由定义2，给定 f 是 L-光滑的，对任意的 $x, y \in \mathbb{R}^n$，我们有：

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2 \tag{393}$$

将定义中的 x 和 y 互换，得到：

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|_2^2 \tag{394}$$

将两个不等式相加可得：

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|_2^2 \tag{395}$$

注意到不等式左侧的内积无论如何取值，该不等式均成立。 根据 Cauchy-Schwarz 不等式，当 $y - x$ 与 $\nabla f(x) - \nabla f(y)$ 平行时左侧内积取到最大值，即 $\|\nabla f(x) - \nabla f(y)\|_2\|x - y\|_2$，代入可得：

$$\|\nabla f(x) - \nabla f(y)\|_2\|x - y\|_2 \leq L\|x - y\|_2^2 \tag{396}$$

化简后即得证。

这里对光滑性和 $Lipschitz$ 连续性进行一些比较：

- $Lipschitz$ 连续性关注的是函数值变化的速度，即函数值的“陡峭程度”，而光滑性关注的是梯度变化的速度，即函数的“曲率”或二阶变化。
- $Lipschitz$ 连续性表示函数变化不会太快，确保函数的整体平滑性，而光滑性表示梯度变化不会太快，确保函数曲面没有急剧的弯曲。

次梯度

次梯度（subgradient）是凸函数导数的推广形式。某些凸函数在特定区域内可能不存在导数，但我们依旧可以用次梯度来表示该区域内函数变化率的下界。形式上，对于凸函数 $f(x)$，在任意点 x_0 处的次梯度 c 必须满足以下不等式：

$$f(x) - f(x_0) \geq c(x - x_0) \tag{397}$$

根据微分中值定理的逆命题，c 通常在 $[a, b]$ 之间取值，其中 a, b 是函数 $f(x)$ 在 x_0 处的左右导数，即：

$$a = \lim_{x \to x_0^-} \frac{f(x) - f(x_0)}{x - x_0},\ b = \lim_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0} \tag{398}$$

此时，次梯度 c 的集合 $[a, b]$ 被称为次微分，即 $\partial f(x_0)$。当 $a = b$ 时，次梯度 c 退化为导数。

次梯度在机器学习领域广泛应用，特别是在训练支持向量机（SVM）和其他具有非可微损失函数的模型中。它们还构成了随机次梯度方法的基础，这些方法在处理大规模机器学习问题时非常有效。

对偶空间

线性泛函（linear functional）是指从向量空间 V 到对应标量域 k 的线性映射，满足加法和数乘的性质，即对于任意向量 $x, y \in V$ 和标量 $\alpha \in k$，有：

$$\begin{aligned} f(x + y) &= f(x) + f(y) \\ f(\alpha x) &= \alpha f(x) \end{aligned} \tag{399}$$

所有从 V 到 k 的线性泛函构成的集合称为 V 的对偶空间（dual space），记为 $V^* = \mathrm{Hom}_k(V, k)$，对偶空间中的元素称为对偶向量。

Legendre变换

将函数转换为另一种函数，常用于改变其定义域和属性，使问题更简单或更易分析。Legendre 变换（Legendre transform）常用于将一组独立变量转换为另一组独立变量，特别是在经典力学和热力学中。以下是 Legendre 变换的基本概念和步骤：

1. **定义函数**：假设有一个凸函数 $f(x)$，其自变量为 x。
2. **定义共轭变量**：定义新的变量 p，它是原函数 $f(x)$ 的导数，即 $p = \frac{df(x)}{dx}$。
3. **定义共轭函数**：定义新的函数 $g(p)$，其形式为：$g(p) = x \cdot p - f(x)$。这里，x 是 $f(x)$ 的自变量，同时也是 $g(p)$ 的隐含变量。
4. **变换关系**：通过 Legendre 变换，从原来的函数 $f(x)$ 得到新的函数 $g(p)$，这个新的函数 $g(p)$ 依赖于共轭变量 p。

共轭函数

凸共轭（convex conjugate）是 Legendre 变换的一种推广，因此也被称为 Legendre-Fenchel 变换（Legendre-Fenchel transform）。通过凸共轭变换，原函数可以转换为凸函数，从而利用凸函数的性质来解决原问题。

形式上，对于函数 $f(x)$，其共轭函数 $f^*(y)$ 定义为：

$$f^*(y) = \sup_{x \in \mathrm{dom}(f)} (y^T x - f(x)) \tag{400}$$

其中，$\mathrm{dom}(f)$ 是函数 $f(x)$ 的定义域。

共轭函数具有以下一些有用的性质：

1. **凸性**：函数 $f(x)$ 的共轭函数 $f^*(y)$ 一定是凸函数。证明如下：

$$\begin{aligned} f^*(\lambda y_1 + (1-\lambda) y_2) &= \sup_{x \in \mathrm{dom}(f)} \{x^T(\lambda y_1 + (1-\lambda) y_2) - f(x)\} \\ &\leq \lambda \sup_{x \in \mathrm{dom}(f)} \{x^T y_1 - f(x)\} + (1-\lambda) \sup_{x \in \mathrm{dom}(f)} \{x^T y_2 - f(x)\} \\ &= \lambda f^*(y_1) + (1-\lambda) f^*(y_2) \end{aligned} \tag{401}$$

其中的不等式利用了凸性的性质。

2. **逆序性**：对于定义域中所有元素 x，若 $f(x) \leq g(x)$，则 $f^*(y) \geq g^*(y)$。证明如下：

由于 $f(x) \leq g(x)$，因此 $x^T y - f(x) \geq x^T y - g(x)$。两边同时取上确界，根据定义有：

$$f^*(y) = \sup_{x \in \mathrm{dom}(f)} \{x^T y - f(x)\} \geq \sup_{x \in \mathrm{dom}(f)} \{x^T y - g(x)\} = g^*(y) \tag{402}$$

3. **极值变换**：若 f 可微，则对于 $\forall y$，有：

$$f^*(y) \leq f^*(\nabla f(x)) = \nabla f^*(x)^T x - f(x) = -[f(x) + \nabla f(x)^T (0 - x)] \tag{403}$$

此性质即书中的（1.10），完整证明如下：

为了在 f^* 的定义中找到上确界，对右侧的 x 求导，并将其设置为零以找到极大值点：

$$\frac{d}{dx}(x^T y - f(x)) = y - \nabla f(x) = 0 \tag{404}$$

此时有 $y = \nabla f(x)$，得证。

σ-代数

σ-代数（或 σ-域）是测度论和概率论中的一个重要概念。σ-代数是一个满足特定封闭性质的集合族，使我们能够对这些集合定义一致的测度（如概率）。具体来说，σ-代数是一个集合族，满足以下三个性质：

1. **包含全集**：如果 $\mathcal{F}$ 是定义在集合 X 上的一个 σ-代数，那么 X 本身属于 $\mathcal{F}$，即 $X \in \mathcal{F}$。
2. **对补集封闭**：如果 A 是 $\mathcal{F}$ 中的一个集合，那么它的补集 $X \setminus A$ 也属于 $\mathcal{F}$，即 $A \in \mathcal{F} \implies X \setminus A \in \mathcal{F}$。

3. **对可数并封闭**：如果 $A_1, A_2, A_3, \ldots$ 是 $\mathcal{F}$ 中的集合，那么它们的可数并集 $\bigcup_{i=1}^{\infty} A_i$ 也属于 $\mathcal{F}$，即 $A_i \in \mathcal{F}$ 对所有 $i \in \mathbb{N}$，则 $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$。

σ-代数在测度论中尤为重要，因为它为定义测度提供了必要的框架。测度是定义在 σ-代数上的集合函数，用于度量集合的“大小”。在概率论中，σ-代数用于定义事件空间，从而定义概率测度。

过滤

σ-代数 $\mathcal{F}$ 是一个固定的集合族，满足特定的封闭性质，表示我们在某一时刻可以知道的所有信息。过滤（filtration）是关于随着时间推移而观察信息的概念，通常与随机过程（stochastic processes）相关。具体来说，过滤是一个按时间参数索引的 σ-代数序列 $\{\mathcal{F}_t\}_{t \in T}$，表示随时间变化的可观测事件的集合，满足以下性质：

1. **每个 $\mathcal{F}_t$ 是一个 σ-代数**：对于每个时刻 t，$\mathcal{F}_t$ 是定义在某个固定集合 X 上的一个 σ-代数。
2. **单调性**：对于任意的 $t_1 \leq t_2$，有 $\mathcal{F}_{t_1} \subseteq \mathcal{F}_{t_2}$。这意味着随着时间的推移，所包含的信息只会增加，不会减少。

鞅

鞅（Martingale）是概率论中的一个重要概念，用于描述某些类型的随机过程。鞅过程的特点是，其未来期望值在已知当前信息的条件下等于当前值。

形式化定义

设 $\{X_t\}$ 是一个随机过程，$\{\mathcal{F}_t\}$ 是一个随时间 t 变化的过滤（即包含随时间增加的所有信息的 σ-代数的序列）。当这个随机过程 $\{X_t\}$ 是鞅时，必须满足以下条件：

1. **适应性（Adaptedness）**：对于每一个 t，X_t 是 $\mathcal{F}_t$-可测的（即 X_t 的值在时间 t 时刻是已知信息的函数）。
2. **积分性（Integrability）**：对于所有 t，$\mathbb{E}[|X_t|] < \infty$。
3. **鞅性质（Martingale Property）**：对于所有 t 和 $s \geq t$，有 $\mathbb{E}[X_s \mid \mathcal{F}_t] = X_t$。这意味着在已知当前时刻 t 的信息 $\mathcal{F}_t$ 条件下，未来某个时刻 s 的期望值等于当前时刻 t 的值。

直观解释

鞅的定义保证了在已知当前信息的条件下，未来值的期望等于当前值，这反映了一种“无偏性”。因此，鞅过程可以被看作是一种“公平游戏”。设想一个赌徒在赌场中进行赌博，如果这个赌徒的资金变化形成一个鞅过程，那么在任何时刻，给定当前的资金情况，未来资金的期望值都是当前的资金，表示没有系统性的赢或输的趋势。

举例说明

考虑一个简单的随机游走过程，其中 $X_{t+1} = X_t + Z_{t+1}$，其中 Z_{t+1} 是一个独立同分布的随机变量，取值为 $+1$ 或 -1，且概率各为 50%。在这种情况下，如果设 $X_0 = 0$，那么 $\{X_t\}$ 是一个鞅，因为每一步的期望值都是零。

鞅的类型

除了标准的鞅，还有两个相关的概念：

1. **超鞅（Submartingale）**：若对于所有 t 和 $s \geq t$，有 $\mathbb{E}[X_s \mid \mathcal{F}_t] \geq X_t$，则称 $\{X_t\}$ 为超鞅（或上鞅）。
2. **亚鞅（Supermartingale）**：若对于所有 t 和 $s \geq t$，有 $\mathbb{E}[X_s \mid \mathcal{F}_t] \leq X_t$，则称 $\{X_t\}$ 为亚鞅（或下鞅）。

一个区分超鞅和亚鞅的记忆方法是：“生活是一个超鞅：随着时间的推进，期望降低。”

鞅差序列

鞅差 D_t 定义为 $D_t = X_t - X_{t-1}$，鞅差序列（Martingale Difference Sequence）$\{D_t\}$ 则满足以下条件：

1. **适应性（Adaptedness）**：对于每一个 t，D_t 是 $\mathcal{F}_t$-可测的。

2. **零条件期望（Zero Conditional Expectation）**：对于所有 t，有 $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] = 0$，即在已知过去信息 $\mathcal{F}_{t-1}$ 的条件下，D_t 的条件期望为零。这意味着当前的观察值不提供对未来观察值的系统性偏差，即每一步的变化是纯随机的。

虽然鞅差序列中的每个元素的条件期望为零，但这并不意味着这些元素是独立的。相反，它们可以有复杂的依赖关系。鞅差序列的关键性质是每个元素在条件期望下为零，这使得它在分析鞅和集中不等式（如 Bernstein 不等式）中非常有用。

KL 散度

KL 散度（Kullback-Leibler 散度），也称为相对熵，是一种用于衡量两个概率分布之间差异的非对称度量，在信息论和统计学中广泛应用。KL 散度衡量的是在使用近似分布时，相比于使用真实分布，所增加的“信息损失”或“不确定性”。

定义

假设有两个概率分布 P 和 Q，它们定义在同一个概率空间上。P 通常被认为是“真实”分布，而 Q 是近似分布。KL 散度 $D_{KL}(P\|Q)$ 表示为：

$$D_{KL}(P\|Q) = \sum_{x} P(x) \ln \frac{P(x)}{Q(x)} \tag{405}$$

对于连续分布：

$$D_{KL}(P\|Q) = \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{q(x)} \, dx \tag{406}$$

其中，$P(x)$ 和 $Q(x)$ 分别是分布 P 和 Q 在 x 处的概率密度函数（或概率质量函数）。

性质

1. **非负性**：KL 散度总是非负的，即 $D_{KL}(P\|Q) \geq 0$，只有当 P 和 Q 完全相同时，KL 散度才为零。

非负性的证明

KL 散度的非负性可以通过 Jensen 不等式来证明。首先，考虑离散情况下的 KL 散度定义：

$$D_{KL}(P\|Q) = \sum_{x} P(x) \ln \frac{P(x)}{Q(x)} \tag{407}$$

由于对数函数是一个凹函数，可以应用 Jensen 不等式。对于凹函数 f 和随机变量 X，有：

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)] \tag{408}$$

将 $f(x) = \ln(x)$，并令 $X = \frac{Q(x)}{P(x)}$。则有：

$$\ln(\mathbb{E}[\frac{Q(x)}{P(x)}]) \geq \mathbb{E}[\ln(\frac{Q(x)}{P(x)})] \tag{409}$$

因为 $\sum_x P(x) = 1$ 且 $Q(x) \geq 0$，所以：

$$\mathbb{E}[\frac{Q(x)}{P(x)}] = \sum_x P(x) \frac{Q(x)}{P(x)} = \sum_x Q(x) = 1 \tag{410}$$

于是，有：

$$0 = \ln(1) \geq \sum_x P(x) \ln(\frac{Q(x)}{P(x)}) \tag{411}$$

即：

$$D_{KL}(P\|Q) = \sum_x P(x) \ln(\frac{P(x)}{Q(x)}) \geq 0 \tag{412}$$

2. **非对称性**：$D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$，即 KL 散度不是对称的，交换 P 和 Q 一般会导致不同的结果。

应用

- **机器学习**：在训练过程中，KL 散度常用于优化目标函数，例如变分自编码器（VAE）和生成对抗网络（GAN）。通过最小化 KL 散度，可以使近似分布 Q 尽可能接近真实分布 P，从而提高模型的准确性和效率。
- **信息论**：用于测量编码方案的效率，评估数据压缩方案等。
- **统计学**：用于假设检验和模型选择。

先验和后验

先验（Prior）和后验（Posterior）是贝叶斯统计中的两个核心概念，用于描述不确定性和信息更新的过程。

先验概率（Prior Probability）

定义：先验概率是指在获得新数据之前，根据已有的知识或经验对某一事件或参数的初始估计。先验概率反映了在观察到新数据之前，我们对某一事件或参数的不确定性。

表示方法：用 $P(\theta)$ 表示，其中 θ 代表参数或事件。

作用：先验概率提供了一个起点，在进行贝叶斯推断时，它与新的数据结合，更新我们的认知。

后验概率（Posterior Probability）

定义：后验概率是指在获得新数据之后，根据贝叶斯定理更新的某一事件或参数的概率分布。后验概率反映了在观察到新数据之后，我们对某一事件或参数的不确定性。

表示方法：用 $P(\theta \mid D)$ 表示，其中 θ 代表参数或事件， D 代表新观察到的数据。

计算方法：根据贝叶斯定理，后验概率可以通过先验概率、似然函数和边际似然计算得到：

$$P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)} \tag{413}$$

其中：

- $P(\theta \mid D)$ 是后验概率。
- $P(D \mid \theta)$ 是似然函数，表示在给定参数 θ 时观察到数据 D 的概率。
- $P(\theta)$ 是先验概率。
- $P(D)$ 是边际似然，表示观察到数据 D 的总体概率。

拓扑向量空间

拓扑向量空间（Topological Vector Space，简称 TVS）是一个定义在拓扑域 $\mathbb{K}$（通常是带有标准拓扑的实数或复数）上的向量空间，该空间被赋予了一个拓扑结构，使得向量加法 $\cdot + \cdot : X \times X \to X$ 和标量乘法 $\cdot : \mathbb{K} \times X \to X$ 是连续函数（这些函数的定义域赋予了乘积拓扑）。这样的拓扑被称为 X 上的**向量拓扑**或**TVS 拓扑**。

拓扑向量空间是数学分析和函数空间理论中的重要概念，它们将向量空间的代数结构与拓扑空间的结构相结合，从而使我们能够更好地理解向量空间中的连续性和收敛性。

超平面

超平面（Hyperplane）是指一个比所在拓扑向量空间少一维的平滑仿射子空间。
半空间（Half Space）是指拓扑向量空间被超平面划分出的两个区域之一。

假设有一个超平面，其由以下方程定义：

$$\mathbf{n} \cdot \mathbf{x} = c \tag{414}$$

其中，$\mathbf{n}$ 是垂直于超平面的法向量，$\mathbf{x}$ 是空间中的一个点，c 是一个常数。

两个半空间分别由以下不等式定义：

$$\mathbf{n} \cdot \mathbf{x} \geq c \tag{415}$$

和

$$\mathbf{n} \cdot \mathbf{x} \leq c \tag{416}$$

这些不等式中的每一个代表了超平面两侧的一个半空间，满足其中一个不等式的点位于相应的半空间中。

紧空间

紧空间（Compact Space）在数学中是一种具有特殊性质的空间，即它在某种意义上表现得像“有限的”，即使它可能看起来非常大，甚至是无限的。

一个空间被称为紧致的，如果可以用有限数量的小而重叠的片段完全覆盖整个空间。换句话说，即使这个空间本身可能非常大或无限大，但紧致性意味着总能用有限数量的部分来描述它的全貌。

紧空间可以理解为一种“有限”或“被包含”的空间。这种空间不会让你“无限延伸”，而是会将你限制在某个范围内。想象你在一个小岛上，无论你走到哪里，总会遇到岛的边缘——你不能无限制地前进，总有一个尽头。这类似于紧空间。

相反地，如果你在一片无边无际的沙漠中，可以一直走下去而永远不会到达尽头，这类似于非紧空间。在紧空间中，总有一种“有限”的感觉，而在非紧空间中，感觉像是没有尽头的延伸。

Taylor展开

Taylor展开（Taylor Expansion）是用多项式来近似一个函数的工具。它表示一个函数在某一点附近的值为该函数在该点的导数信息的线性组合，从而通过简单的多项式来逼近复杂的函数。

定义：

给定一个在某点 a 处可导多次的函数 $f(x)$，它的 **Taylor 展开** 在点 a 处的表达式为：

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n(x) \tag{417}$$

其中：

- $f^{(n)}(a)$ 表示函数 $f(x)$ 在点 a 处的第 n 阶导数，
- $R_n(x)$ 是剩余项（余项），它表示截断后，未被包含的误差部分。

当 x 足够接近 a 时，截取足够多项的 Taylor 展开可以非常准确地逼近函数值。

特殊情况：麦克劳林（Maclaurin）展开

当 $a = 0$ 时，Taylor 展开被称为 **麦克劳林展开**，形式为：

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \dots \tag{418}$$

例子：

1. **指数函数的 Taylor 展开**（以 $a = 0$ 为例，即 麦克劳林展开）：

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \tag{419}$$

2. **正弦函数的 Taylor 展开**（在 $a = 0$ 处）：

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots \tag{420}$$

通过 Taylor 展开，我们可以在某个点附近用有限项多项式来近似复杂的函数。这在数值计算和分析中非常有用。

参考文献

Abernethy, Jacob, et al. "Optimal strategies and minimax lower bounds for online convex games." Proceedings of the 21st annual conference on learning theory. 2008.

Auer, Peter. "Using confidence bounds for exploitation-exploration trade-offs." Journal of Machine Learning Research 3.Nov (2002): 397-422.

Bouneffouf, Djallel. "Finite-time analysis of the multi-armed bandit problem with known trend." 2016 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2016.

Bubeck, Sébastien, Ronen Eldan, and Yin Tat Lee. "Kernel-based methods for bandit convex optimization." Journal of the ACM (JACM) 68.4 (2021): 1-35.

Boyd, Stephen, and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.

Devroye, Luc, László Györfi, and Gábor Lugosi. A probabilistic theory of pattern recognition. Vol. 31. Springer Science & Business Media, 2013.

Feller, William. "An introduction to probability theory and its applications." (1971).

Flaxman, Abraham D., Adam Tauman Kalai, and H. Brendan McMahan. "Online convex optimization in the bandit setting: gradient descent without a gradient." arXiv preprint cs/0408007 (2004).

Hazan, Elad, Amit Agarwal, and Satyen Kale. "Logarithmic regret algorithms for online convex optimization." Machine Learning 69.2 (2007): 169-192.

Kearns, Michael J., and Umesh Vazirani. An introduction to computational learning theory. MIT press, 1994.

Lai, Tze Leung, and Herbert Robbins. "Asymptotically efficient adaptive allocation rules." Advances in applied mathematics 6.1 (1985): 4-22.

McAllester, David A. "PAC-Bayesian stochastic model selection." Machine Learning 51.1 (2003): 5-21.

Mohri, Mehryar. "Foundations of machine learning." (2018).

Nakkiran, Preetum, et al. "Deep double descent: Where bigger models and more data hurt." Journal of Statistical Mechanics: Theory and Experiment 2021.12 (2021): 124003.

Penot, Jean-Paul. "On regularity conditions in mathematical programming." Optimality and Stability in Mathematical Programming (1982): 167-199.

Robbins, Herbert. "Some aspects of the sequential design of experiments." (1952): 527-535.

Thompson, William R. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples." Biometrika 25.3-4 (1933): 285-294.

Wainwright, Martin J. High-dimensional statistics: A non-asymptotic viewpoint. Vol. 48. Cambridge university press, 2019.

Wang, Guanghui, Shiyin Lu, and Lijun Zhang. "Adaptivity and optimality: A universal algorithm for online convex optimization." Uncertainty in Artificial Intelligence. PMLR, 2020.

Zhang, Lijun, Shiyin Lu, and Zhi-Hua Zhou. "Adaptive online learning in dynamic environments." Advances in neural information processing systems 31 (2018)

Zhang, Lijun, Tie-Yan Liu, and Zhi-Hua Zhou. "Adaptive regret of convex and smooth functions." International Conference on Machine Learning. PMLR, 2019.

Zinkevich, Martin. "Online convex programming and generalized infinitesimal gradient ascent." Proceedings of the 20th international conference on machine learning (icml-03). 2003.