arXiv:2601.06629v1 [cs.DS] 10 Jan 2026

学习型索引算法复杂度的下界

Luis Alberto Croquevielle 伦敦帝国理工学院，伦敦，英国
Roman Sokolovskii 伦敦帝国理工学院，伦敦，英国
Thomas Heinis
伦敦帝国理工学院，伦敦，英国

摘要

学习型索引结构旨在通过训练机器学习模型来逼近与数据库属性相关联的秩函数，从而加速查询。尽管在实践中效果显著，但其理论局限性尚未得到充分理解。我们提出了一个通用框架，用于证明学习型索引的查询时间下界，并以其空间开销表示，且根据用于逼近的模型类进行参数化。我们的表述涵盖了广泛的学习型索引，包括大多数现有设计，并将其视为分段模型预测器。

我们分两步解决查询时间下界问题：首先，当数据库属性取自某一概率分布时，我们使用概率工具来控制抽样的影响。接着，我们分析逼近理论问题：如何使用给定模型类中的逼近器来最优地表示累积分布函数。在此框架下，我们在一系列建模和分布假设下推导出下界，并特别关注分段线性模型类和分段常数模型类的情形；它们在实际实现中十分常见。

我们的分析展示了如何利用逼近理论中的工具（如量化和柯尔莫哥洛夫宽度），来形式化学习型索引结构固有的时空权衡。由此得到的界揭示了这些方法的核心局限。

2012年 ACM 学科分类 计算理论 → 前驱查询；计算理论 → 数据管理的数据结构与算法

关键词与短语 学习型索引、随机过程、逼近理论

数字对象标识符 10.4230/LIPIcs.CVIT.2016.23

1 引言

高效查询处理是数据库系统中的一项基础挑战。对于关系型数据库，重点研究领域包括连接优化技术 [31, 17]，基数估计方法 [2, 27, 19]，以及面向多维数据的空间索引 [10, 14]。一般而言，降低查询延迟需要权衡其他资源，例如空间开销或 I/O 操作次数。

对于一维索引，B+树 [12] 等基于树的结构在实践中是最主要的选择。这些方法旨在针对单个有序属性回答点查询和范围查询。通常，属性值（或键）存储在排序结构A中，查询则表示为对 A执行的搜索操作。具体而言，要查询值 q，需要在 A中定位 q 的位置；而要查询区间 $[q, q']$，则需要找到第一个键 $\geq q$，然后向后扫描，直到遇到键 $> q'$。

近年来，机器学习已成为一种颇具前景的查询处理工具。一维索引的一个典型例子是递归模型索引 [18]，，它维护着类似于 B+ 树的层次结构，但在内部节点中嵌入预测模型，以估算排序数据子集中的键位置。这项工作启发了

©作者：请填写 \Copyright 宏；依据知识共享署名 4.0 国际许可协议授权

第42届非常重要主题会议（CVIT 2016）。编辑：John Q. Open 和 Joan R. Access；文章编号：23；页码：23:1–23:34 莱布尼茨国际信息学会议录 达格施图尔城堡 – 莱布尼茨信息学中心，达格施图尔出版社，德国
LIPICS

一系列广泛的研究，致力于开发“学习型索引” [11, 16, 15, 5], ，将这一思路扩展到动态工作负载 [5, 29], 、流式数据 [30, 21], 和多维查询 [25, 6, 20]。

尽管实证研究凸显了学习型索引相较于传统方法的潜在优势 [23, 28], ，但学习型索引通常缺乏形式化保证。研究还发现，与 B+树等成熟方法的一贯表现不同，学习型索引的性能会随数据集和查询工作负载而变化。如何为学习型索引证明算法保证，以及如何刻画这些方法的数据依赖行为，都是重要的理论问题，但目前仅得到部分探索。

本文提出了一个形式化框架，用以涵盖广泛的学习型索引结构。在此框架下，我们建立了查询时间下界，并明确刻画了时间复杂度与空间复杂度之间的权衡。这些下界在一般假设下成立，适用于范围广泛的数据生成过程。作为一个有趣的应用，这些下界使我们能够确定一个参数范围，在该范围内，学习型索引相较于传统方法不具备渐近优势。

本文组织如下：第 2 节将更详细地介绍相关工作和我们自身的贡献。第 3 节介绍数学设定和符号。第 4 节提出我们的一般理论方法，第 5 节和第 6 节则将该框架应用于具体的函数逼近场景。最后，第 7 节给出结论性说明，并讨论未来工作。

2 相关工作与贡献

下一节将介绍完整的符号体系，但我们在此先给出一些初步定义。我们将数据库关系中的属性 A 建模为由一系列数值构成： X_1，……， X_n。我们分析的核心研究对象是**秩**函数，其定义为**秩**$(q) = \sum_{i=1}^{n} \mathbf{1}_{(-\infty, X_i]}(q)$，用于统计小于或等于给定查询值 q的键的数量。在学习型索引的语境中，查询时间通常指计算**秩**(q)所需的时间。

B+树等经典方法不考虑数据的生成过程。相比之下，学习型索引通过对数据生成过程作出假设来获得性能保证，例如假设$\{X_i\}$ 相对于某个分布是独立同分布的。鉴于相关研究领域的情况，这一点并不意外，尤其是统计学习理论，其中泛化保证通常依赖于训练数据和测试数据均满足独立同分布假设。

作为基线，B+树支持在 $O(\log n)$ 时间内执行查找、插入和删除，并使用$O(n)$ 空间，其最坏情况保证独立于数据生成过程而成立。学习型索引的早期理论分析始于 PGM-Index [9]，结果表明，在静态场景下，其性能可与 B+树相当，但对于动态工作负载，其查找复杂度为 $O(\log^2 n)$。在对数据生成过程作出（限制性）假设的条件下，研究表明其空间复杂度相较于 B+树可获得常数级改进 [8]。

近期的研究表明，对于静态数据库，某些特定的学习型索引在期望意义下的渐近性能可以优于 B+树。Zeighami 和 Shahabi [32]证明，使用线性空间可以实现 $O(\log(\log n))$ 的期望查找时间。Croquevielle 等人 [3] 进一步将其改进为 $O(1)$ 的期望时间。对于动态数据库，Zeighami 和 Shahabi [33] 考虑了一种数据分布可能随时间变化的模型，其中参数 δ 用于刻画分布偏移的程度。他们提出了一种学习型索引，在 $\delta = 0$ (即没有分布偏移)时，支持以 $O(\log(\log n))$ 的期望时间执行插入操作和查询，其渐近性能优于 B+树。对于 $\delta > 0$，则需额外付出 $O(\log(\delta n))$ 的代价，因此渐近复杂度与 B+树相同。

Zeighami 和 Shahabi [34], 最近提出了另一条研究路线，从信息论视角研究学习型数据库操作的下界。在他们的框架中，学习型索引被建模为一个参数化函数 f_θ，其参数 θ 以总计 σ 位的预算进行编码。作者推导出模型大小 σ的下界，并指出，如果模型过小，则至少存在一个输入序列X_1，..... X_n，使其无法达到所需的近似精度。

我们的工作采用了根本不同的方法。我们不分析一般函数 f_θ，而是聚焦于一种结构化的索引算法类别 $\mathcal{I}$，它涵盖了大多数学习型索引构造中常见的设计元素。与 [34], 的一般情形相比，这缩小了我们工作的范围，但也使我们能够得出更精确的下界，且这些下界与文献中通常定义的学习型索引更加直接相关。在这一类别中，我们建立了查询时间下界，并将其表示为索引结构所用空间开销的函数。

更具体地说，我们将学习型索引建模为有向无环图（DAG），其中内部节点负责引导查找，汇点节点则编码预测模型并存储键。我们提出了一个通用框架，用于证明查询时间下界，并将其表示为模型类复杂度、汇点节点数量 K、数据集大小 n以及所选局部搜索算法（例如线性搜索或二分搜索）的函数。

对于涵盖大多数已知设计的广泛学习型索引类，我们证明，在最坏情况下，查询时间的下界为 $\Omega(\log(n/K))$。一个重要推论是，为了获得相对于 B+树等传统结构的渐近改进（传统结构保证 $O(\log n)$ 的查询时间和线性空间），学习型索引必须使用接近线性的空间。具体而言，如果汇点节点的数量满足 $K = O(n^\alpha)$（对于任意 $0 < \alpha < 1$），则最坏情况查询时间仍为 $\Omega(\log n)$——与 B+树相匹配。因此，只有当空间使用量接近线性时，学习型索引才有望在最坏情况下获得渐近收益。当模型类受限于分段常数函数时，这一结论还同样适用于平均情况查询时间，其中期望取自查询分布。

3 预备知识

3.1 将数据视为随机过程

我们考虑一个关于数据库属性的概率模型。形式化地，我们将该属性建模为随机过程 $X = \{X_i\}_{i\in\mathbb{N}}$，其中每个 X_i 都是取实数值的随机变量。在时间 $n \in \mathbb{N}$，该属性由序列中前 n 个值按递增顺序排列而成。我们将排序后的值记为 $X_{(1)} \leq \ldots \leq X_{(n)}$，因此该属性表示为一个数组 A_n，对于所有 $i \in \{1, \ldots, n\}$，都有 $A_n[i] = X_{(i)}$。这一框架同时适用于最坏情况和平均情形分析：我们可以考察单个序列 $X_1, \ldots, X_n$ 来研究最坏情况行为，或者纳入支配该随机过程的概率测度 $\mathbb{P}$ 来分析期望性能。

在完全一般的情况下，不会有适用的复杂度下界。例如，如果 $X_i = i$，那么秩函数 **秩** 可在 $O(1)$ 时间和空间内计算，而 PGM-Index 等现有结构很容易达到这一性能。在本工作中，我们在如下假设下推导下界： X_i 是按照同一个累积分布函数（CDF） F抽取的独立随机变量。我们进一步假设，查询值 q 独立于 X_i 生成，其依据的概率测度 μ 满足

其支撑为 $\mathcal{Q} \subseteq \mathbb{R}$。[1] 我们强调，独立性假设仅用于平均情形分析；我们的最坏情形界无条件成立，尽管推导过程中将独立性作为一种技术工具。

全文中，我们用 $\mathbb{E}_X[\cdot]$ 表示相对于控制键序列 $\{X_i\}_{i=1}^n$ 的测度$\mathbb{P}$ 的期望。记号 $\mathbb{E}_q[\cdot]$ 表示相对于查询分布 μ 的期望，并将键值 $\{X_i\}_{i=1}^n$ 视为固定（即，以数据的一次实现为条件）。联合计算期望时，我们记作 $\mathbb{E}_{X,q}[\cdot],$，这对应于相对于乘积测度 $\mathbb{P} \otimes \mu$ 的积分。

3.2 计算模型与学习问题

如第 2 节所述，我们考虑属性 A_n 上的**秩** 函数，其定义为

$$\mathbf{rank}_n(q) = \sum_{i=1}^{n} \mathbf{1}_{(-\infty, X_i]}(q)\,,$$

其中，$\mathbf{1}_U$ 表示集合 U 的示性函数。因此，**秩**$_n(q)$ 将查询值$q \in \mathbb{R}$ 映射为小于或等于它的键的数量。根据上下文通常可以明确其含义，我们省略下标 n，直接写作**秩**。对于学习型索引，目标是在尽量减少计算所需时间和空间开销的同时，准确地近似**秩**(q)。

我们采用均匀代价准则下的标准 RAM 模型 [1],：每条基本指令需要一个时间单位，每个存储任意整数的寄存器占用一个空间单位。时间复杂度定义为使用索引计算**秩**(q)所需的基本指令数。空间复杂度衡量索引使用的冗余内存，不包括存储属性 A_n 本身所需的空间。

设 I 是一个索引构建过程，它根据属性 A_n构建数据结构 $I(A_n)$。对于查询值 q，令 $T(I(A_n)\ q)$表示使用数据结构 $I(A_n)$计算**秩**(q)所需的时间，并令 $S(I(A_n))$表示 $I(A_n)$使用的空间开销。由于上下文通常已足够明确，我们通常用 $T(q)$代替 $T(I(A_n)\ q)$。

3.3 学习型索引结构

我们将学习型索引建模为有向无环图（DAG），如图 1 所示。内部节点实现路由函数，汇点节点则包含预测模型并存储键。每个查询都会沿着一条从根节点到汇点节点的路径前进，并在汇点节点处通过局部搜索来修正其秩的预测值。这种结构涵盖了广泛的学习型索引，其中大多数都包含分层路由，例如 RMI [18], ALEX [5], 和 PGM-Index [9]。

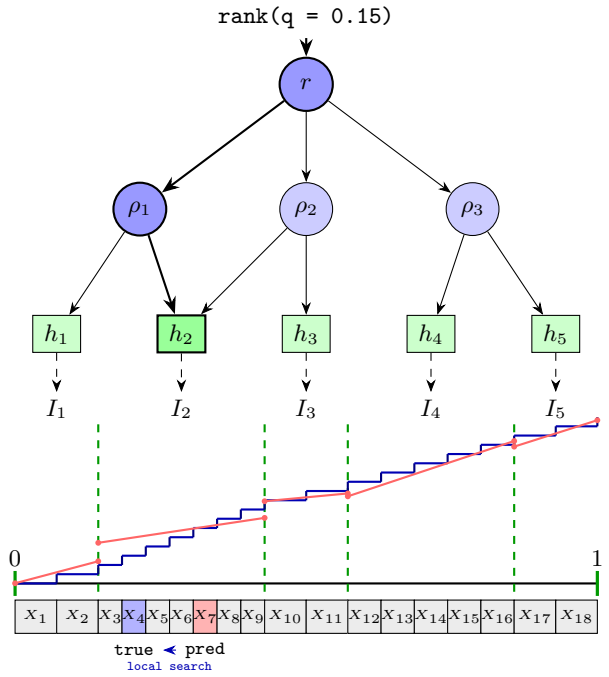
为推导下界，我们重点关注局部校正步骤。无论查询如何路由，最终都要在有序数组中围绕预测值 $h(q)$ 搜索，以恢复 q 的精确秩。由于我们的目标是得到不受内部结构影响的下界——而不同设计的内部结构可能差异很大——因此，在我们的分析中，这种局部搜索的成本成为决定查询复杂度的主要因素。

因此，我们将索引抽象为由其汇点模型定义的分段函数，并令 K 表示汇点的数量。这对应于将查询域划分为 K 个区域$\{I_v\}_{v=1}^K$，每个区域都关联一个模型 h_v。尽管我们没有分析

[1] 形式上，必须为数据和查询分布指定可测空间。为简单起见，我们假设所有相关空间都配备了 Borel σ-代数，并且我们进行积分的所有函数都是可测的；这些假设是标准做法，不会引入实质性限制。

路由结构，但我们将其纳入模型，以强调无论图拓扑或路由函数的复杂程度如何，我们的下界都适用。理解这种结构如何影响查询时间仍是一个开放问题，并可能带来更精确的界。

我们将以此方式定义的学习型索引类别记作$\mathcal{I}$，并将其汇点模型属于函数类$\mathcal{L}$（例如，$\mathcal{L}$ 可以是多项式函数集合）且最多使用 K 个汇点的学习型索引记作$\mathcal{I}_{K,\mathcal{L}}$。在本文其余部分，我们将 K 用作空间使用量的下界（及代理量），并着重探讨$\mathcal{L}$ 的近似能力如何与 K 相互作用，以刻画性能。



根节点： 所有查询都从此处进入，并沿 DAG 路由。图中以较粗的遍历线突出显示了 $q = 0.15$ 的一条示例路径。

内部节点： 每个内部节点v 都会应用路由函数 $\rho_v(q)$，将查询引导至其某个子节点。

汇点节点： 每个汇点节点 v 都包含一个模型 $h_v(q)$，用于估计子区间 I_v内的**秩**(q)。

在此示例中，真实秩函数以蓝色阶梯函数表示，预测模型 h_v 则以红色线性片段表示。

局部搜索： 估计值 $h_v(q)$ 通过对一个……进行局部搜索，从而得到修正。该子数组存储在汇点节点中。

图 1 将学习型索引建模为有向无环图（DAG）的示意图。

4 下界的一般框架

现在，我们提出一个用于推导学习型索引查询时间下界的一般框架。我们将查询时间 $T(q)$ 视为使用在数组 A_n 上构建的索引计算**秩**(q)所需的操作数。下界通过汇点节点的数量 K，以空间开销 $S(R(A_n))$ 的形式表示。分析基于以下三个思路：

(a) 将查询时间 $T(q)$ 与预测误差 $\varepsilon(q) = |h_v(q) - \text{秩}(q)|$联系起来。我们的分析聚焦于查询执行的最后一步：模型预测后的局部搜索。由于这一步始终是确认或纠正预测秩所必需的，因此其开销为 $T(q)$ 提供了有效的下界，且该下界随 $\varepsilon(q)$ 的大小而增大。(b) 通过使用逼近秩$(q) \approx nF(q)$，将对预测误差 $\varepsilon(q)$ 的分析转化为针对累积分布函数 F 的函数逼近问题。(c) 一种概率分析，对逼近秩$(q) \approx nF(q)$ 进行了形式化，并使我们能够在各种概率假设下推导出关于 $T(q)$ 的下界。

我们首先介绍组件 (b) 和 (c)，并将 (a) 的讨论留到本节末尾，届时我们将为 $T(q)$ 给出一般下界。为清晰起见，本节阐述过程中省略了大多数证明，所有省略的技术细节均收录于附录。

4.1 从学习型索引到逼近误差

第3.3节中定义的学习型索引可以看作一种结构，它通过分段定义的预测函数来近似秩(q)

$$h(q) = \sum_{v=1}^{K} h_v(q)\mathbf{1}_{I_v}(q)\,, \tag{1}$$

其中每个 h_v 都是一个预测模型，而 $I_v \subseteq \mathcal{Q}$ 是其对应的子域。为方便记号，我们将索引 I 与其诱导的预测模型 h视为相同；因此，我们用$h \in \mathcal{I}_{K,\mathcal{L}}$ 表示： h 由类别 $\mathcal{L}$中的函数在 K 个互不相交的区间上分段定义。

给定预测函数 h，预测误差为 $\varepsilon(q) = |\textbf{秩}(q) - h(q)|$。为了从函数逼近的角度分析 $\varepsilon(q)$，我们使用经验累积分布函数（ECDF）。回顾一下，我们的基本假设是，键 $X_1, \ldots, X_n$是从累积分布函数为 F 的分布中独立同分布地抽取的。因此，ECDF 定义为

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty, X_i]}(x)\,.$$

于是，秩函数可表示为$\textbf{秩}(q) = nF_n(q)$。现在，无论是分析平均情况还是最坏情况，都需要考虑在给定 $X_1, \ldots, X_n$ 的条件下，所有查询值 q 上的期望误差。由于 q是独立于$\{X_i\}$生成的，因此我们有

$$\mathbb{E}_q\big[\varepsilon(q)\big] = \int_{\mathcal{Q}} \varepsilon(q)d\mu = \int_{\mathcal{Q}} |\mathbf{rank}(q) - h(q)|d\mu = \int_{\mathcal{Q}} |nF_n(q) - h(q)|d\mu = \|nF_n - h\|_\mu\,, \tag{2}$$

其中 $\|\cdot\|_\mu$ 表示 $L^1(\mu)$ 中的绝对范数，即$\mathcal{Q}$上 μ-可积函数构成的空间。换句话说，平均而言，最佳的 h函数就是使其与$\textbf{秩} = nF_n$之间的绝对距离最小的函数。这启发我们引入最小逼近误差：

$$R_{K,\mathcal{L}}(G) = \inf_{h \in \mathcal{I}_{K,\mathcal{L}}} \|G - h\|_\mu\,,$$

对于任意 $G \in L^1(\mu)$。当 $G = \textbf{秩} = nF_n$时，可得

$$R_{K,\mathcal{L}}(\mathbf{rank}) = R_{K,\mathcal{L}}(nF_n) = nR_{K,\mathcal{L}}(F_n)\,, \tag{3}$$

根据近似误差的线性缩放性质（见附录A中的引理7）。这使我们能够将 $\mathbb{E}_q[\varepsilon(q)]$ 的分析归结为一个逼近问题。核心思路是， $\mathbb{E}_q[\varepsilon(q)]$ 可以由 F的最优 $L^1(\mu)$-逼近给出下界，误差至多为一个由抽样随机性导致的小偏差项。该偏差由函数 $\delta_n(x) = F(x) - F_n(x)$刻画，其中 F_n 为经验分布函数。以下结果将这一关系精确化。证明很直接，可见附录A。

▶ **命题 1.** 设 $X_1, \ldots, X_n$ 是来自累积分布函数为 F的分布的独立同分布样本，并令$q \sim \mu$ 从 $\{X_i\}$中独立抽取。那么对于任意预测模型 $h \in I_{K,\mathcal{L}}$ 均有

$$\mathbb{E}_q\big[\varepsilon(q)\big] = \|nF_n - h\|_\mu \geq n\left[R_{K,\mathcal{L}}(F) - \|\delta_n\|_\mu\right]\,,$$

前提是 L 在标量乘法下保持不变。

该不等式表明，期望预测误差以下界为确定性的逼近项 $nR_{K,\mathcal{L}}(F)$，减去依赖数据的波动项 $n\|\delta_n\|_\mu$。控制这一波动是我们分析中的主要概率步骤，将在下一小节中讨论。除了 $\mathbb{E}_q[\varepsilon(q)]$ 之外，我们还需要对 $\mathbb{E}_q[\log_2 \varepsilon(q)]$ 给出界，因为它会在指数搜索的分析中出现。为此，我们定义查询时间的一个代理量：$\sim T(q) = \log_2(\max\{2, \varepsilon(q)\})$，并为其期望建立一个下界。这需要对 F、μ 和 $\mathcal{L}$ 进行仔细分析并作出具体假设。以下结果展示了可以推导出的界的类型（证明见附录 B）。

▶ **命题 2.** 令 $q \sim \mu$ 与键 $\{X_i\}_{i=1}^n$ 相互独立。假设累积分布函数 F 和 μ 分别具有密度函数 f 和 g，且满足 $0 < c_F \le f(q)\ g(q) \le C_F < \infty$ 对所有 $q \in Q$ 均成立。那么，存在 $X_1, \ldots, X_n$ 的一种实现，使得：

$$\mathbb{E}_q\left[\widetilde{T}(q)\right] \ge C_1\left[\log_2\left(nR_{K,\mathcal{L}}(F)\right) - C_2\right],$$

其中 $C_1, C_2 > 0$ 与 n 和 K 相互独立，而 $L = P_0$——常函数类。

命题 2 提供了命题 1 的对数形式类比，可用于~分析指数搜索。在 $T(q)$ 中使用 $\max\{2, \cdot\}$ 可确保查询时间的代理量至少为 1。常数 C_1 和 C_2 取决于 c_F 和 C_F，可能会带来不利影响。不过，它们与 n 和 K 无关，也不会改变渐近分析。此外，条件 $0 < c_F \le f, g \le C_F < \infty$ 不必在整个 $\mathcal{Q}$ 上都成立；只需在子区间 $\mathcal{Q}' \subseteq \mathcal{Q}$ 上成立即可。与命题 1 相比，这一变体的主要局限在于约束 $\mathcal{L} = P_0$，以及该界限仅对特定实现 $X_1, \ldots, X_n$ 成立这一事实：它未必对所有 $\{X_i\}_{i=1}^n \sim \mathbb{P}$ 都成立。

4.2 控制随机性：F_n 的集中

命题 1 和命题 2 中的下界取决于 $R_{K,\mathcal{L}}(F)$ 以及 δ_n 的行为。一方面，$R_{K,\mathcal{L}}(F)$ 是一个确定性量，仅取决于 F、模型类 $\mathcal{L}$ 和测度 μ，可以借助逼近理论进行研究。另一方面，我们可以使用概率工具来控制项 $\|\delta_n\|_\mu$，确保它足够小，使误差的主要来源是对 F 的逼近，而非抽样变异。

一般而言，我们希望说明 δ_n 的行为类似于 $1/r(n)$，其中某个函数 $r(n)$ 用于量化 F_n 到 F 的收敛速率。这样，我们便能将所有随机性集中到单个项 $1/r(n)$ 中，并用 F 的确定性近似误差来表示关于 $\mathbb{E}_q[\varepsilon(q)]$ 的下界。为使这一思路更加严谨，我们依赖两类结果（所有证明均见附录 D）：期望集中和最坏情况集中。

期望集中 这一类的主要结果由引理 3 给出，我们将结合命题 1 使用该结果。

▶ **引理 3** (期望界)。 假设 $X_1, \ldots, X_n$ 相互独立同分布，并令 μ 为 Q 上任意一个与 $\{X_i\}$ 独立的概率测度。则有：
(a) $\mathbb{E}_X[\|\delta_n\|_\infty] \le \sqrt{\frac{\pi}{2n}}$。 **(b)** 因此，
$\mathbb{E}_X[\|\delta_n\|_\mu] \le \sqrt{\frac{\pi}{2n}}$。 —

这些界使我们能够推理 δ_n 关于随机样本 $\{X_i\}$ 的期望集中情况，从而从平均情况的角度看待近似误差——同样，无需对查询分布 μ 作出任何假设。

最坏情况集中 这类结果的主要结论由引理 4 给出，我们用它来分析学习型索引在最坏情况下的算法性能。

▶ **引理 4**（小偏差的存在性）。假设 X_1、……、X_n 独立同分布，且令 μ 为 Q 上任意一个与 $\{X_i\}$ 独立的概率测度。那么，有：
（a）存在 $\{X_i\}_{i=1}^n$ 的一个实现，使得 $\|\delta_n\|_\mu \le \sqrt{\frac{\pi}{2n}}$。（b）如果 μ 等于数据生成分布，即 $d\mu = dF$，那么存在 $\{X_i\}_{i=1}^n$ 的一个实现，使得 $\|\delta_n\|_\mu \le \frac{1}{\sqrt{6}\,n}$。

这表明，对 μ 施加更强的假设可以得到更紧的界——具体而言，得到$1/n$ 速率，而非一般的 $1/\sqrt{n}$。相比之下，对任意 μ 都成立的界必须更加保守，以涵盖不利情形：它们在大多数情况下可能较为宽松，但若无额外假设便无法改进。尽管全面刻画 μ 的不同选择如何影响收敛速率超出了本文范围，我们仍纳入 $d\mu = dF$ 情形，因为它反映了学习型索引文献中的常见设定。

4.3 从预测误差到查询时间

现在，我们通过将预测误差转换为实际查询时间，完成我们的框架。如第3.3节所述，查询处理的最后阶段包括执行局部搜索，将预测的秩 $h(q)$ 修正为真实值 $\textbf{秩}(q)$。此步骤的成本取决于预测误差 $\varepsilon(q) = |\textbf{秩}(q) - h(q)|$，并且会因所采用的算法而异。我们分析三种最常见的搜索策略：线性搜索、指数搜索和二分搜索。

▶ **命题 5**（通过预测误差得到的查询时间下界）。设 $q \in Q$ 是一个查询，其预测秩为 $h(q)$，实际秩为 $\textbf{秩}(q)$，且令 $\varepsilon(q) = |\textbf{秩}(q) - h(q)|$ 表示预测误差。那么，在相应的搜索策略下：

（a）线性搜索：$T(q) \ge \varepsilon(q)$。（b）指数搜索：$T(q) \ge \log_2(\max\{2, \varepsilon(q)\})$。（c）二分搜索：$\sup_q T(q) \ge \log_2(\sup_q \varepsilon(q))$。

证明。 我们依次分析每种策略。

(a) **线性搜索。** 此方法从预测位置 $h(q)$ 开始顺序扫描，直至到达真实位置 $\textbf{秩}(q)$，如图 2 所示。实际上，每一步搜索可以前进 2、3 个或任意固定数量的位置，但这只会使总数相差一个常数因子，不会改变渐近界。因此，我们假设每次比较前进一步。于是，总步数至少为 $\varepsilon(q) = |\textbf{秩}(q) - h(q)|$，由此得到下界 $T(q) \ge \varepsilon(q)$。

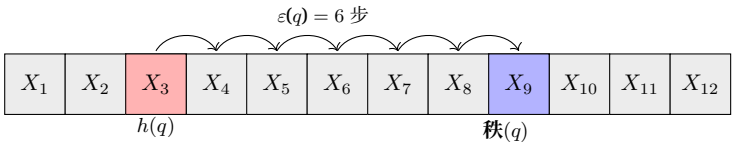


图 2 线性搜索从预测位置 $h(q)$ 开始逐步扫描，直至到达真实秩 $\textbf{秩}(q)$。总成本与预测误差 $\varepsilon(q)$ 成正比。

(b) **指数搜索。** 从 $h(q)$ 开始，算法以几何级数扩展（通常将偏移量加倍），直到在一个窗口内将真实位置 $\textbf{秩}(q)$ 框定，如图 3 所示。随后在该窗口内执行二分搜索。到达位置 $\ge \textbf{秩}(q)$ 所需的扩展步骤数为 $\lceil \log_2 \varepsilon(q) \rceil + 1 \ge \log_2 \varepsilon(q)$。其成本为

最终二分搜索的成本也与窗口大小呈对数关系。由于查询时间不可能少于 1 次操作，因此我们写作：

$$T(q) \geq \widetilde{T}(q) = \log_2\left(\max\{2, \varepsilon(q)\}\right),$$

其中，$\widetilde{T}$ 对应于前文引入的代理函数。

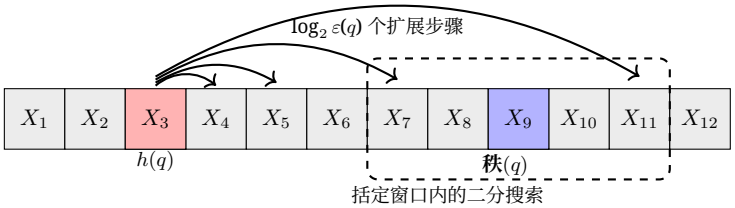


图 3 指数搜索会围绕预测位置 $h(q)$ 扩展几何窗口，直到真实秩 **秩**(q) 落入范围内。其代价关于预测误差 $\varepsilon(q)$呈对数级。

(c) **二分搜索**。在这种情况下，算法依赖于一个固定窗口大小，该窗口按构造保证包含真实秩。此窗口的大小必须足以覆盖所有 q 中的最大可能预测误差，即 $\sup_q \varepsilon(q)$。由于在大小为 M 的窗口中进行二分搜索，在最坏情况下至少需要 $\log_2 M$ 步，因此我们得到：

$$\sup_q T(q) \geq \log_2(\sup_q \varepsilon(q)).$$

如图 4 所示，尽管对于这个特定的 q 值，预测误差较小，但搜索范围仍必须足够宽，以容纳最坏情况误差。

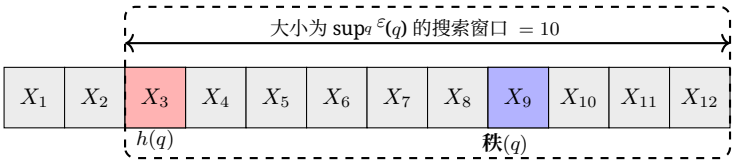


图 4 二分搜索需要一个保证包含真实位置**秩**(q) 的搜索区间。因此，其代价取决于最坏情况预测误差，并与区间大小的对数成正比。

◀

4.4 查询时间下界

现在，我们可以结合前面各小节的结果，推导第 4.3 节所述三种搜索策略下查询时间 $T(q)$ 的具体下界。我们考虑三种自然的下界类型，它们在保证力度和所需假设上各不相同。按一般性从低到高排列，分别是：

(A) 完全期望界（平均情况）：

$$\mathbb{E}_{X,q}[T(q)] \geq C(n, K, \mathcal{L}, F, \mu).$$

这是保证力度最强的一类结果，其结论适用于随机数据集$\{X_i\}_{i=1}^n$ 和查询值 $q \sim \mu$两者的联合期望。它通常需要较强的假设（例如独立同分布抽样），但能给出整个系统的平均情况界。

(B) 条件期望界（混合情况）：

$$\sup_{\{X_i\}_{i=1}^n} \mathbb{E}_q[T(q)] \geq C(n, K, \mathcal{L}, F, \mu).$$

这对应于数据上的最坏情况界和查询上的平均情况界。

(C) 逐点界（最坏情况）：

$$\sup_{\{X_i\}_{i=1}^n, q\in\mathcal{Q}} T(q) \geq C(n, K, \mathcal{L}, F, \mu)\,.$$

这是一种纯粹的最坏情况保证，断言存在某个数据集和查询，使得查询时间很长。它所需的假设最少。

这些情形形成了一个自然的层级：类型 (A) 的界蕴含类型 (B) 的界，而类型 (B) 的界又蕴含类型 (C) 的界。这种蕴含结构直接源于概率方法，反映了普遍性与假设强度之间的权衡。重要的是，这些界对模型类 $\mathcal{I}_{K,\mathcal{L}}$ 一致成立，即适用于 $\mathcal{I}_{K,\mathcal{L}}$ 中的每一种学习型索引结构。

表1总结了在不同搜索策略、关于查询分布 μ 的假设以及上述三种类型的下界 (A)–(C) 下，查询时间 $T(q)$ 的相应下界。这些结果表明，线性搜索可达到最强的界，其次是指数搜索，最后是二分搜索；对于二分搜索，我们只能得出最坏情况保证。形式化陈述和证明见附录 E；其中大多数可直接由本节前面建立的一般框架推出。

请注意，指数搜索的 (B) 类下界以渐近形式表示，而未给出 t
命题 2 中的常数 C_1，以突出其与其余部分的结构相似性 .

表1 在不同搜索策略、下界类型以及关于 μ 的假设下，查询时间 $T(q)$ 的下界。(A) 类下界进一步假设 $\{X_i\}$ 是独立同分布的，且 q 与 $\{X_i\}$ 相互独立。(B) 类下界假设 $\mathcal{L} = P_0$，q 与 $\{X_i\}$ 相互独立，并满足条件 $0 < c_F \leq f, g \leq C_F < \infty$。

下界	搜索	下界类型	假设	下界
L1	线性	（A）（平均/平均）	任意 μ	$n\left[R_{K,\mathcal{L}}(F) - \sqrt{\frac{\pi}{2n}}\right]$ —
L2	线性	(B) (最坏/平均)	$d\mu = dF$	$n\left[R_{K,\mathcal{L}}(F) - \frac{1}{\sqrt{6}\,n}\right]$
E1	指数	(B) (最坏/平均)	任何 μ	$\log_2\left(n\ R_{K,\mathcal{L}}(F)\right) - C_2$
E2	指数	(C) (最坏/最坏)	$d\mu = dF$	$\log_2\left(n\left[R_{K,\mathcal{L}}(F) - \frac{1}{\sqrt{6}\,n}\right]\right)$
B1	二进制	(C) (最坏/最坏)	任意 μ	$\log_2\left(n\left[R_{K,\mathcal{L}}(F) - \sqrt{\frac{\pi}{2n}}\right]\right)$
B2	二进制	(C)（最坏/最坏）	$d\mu = dF$	$\log_2\left(n\left[R_{K,\mathcal{L}}(F) - \frac{1}{\sqrt{6}\,n}\right]\right)$

这些结果凸显了学习型索引中模型复杂度、数据集大小、搜索策略和查询性能之间的权衡。请注意，表1考虑了查询分布的两种典型情形：(i) 任意 μ，以及 (ii) 匹配情形 $d\mu = dF$，即查询与数据集来自相同的分布。后一种情形在实践中尤为重要，因为它广泛用于学习型索引的评估 [9, 23, 21]。

总而言之，这一框架提供了一种通用且灵活的方法，用于推导查询时间下界。在接下来的章节中，我们将分析学习型索引结构中常用的具体模型类所对应的误差项 $R_{K,\ \mathcal{L}}(F)$。第5节重点讨论 $\mathcal{L} = P_0$，它对应分段常数模型 [32, 3]，；第6节则研究 $\mathcal{L} = P_1$，它对应分段线性近似器，也是实践中应用最广泛的选择。

5 分段常数模型的下界

第 4 节中推导的一般下界，只有在我们为近似误差 $R_{K,\ \mathcal{L}}(F)$ 得出明确公式后，才能充分体现其意义。这需要针对模型类 $\mathcal{L}$ 的特定选取，为该项推导下界。

几乎所有学习型索引结构都依赖多项式模型，因为这类模型需要存储的参数较少，而且训练和评估速度快。因此，我们重点关注形式为 $\mathcal{L} = P_m$ 的模型类，其中 P_m 表示次数至多为 m 的多项式类。在本节中，我们分析 $\mathcal{L} = P_0$ 的情形，它对应于分段常数模型。接下来，在第6节中，我们讨论实践中最广泛使用的情形：$\mathcal{L} = P_1$，对应于分段线性模型。

我们分析的核心思路是：如下文所示，对于分段常数模型，逼近累积分布函数 F 的问题可归结为求随机变量在 $[0,1]$ 上分布时的最优量化。这样，我们便可以借鉴量化理论中的基本结果，来刻画近似误差 $R_{K,\ P_0}(F)$。

本节的结果适用于一大类分布 F 和 μ——具体而言，是指两者都可由相对于勒贝格测度的密度函数刻画的情形。我们分别将这些密度函数记为 f 和 g，并同时考虑 $f = g$ 情形（第5.1节）以及一般情形（详见附录 F）。在这两种情形下，我们都证明近似误差随 K 线性递减；区别在于，对于 $f = g$，这一结论对任意 K都成立，而对于 $f \neq g$，该结果仅在 $K \to \infty$时渐近成立。

更具体地说，在第5节和第6节中，我们始终关注 F 为 Lipschitz 连续的情形。这一正则性假设相对温和，几乎涵盖了所有常见的连续分布（例如高斯分布、指数分布和均匀分布）。同时，它保证密度 f的存在，并确保下界反映的是模型类的局限性（由 K 和 $\mathcal{L}$决定），而非高度病态的目标函数所造成的假象。

5.1 相同的数据集分布和查询分布 $f = g$

对于分段常数近似函数 $h \in \mathcal{I}_{K,\ P_0}$ 和$f = g$这一相对简单的情况，对于任意 K，最小逼近误差 $R_{K,\ P_0}(F)$ 都可以精确计算。我们首先证明，这一问题设定可归结为对均匀分布随机变量进行最优量化，然后借助量化理论中的一个著名结果，计算相应最优量化器的误差。

令 $h \in \mathcal{I}_{K,\ P_0}$。由于 F 是一个取值范围为 $[0,1]$, 的累积分布函数，不失一般性，我们可以假设对于所有 q， $0 \leq h(q) \leq 1$ 。因此， h 可以写为

$$h(q) = \sum_{k=1}^{K} c_k \mathbf{1}_{I_k}(q)\,,$$

其中， K 区间 $\{I_k\}$ 划分了 $\mathcal{Q}$，而 $c_k \in [0,1]$ 是相应的近似常数。这意味着当 $f = g$时，绝对误差 $\|F - h\|_\mu$ 可以写为

$$\|F - h\|_\mu = \int_{\mathcal{Q}} |F(q) - h(q)|\, f(q) dq = \sum_{k=1}^{K} \int_{I_k} |F(q) - c_k|\, f(q) dq = \sum_{k=1}^{K} \int_{J_k} |u - c_k|\, du\,, \quad (4)$$

其中 $J_k = F(I_k) = \{F(q) :\ q \in I_k\}$（这是一个连续区间），最后一个等式是将 $u = F(q)$ 和 $du = f(q)dq$代入积分得到的。由于 F是累积分布函数，我们知道 $\{J_k\}$ 是 $[0,1]$的一个划分。若现在定义 $Q_K(u) = \sum_{k=1}^{K} c_k \mathbf{1}_{J_k}(u)\,,$，则

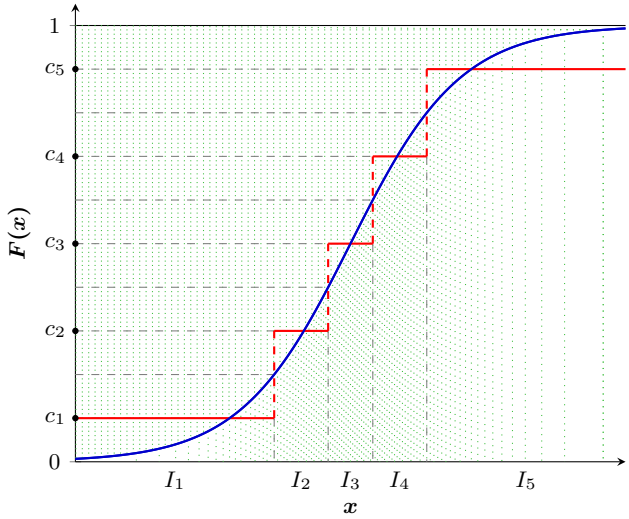


图5 当 $f = g$ 时（即数据集和查询来自同一分布），用含有 $K = 5$ 个分段的分段常数函数 h（红色）对累积分布函数 F（蓝色）进行最优逼近。此时，最优逼近值 $\{c_k\}$ 等间距分布（在 y 轴上标出），而对累积分布函数进行逼近等价于沿 y 轴对 Uniform$([0,1])$ 进行量化。量化边界（未标出的黑色水平虚线）通过 F^{-1} 映射到 x 轴上，将 $\mathcal{Q}$ 沿着 F 的五分位数（黑色垂直虚线）划分为 $K = 5$ 个等概率区域 $\{I_k\}$。绿色点线展示了经由 $F(X)$ 变换后，$X \sim F$ 的分布如何变为均匀分布。

（4）可以表示为

$$\|F - h\|_\mu = \mathbb{E}\big[|U - Q_K(U)|\big]\,, \tag{5}$$

其中，Q_K 是定义在 $[0,1]$ 上的分段常数函数，期望是对均匀分布 Uniform$([0,1])$ 的随机变量 U 取的。注意，(4)–(5) 体现了当 $X \sim F$ 时，$F(X)$ 服从均匀分布这一事实。

换言之，最小化 $\|F - h\|_\mu$ 等价于选择一组区间 $\{J_k\}$ 及其对应的点 $\{c_k\}$，使U 的某个实现与其对应值 c_k之间的期望距离最小。我们认识到，这是在绝对误差下对 U 进行最优量化的问题。

量化理论中一个众所周知的结果是，对于 Uniform$([0,1])$，均匀量化是最优的。具体而言，量化点应等距分布，如 $\{c_k = \frac{2k-1}{2K} :\ k = 1, \ldots, K\}$；量化区间 $\{J_k\}$ 应将定义域 $[0,1]$ 均匀划分。对应的最小量化误差等于 [13, 例 5.5]

$$T_{K,r}(U) = \inf_{Q_K} \{\mathbb{E}\left[|U - Q_K(U)|^r\right] : Q_K \text{ has } K \text{ quantization points}\} = \frac{1}{K^r(1+r)2^r}\,;$$

这意味着，用我们的记号表示，

$$R_{K,P_0}(F) = T_{K,1}(U) = \frac{1}{4K}\,. \tag{6}$$

值得注意的是，对于任意 F，最小逼近误差 (6) 以及最优的$\{c_k\}$ 和$\{J_k\}$ 都相同；对 F 的唯一依赖体现在区域 $\{I_k\}$ 上，这些区域应在 F 的 K-分位数处划分 $\mathcal{Q}$，从而得到大小相等的$\{J_k\}$。

我们在图5中以图形方式展示了这一设定，其中给出了一个 CDF 曲线示例$F(x)$（蓝色），并用分段常数函数 h（红色）对其进行了最优逼近，该函数包含 $K = 5$

个分段。逼近点 $\{c_k\}$ 沿 y 轴等间距排列，而这些点之间的中点经 F^{-1} 映射到 x 轴后，标出了 F 的五分位数，从而将 $\mathcal{Q}$ 划分为 $K = 5$ 个等概率区域 $\{I_k\}$。我们在式（4）–（5）中使用的、通过分布自身的累积分布函数对其进行“均匀化”的过程，由绿色虚线示出：蓝色累积分布函数曲线下方的密度与 f 成正比；而在累积分布函数曲线上方，将其映射到 y 轴后，密度则变为均匀分布。

6 分段线性模型的下界

如第5节所示，基于分段常数近似的学习型索引可以通过量化理论进行简洁的下界分析。在该节中，我们分析了单个（Lipschitz 连续）累积分布函数 F 的近似误差 $R_{K,\mathcal{L}}(F)$。本节将把分析扩展到表达能力更强的模型类，例如分段线性函数。我们的研究核心对象将是函数类 $\mathcal{F}$ 上的最坏情况逼近误差：

$$\sup_{F\in\mathcal{F}} R_{K,\mathcal{L}}(F)\,.$$

通过给这个量建立下界，我们可以证明，$\mathcal{F}$ 中存在一些函数，无法被任何 $h \in \mathcal{I}_{K,\mathcal{L}}$很好地逼近。与本文其他部分一致，我们聚焦于利普希茨连续累积分布函数这一类，并证明对于 $\mathcal{L} = P_1$，在这种最坏情况意义下，下界 $\Omega(1/K)$ 仍然成立，与经典上界相匹配（见附录 G.1）。

我们通过两种互补的方法建立这一结果。在第 6.1 节中，我们描述了一个显式的利普希茨连续累积分布函数族 $\{F_M\}$，对于该函数族，$R_{K,P_1}(F) = \Omega(1/K)$成立。这表明，类别 $\mathcal{I}_{K,P_1}$ 无法保证在所有利普希茨连续函数上都具有一致的 $o(1/K)$ 近似误差，并且该构造适用于一大类测度 μ。

在第 6.2 节中，我们通过柯尔莫哥洛夫宽度来研究这一问题；这是逼近理论中的重要工具。该方法通常假设 μ 为勒贝格测度，并且只允许存在性证明（而非构造性证明）。尽管如此，它仍是一个非常灵活的框架：它使我们能够分析超出 P_1 的更广泛模型类，并为更规则的函数空间推导出更紧的界限。总的来说，这两种方法凸显了学习型索引结构所面临的逼近极限的互补方面。

6.1 一个对抗性累积分布函数构造

为说明分段线性近似器的局限性，我们构造了一个具有以下性质的累积分布函数族$\{F_M\}_{M\in\mathbb{N}}$：对于任意具有 $\alpha > 0$ 的 $M > (1+\alpha)K$，近似误差满足 $R_{K,P_1}(F_M) = \Omega(1/K)$。该构造详见附录 G.2，其基础是在 M 个互不相交的区间上拼接二次函数 x^2 的缩放副本。虽然这里是在 μ 为均匀分布的假设下进行阐述的，但该论证可自然推广到更广泛的类别：具体而言，任意测度 μ，只要其密度 g 在某个区间上满足$g(q) \geq \lambda > 0$，都适用。在此类情况下，下界仍然成立，但其常数取决于 λ。

因此，这一构造为 Lipschitz 连续累积分布函数类别中的 $R_{K,P_1}(F)$ 提供了一个最坏情况下界。重要的是，它还凸显了学习型索引结构的一个基本特性：逼近误差取决于难以逼近区域的位置和数量。因此，必须根据数据调整划分和局部模型。预先固定模型类和预算 K 无法保证始终获得良好的逼近效果。

6.2 柯尔莫哥洛夫宽度与逼近下界

为了进一步理解学习型索引模型的局限性，而不局限于显式构造，我们转而采用逼近理论中一个更一般的框架：柯尔莫哥洛夫宽度。这一概念提供了一种方法，用于量化复杂度有界的模型对整个函数类 $\mathcal{F}$ 的逼近能力。形式上，柯尔莫哥洛夫宽度 $d_\kappa(\mathcal{F})$ 表示在$\mathcal{F}$上所能达到的最佳最坏情况逼近误差：

$$d_\kappa(\mathcal{F}) = \inf_{H \in \mathcal{H}_\kappa} \sup_{F \in \mathcal{F}} \inf_{h \in H} \|F - h\|_\mu \,,$$

其中，最外层的下确界取遍维数为 κ的所有线性子空间的集合$\mathcal{H}_\kappa$ 。

我们将 $d_\kappa(\mathcal{F})$与当前语境联系如下。在全文中，我们一直考虑使用$\mathcal{I}_{K,\mathcal{L}}$中的函数进行逼近。该类别不一定是 $L^1(\mu)$的线性子空间，因为两个 K分段函数之和不一定是 K分段函数。尽管如此，如果我们固定$\mathcal{P} = \{I_k\}_{k=1}^K$ 对Q的一个划分，则其子类

$$\mathcal{I}_{\mathcal{P},\mathcal{L}} = \left\{ h : h(q) = \textstyle\sum_k h_k(q) \mathbf{1}_{I_k}(q),\ h_k \in \mathcal{L} \right\}$$

确实构成一个线性子空间，前提是$\mathcal{L}$ 本身是 $L^1(\mu)$的线性子空间。在此类情况下——例如，当 $\mathcal{L}$ 由有界次数多项式组成时——任意函数类$\mathcal{I}_{\mathcal{P},\mathcal{L}}$ 的维数至多为 $\kappa = K \cdot \dim(\mathcal{L})$。这一刻画导出了以下不等式。

▶ **引理6**。 对于任意F、$L \subseteq L^1(\mu)$，其中L是有限维线性子空间，有 s

$$\sup_{F \in \mathcal{F}} R_{K,\mathcal{L}}(F) \geq d_\kappa(\mathcal{F}), \qquad where\ \kappa = K \cdot \dim(\mathcal{L}) \,.$$

证明见附录 G。现在我们关注$\mathcal{F}$ 为 Lipschitz连续函数类别的情形，记作 $W^{1,\infty}$。这一记号源自索伯列夫空间理论，该理论为刻画函数光滑性提供了便捷的形式化方法，且大多数 Kolmogorov 宽度结果通常都在这一理论框架下表述。当 μ 是紧区间 $[a,b]$上的均匀分布时，有（参见 [22, 第14章，定理3.8]）：

$$d_\kappa(W^{1,\infty}) = \Omega(1/\kappa) \,. \tag{7}$$

注意，对于$\mathcal{L} = P_m$这一度数为 m的多项式类别，我们有 $\kappa = (m+1)K$，因为每个分段包含 $m+1$ 个参数。分段线性模型对应于 $m = 1$，因此有 $\kappa = 2K$。因此，对于固定的 m，引理6和公式（7）意味着 $\sup_{F \in W^{1,\infty}} R_{K,P_m}(F) = \Omega(1/K)$。

这对学习型索引结构有直接影响：即使使用 K 个自适应分段和有界次数的多项式预测器的最佳模型，也无法对一般的 Lipschitz 连续累积分布函数实现 $o(1/K)$ 的近似误差。[2]

这些结果与第 6.1 节中得到的 $\Omega(1/K)$ 结果一致。虽然第 6.1 节中的显式构造可以容纳广泛的分布 μ，包括非均匀密度，但 Kolmogorov 宽度方法在近似函数方面提供了更广阔的视角。特别是，它适用于任意分段多项式类$\mathcal{I}_{K,P_m}$ 以及更一般的任意函数 κ 的线性张成，例如截断傅里叶级数。

[2] 这一结论需要谨慎看待。尽管 Kolmogorov 宽度界针对类别 $W^{1,\infty}$中的最坏情况，但并不能直接断定实现 $\Omega(1/\kappa)$ 界的最坏情况函数中包含任何累积分布函数。原则上，人们可能会担心这些最坏情况函数中不包含任何有效累积分布函数。在附录 G.3 中，我们给出了严格论证，证明任何难以逼近的函数 $F \in W^{1,\infty}$ 都可以转换为一个同样难以逼近的累积分布函数。

Kolmogorov 宽度框架明确表明，$\Omega(1/K)$ 下界本质上是一个最坏情况结果。对于更光滑的子类$\mathcal{F} \subseteq W^{1,\infty}$，该理论可以得到更精确的速率：例如，对于 $W^{2,\infty}$（即具有 Lipschitz 连续导数的函数），可以得到$d_\kappa(W^{2,\infty}) = \Theta(1/\kappa^2)$。这凸显了更强的正则性假设如何实现更快的逼近。相比之下，分段常数模型的下界（第5节）一致适用于所有 Lipschitz 连续累积分布函数，并不会因额外的正则性而改进。

结论性说明。 本节的一个关键结论是，对于类别 $\mathcal{F} = W^{1,\infty}$中的 Lipschitz连续函数，在类别$\mathcal{I}_{K,P_m}$ 的学习型索引模型下，最坏情况逼近误差对于任意固定的 m均满足 $\Omega(1/K)$。这与第5节针对分段常数逼近得出的 $\Omega(1/K)$ 速率相似，但这里仅在 $\mathcal{F}$ 上以最坏情况意义成立，而不是对所有 $F \in \mathcal{F}$ 都一致成立。特别是，该结果并不适用于更光滑的子类，例如 $W^{2,\infty}$，因为在这些子类中逼近效果可能更好；但它仍然为学习型索引的一般性能构成了一个有意义的下界。

7 讨论与结论

本文建立了一个利用学习型索引结构对查询时间给出下界的分析框架（第3–4节）。我们考察了两类最常用于实践的近似函数——分段常数函数和分段线性近似器，并将查询时间与数据生成分布的累积分布函数的逼近误差联系起来。对于分段常数函数，我们证明，对于一大类累积分布函数，逼近误差会随分段数量 K 线性下降（第5节）。对于分段线性近似器，情况更为微妙：我们证明，逼近误差取决于累积分布函数的正则性，并且一般而言，使用更丰富的逼近模型并不能保证逼近误差的下界得到改善（第6节）。然而，在对数据生成累积分布函数的光滑性作出额外假设的情况下，逼近误差的下界还可以进一步降低。

更具体地说，第5节和第6节确立了一个关于逼近误差的 $\Omega(1/K)$ 下界；对于 Lipschitz 连续累积分布函数类，这一界一般无法改进。对于分段常数模型，该结论一致成立（第5节）；对于分段线性模型，该下界则从对累积分布函数的选择所得到的最坏情况意义上成立（第6节）。在这两种情况下，$\Omega(1/K)$ 的行为都支配着最坏情况区间。将这些结果与表1中总结的分析相结合（例如，参见 E2 行和 B2 行），我们得到最坏情况查询时间的 $\Omega(\log(n/K))$ 下界。如第2节所述，这意味着分段数量K （从而空间使用量）必须在 n 中近似线性增长，才能相较于传统方法获得渐近优势；这一结论适用于广泛的学习型索引类。

将我们的界与 [34] 中的信息论结果进行比较是很有启发性的。他们的最坏情形界要求有限的查询域 $\mathcal{Q}$，而我们的框架不受此限制。此外，我们的分析为典型的学习型索引设计提供了更精确且更具适用性的见解，尤其是那些基于分段常数模型或线性模型的设计。在平均情形下，他们的结果表明，要实现期望误差 η，需要 $\log(1/\eta)$ 位的模型容量。相比之下，我们的界在独立性假设下表明，其缩放为 a$1/\eta$，当 $\eta \to 0$ 时该值更大。此外，当查询分布与数据分布相同——即 $d\mu = dF$——时，我们表明参数数量可以按 $\Omega(n)$ 缩放，而在他们的框架中则为 $\Omega(\sqrt{n})$。

除了理论意义之外，我们的结果还为学习型索引的设计提供了一些实践启示：首先，分段数量 K 不应

在不考虑底层数据的情况下选定，因为对于给定的 K，存在一些“对抗性”分布会导致较大的逼近误差（参见第 6.1 节）。其次，第 6.2 节的结果表明，近似能力不仅取决于模型类的表达能力，也取决于累积分布函数的光滑性。这表明，以依赖数据的方式选择模型类——使其适应累积分布函数的正则性——可能有助于设计出更高效的索引；这一想法值得进一步探索。

仍有若干值得探索的方向：将平均情形分析扩展到二分搜索；放宽对数据的独立性假设；考虑更广泛的索引结构类别；以及分析学习型索引在更新或动态工作负载下的性能。我们希望这项工作能提供一些基础工具，助力进一步探索这些想法。

参考文献

1 Alfred V. Aho 和 John E. Hopcroft。《计算机算法的设计与分析》。Addison-Wesley Longman Publishing Co., Inc.，美国，第一版，1974。2 Albert Atserias、Martin Grohe 和 Dániel Marx。关系连接的规模界与查询计划。《SIAM计算杂志》，42(4):1737–1767，2013。 arXiv:https://doi.org/10.1137/110859440, doi:10.1137/110859440。3 Luis Alberto Croquevielle、Guang Yang、Liang Liang、Ali Hadian 和 Thomas Heinis。超越对数界：利用学习型索引实现常数期望时间查询。载于 Sudeepa Roy 和 Ahmet Kara 编，第二十八届国际数据库理论会议（ICDT 2025），第328卷《莱布尼茨国际信息学会议录》（LIPIcs），第19:1–19:21页，德国达格施图尔，2025。Schloss Dagstuhl – Leibniz-Zentrum für Informatik。网址： https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICDT.2025.19, doi:10.4230/LIPIcs.ICDT.2025.19。4 Sándor CSöRgő 和 Julian J. Faraway。Cramér–von Mises统计量的精确分布与渐近分布。《英国皇家统计学会杂志：B辑（方法学）》，58(1):221–234，2018年12月。 arXiv:https://academic.oup.com/jrsssb/article-pdf/58/1/221/49098785/jrsssb_58_1_221.pdf, doi:10.1111/j.2517-6161.1996.tb02077.x。5 Jialin Ding、Umar Farooq Minhas、Jia Yu、Chi Wang、Jaeyoung Do、Yinan Li、Hantian Zhang、Badrish Chandramouli、Johannes Gehrke、Donald Kossmann、David Lomet 和 Tim Kraska。ALEX：一种可更新的自适应学习型索引。载于《2020年ACM SIGMOD数据管理国际会议论文集》，SIGMOD '20，第969–984页，美国纽约州纽约，2020。计算机协会。 doi:10.1145/3318464.3389711。6 Jialin Ding、Vikram Nathan、Mohammad Alizadeh 和 Tim Kraska。Tsunami：面向相关数据和偏斜工作负载的学习型多维索引。VLDB基金会会刊，14(2):74–86，2020年10月。 doi:10.14778/3425879.3425880。7 Lawrence C. Evans。《偏微分方程》。美国数学学会，普罗维登斯，罗得岛州，2010。8 Paolo Ferragina、Fabrizio Lillo 和 Giorgio Vinciguerra。《学习型索引为何如此有效？》。载于 Hal Daumé III 和 Aarti Singh 编，《第37届国际机器学习会议论文集》，第119卷《机器学习研究会议录》，第3123–3132页，线上，2020年7月13–18日。PMLR。网址： http://proceedings.mlr.press/v119/ferragina20a.html。9 Paolo Ferragina 和 Giorgio Vinciguerra。PGM-Index：一种具有可证明最坏情况界的全动态压缩学习型索引。VLDB基金会会刊，13(8):1162–1175，2020年4月。 doi:10.14778/3389133.3389135。10 R. A. Finkel 和 J. L. Bentley。四叉树：用于复合键检索的数据结构。《信息学学报》，4(1):1–9，1974年3月。 doi:10.1007/BF00288933。11 Alex Galakatos、Michael Markovitch、Carsten Binnig、Rodrigo Fonseca 和 Tim Kraska。Fiting-tree：一种数据感知索引结构。载于《2019年国际数据管理会议论文集

数据管理》，SIGMOD '19，第1189–1206页，美国纽约州纽约，2019。计算机协会。doi:10.1145/3299869.3319860。12 Goetz Graefe 等。现代B树技术。数据库基础与趋势，3(4):203–402，2011。doi:10.1561/1900000028。13 Siegfried Graf 和 Harald Luschgy。《概率分布量化基础》。施普林格科学与商业媒体，2000。14 Antonin Guttman。R树：用于空间搜索的动态索引结构。载于《1984年ACM SIGMOD数据管理国际会议论文集》，SIGMOD '84，第47–57页，美国纽约州纽约，1984。计算机协会。doi:10.1145/602259.602266。15 Ali Hadian 和 Thomas Heinis。Shift-table：一种使用模型校正进行范围查询的低延迟学习型索引。载于 Yannis Velegrakis、Demetris Zeinalipour-Yazti、Panos K. Chrysanthis 和 Francesco Guerra 编，《第24届国际数据库技术拓展会议论文集》，EDBT 2021，塞浦路斯尼科西亚，2021年3月23–26日，第253–264页。OpenProceedings.org，2021。doi:10.5441/002/edbt.2021.23。16 Andreas Kipf、Ryan Marcus、Alexander van Renen、Mihail Stoian、Alfons Kemper、Tim Kraska 和 Thomas Neumann。Radixspline：一种单遍学习型索引。载于《第三届利用人工智能技术进行数据管理国际研讨会论文集》，aiDM '20，美国纽约州纽约，2020。计算机协会。doi:10.1145/3401071.3401659。17 Paraschos Koutris、Shaleen Deep、Austen Fan 和 Hangdong Zhao。追求更快的连接算法。载于 Sudeepa Roy 和 Ahmet Kara 编，第二十八届国际数据库理论会议（ICDT 2025），第328卷《莱布尼茨国际信息学会议录》（LIPIcs），第1:1–1:12页，德国达格施图尔，2025。Schloss Dagstuhl – Leibniz-Zentrum für Informatik。网址：https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICDT.2025.1，doi:10.4230/LIPIcs.ICDT.2025.1。18 Tim Kraska、Alex Beutel、Ed H. Chi、Jeffrey Dean 和 Neoklis Polyzotis。学习型索引结构的价值。载于《2018年国际数据管理会议论文集》，SIGMOD '18，第489–504页，美国纽约州纽约，2018。计算机协会。doi:10.1145/3183713.3196909。19 Viktor Leis、Andrey Gubichev、Atanas Mirchev、Peter Boncz、Alfons Kemper 和 Thomas Neumann。查询优化器究竟有多好？VLDB基金会会刊，9(3):204–215，2015年11月。doi:10.14778/2850583.2850594。20 Pengfei Li、Hua Lu、Qian Zheng、Long Yang 和 Gang Pan。Lisa：一种面向空间数据的学习型索引结构。载于《2020年ACM SIGMOD数据管理国际会议论文集》，SIGMOD '20，第2119–2133页，美国纽约州纽约，2020。计算机协会。doi:10.1145/3318464.3389703。21 Liang Liang、Guang Yang、Ali Hadian、Luis Alberto Croquevielle 和 Thomas Heinis。Swix：一种内存高效的滑动窗口学习型索引。ACM数据管理会刊，2(1)，2024年3月。doi:10.1145/3639296。22 G. Lorentz, George 和 Manfred V. Golitschek。《构造逼近：进阶问题》。施普林格出版社，柏林，1996。23 Ryan Marcus、Andreas Kipf、Alexander van Renen、Mihail Stoian、Sanchit Misra、Alfons Kemper、Thomas Neumann 和 Tim Kraska。学习型索引基准测试。VLDB基金会会刊，14(1):1–13，2020年9月。doi:10.14778/3421424.3421425。24 P. Massart。Dvoretzky-Kiefer-Wolfowitz不等式中的紧常数。概率年刊，18(3):1269–1283，1990。doi:10.1214/aop/1176990746。25 Vikram Nathan、Jialin Ding、Mohammad Alizadeh 和 Tim Kraska。学习多维索引。载于《2020年ACM SIGMOD数据管理国际会议论文集》，SIGMOD '20，第985–1000页，美国纽约州纽约，2020。计算机协会。doi:10.1145/3318464.3380579。26 Theodore J Rivlin。《函数逼近导论》。Dover，纽约，1981。

27 Dan Suciu。信息不等式在数据库理论问题中的应用。载于第38届 ACM/IEEE计算机科学逻辑研讨会（LICS 2023），美国马萨诸塞州波士顿，2023年6月26日至29日，第1–30页。IEEE，2023。doi:10.1109/LICS56636.2023.10175769。28 Zhaoyan Sun、Xuanhe Zhou 和 Guoliang Li。学习型索引：全面的实验评估。VLDB基金会会刊，16(8):1992–2004，2023年4月。doi:10.14778/3594512.3594528。29 Jiacheng Wu、Yong Zhang、Shimin Chen、Jin Wang、Yu Chen 和 Chunxiao Xing。具有精确位置的可更新学习型索引。VLDB基金会会刊，14(8):1276–1288，2021年4月。doi:10.14778/3457390.3457393。30 Guang Yang、Liang Liang、Ali Hadian 和 Thomas Heinis。FLIRT：适用于滚动时间窗口的快速学习型索引。载于 Julia Stoyanovich、Jens Teubner、Nikos Mamoulis、Evaggelia Pitoura、Jan Mühlig、Katja Hose、Sourav S. Bhowmick 和 Matteo Lissandrini 编，《第26届国际数据库技术拓展会议论文集》（EDBT 2023），希腊约阿尼纳，2023年3月28日至31日，第234–246页，希腊约阿尼纳，2023年。OpenProceedings.org。doi:10.48786/edbt.2023.19。31 Mihalis Yannakakis。无环数据库模式的算法。载于第七届国际超大型数据库会议论文集——第7卷（VLDB '81），第82–94页。VLDB基金会，1981年。32 Sepanta Zeighami 和 Cyrus Shahabi。关于依赖于分布的学习型索引亚对数查询时间。载于第40届国际机器学习会议论文集（ICML'23）。JMLR.org，2023年。33 Sepanta Zeighami 和 Cyrus Shahabi。通过分布可学习性对分布偏移下的学习型数据库操作进行理论分析。载于第41届国际机器学习会议论文集（ICML'24）。JMLR.org，2024年。34 Sepanta Zeighami 和 Cyrus Shahabi。朝着为学习型数据库操作建立保证误差迈进。载于第十二届国际学习表征会议（ICLR 2024），奥地利维也纳，2024年5月7日至11日。OpenReview.net，2024年。URL: https://openreview.net/forum?id=6tqgL8VluV。

缩放引理和命题1的证明

▶ **引理7**（近似误差的缩放）。 设 $\lambda > 0$，并假设模型类 L 在标量乘法下不变。那么，对于任意函数 $G \in L^1(\mu)$ 和任意 $K \in \mathbb{N}$，

$$R_{K,\mathcal{L}}(\lambda G) = \lambda \cdot R_{K,\mathcal{L}}(G)\,,$$

其中 $R_{K,\mathcal{L}}(G) = \text{下确界}_{h \in \mathcal{I}_{K,\mathcal{L}}} \|G - h\|_\mu$。

证明。 设 $\mathcal{H} = \mathcal{I}_{K,\mathcal{L}}$ 表示定义在 K 个区间上、并使用来自 $\mathcal{L}$ 的函数的分段模型类。由于 $\mathcal{L}$ 在标量乘法下不变，因此 $\mathcal{H}$ 也具有这一性质。也就是说，对于任意$h \in \mathcal{H}$ 和任意 $\lambda > 0$，函数 λh 也属于 $\mathcal{H}$。现在观察：

$$R_{K,\mathcal{L}}(\lambda G) = \inf_{h \in \mathcal{H}} \|\lambda G - h\|_\mu = \inf_{h \in \mathcal{H}} \|\lambda (G - \lambda^{-1} h)\|_\mu = \inf_{h \in \mathcal{H}} \lambda \cdot \|(G - \lambda^{-1} h)\|_\mu\,.$$

利用下确界的正齐次性：

$$R_{K,\mathcal{L}}(\lambda G) = \lambda \cdot \inf_{h \in \mathcal{H}} \|G - \lambda^{-1} h\|_\mu\,.$$

由于$\mathcal{H}$ 在标量乘法下保持不变，因此对于任意 $h \in \mathcal{H}$，都有$\lambda^{-1} h \in \mathcal{H}$。此外，任意 $h \in \mathcal{H}$ 都可以写成 $\lambda^{-1}\tilde{h}$，其中 $\tilde{h} \in \mathcal{H}$。这意味着映射 $h \mapsto \lambda^{-1} h$ 是$\mathcal{H}$上的一个双射，因此

$$R_{K,\mathcal{L}}(\lambda G) = \lambda \cdot \inf_{\tilde{h} \in \mathcal{H}} \|G - \tilde{h}\|_\mu = \lambda \cdot R_{K,\mathcal{L}}(G)\,.$$

◀

命题 1 的证明。 在 $L^1(\mu)$ 中应用三角不等式，对于任意 $h \in \mathcal{I}_{K,\mathcal{L}}$，有：

$$\|F - h\|_\mu \leq \|F - F_n\|_\mu + \|F_n - h\|_\mu \quad \Longrightarrow \quad \|F_n - h\|_\mu \geq \|F - h\|_\mu - \|\delta_n\|_\mu .$$

在最后一个不等式中对 $h \in \mathcal{I}_{K,\mathcal{L}}$ 取下确界，得到：

$$R_{K,\mathcal{L}}(F_n) \geq R_{K,\mathcal{L}}(F) - \|\delta_n\|_\mu .$$

然后，根据缩放恒等式 $R_{K,\mathcal{L}}(nF_n) = nR_{K,\mathcal{L}}(F_n)$（引理7）以及期望预测误差 $\mathbb{E}_q[\varepsilon(q)] = \|nF_n - h\|_\mu \geq R_{K,\mathcal{L}}(nF_n)$ 的定义，即可得出结果。◀

B 命题 2 的证明

本节旨在证明命题 2。为清晰起见，本节中使用的一些辅助结果在附录C中证明。我们首先回顾 $T(q)$ 的定义：

$$\widetilde{T}(q) = \log_2 \left(\max\{2, \varepsilon(q)\}\right) = \log_2 \left(\max\{2, |nF_n(q) - h(q)|\}\right) .$$

最终，我们关注的是对 $\mathbb{E}_q\left[\widetilde{T}(q)\right]$ 给出下界。为此，首先回顾一下，根据假设，F 和 μ 分别具有密度函数 f 和 g，且满足

$$0 < c_F \leq f(q), g(q) \leq C_F < \infty \tag{8}$$

其中常数为 c_F, C_F，并且对所有 $q \in \mathcal{Q}$ 均成立。命题 2（以及本证明）可以直接推广到如下情形：该性质仅在具有正测度的连续区间 $\mathcal{Q}' \subseteq \mathcal{Q}$ 上成立。不过，为了便于说明，我们假设式 (8) 对所有 $q \in \mathcal{Q}$ 均成立。

注意，不失一般性，我们可以假设 $d\mu = dF = f\,dq$。这可由下式推出：

$$\mathbb{E}_q\left[\widetilde{T}(q)\right] = \int_{\mathcal{Q}} \widetilde{T}(q) d\mu = \int_{\mathcal{Q}} \widetilde{T}(q)\, g(q) dq \geq \frac{c_F}{C_F} \int_{\mathcal{Q}} \widetilde{T}(q)\, f(q) dq ,$$

这里我们用到了 $d\mu = g\,dq$ 以及式 (8) 中给出的不等式。因此，如果我们能像命题 2 中所述，对最后一个积分 $\int_{\mathcal{Q}} T(q)\ f(q) dq$ 给出下界，那么该界也适用于 $\mathbb{E}_q\left[\widetilde{T}(q)\right]$，只需额外乘上一个（常数）因子 $\frac{c_F}{C_F}$。因此，在本节中，我们始终用 dF 代替 $d\mu$。

最后，我们假设 $nR_{K,\ P_0} \geq 1$。否则，由于命题 2 的右端为负，命题 2 显然成立。基于这些假设，我们可以开始证明。

命题 2 的证明。 我们将证明分为四个逻辑步骤。

步骤 1：分段展开。 任意 $h \in \mathcal{I}_{K,P_0}$ 均可展开为

$$h(q) = n \sum_{k=1}^{K} c_k \mathbf{1}_{I_k}(q),$$

其中$\{c_k\}$ 为常数，$\{I_k\}$ 为划分 $\mathcal{Q}$ 的区间。现在，关注的量可以给出如下界：

$$\begin{aligned}\mathbb{E}_q\big[\widetilde{T}(q)\big] &= \int_{\mathcal{Q}} \log_2\left(\max\{2,\varepsilon(q)\}\right) dF \\ &= \int_{\mathcal{Q}} \log_2\left(\max\left\{2,|nF_n(q)-h(q)|\right\}\right) dF \\ &\geq \int_{\mathcal{Q}} \log_2\left(\max\left\{2,n|F(q)-n^{-1}h(q)|-n|\delta_n(q)|\right\}\right) dF \\ &= \sum_{k=1}^{K} \underbrace{\int_{I_k} \log_2\left(\max\left\{2,n|F(q)-c_k|-n|\delta_n(q)|\right\}\right) dF}_{E_k}, \end{aligned} \tag{9}$$

其中我们使用三角不等式对 $\varepsilon(q)$进行下界估计：

$$\begin{aligned}\varepsilon(q) &= |nF_n(q)-h(q)| \\ &= |nF_n(q)-nF(q)+nF(q)-h(q)| \\ &\geq n|F(q)-n^{-1}h(q)|-n|F_n(q)-F(q)| \\ &= n|F(q)-n^{-1}h(q)|-n|\delta_n(q)|\,.\end{aligned}$$

现在，不失一般性，我们可以假设 h 在所有区间内都与 F 相交，即对于每个 k ，存在 $q_k \in I_k$ 使得 $F(q_k) = c_k$。否则，将存在另一个不同的 $h' \in \mathcal{I}_{K,P_0}$使得 $|F(q) - n^{-1}h'(q)| \leq |F(q) - n^{-1}h(q)|$对于所有 $q \in \mathcal{Q}$均成立，这意味着

$$\begin{aligned}\mathbb{E}_q\big[\widetilde{T}(q)\big] &\geq \int_{\mathcal{Q}} \log_2\left(\max\left\{2,n|F(q)-n^{-1}h(q)|-n|\delta_n(q)|\right\}\right) dF \\ &\geq \int_{\mathcal{Q}} \log_2\left(\max\left\{2,n|F(q)-n^{-1}h'(q)|-n|\delta_n(q)|\right\}\right) dF\,,\end{aligned}$$

因此，我们为 h' 情形推导出的任何下界同样适用于 h。所以，我们假设存在具有上述性质的$\{q_k\}_{k=1}^K$。

根据这些推导可以清楚地看出，对于我们的证明，只需对式（9）中定义的表达式 $\sum_{k=1}^K E_k$ 进行下界估计，其中假设每个 c_k 都至少在一点 $q_k \in I_k$处与 F 相交。在证明的下一部分，我们将重点对每个单独的项 E_k进行下界估计，该项为

$$E_k = \int_{I_k} \log_2\left(\max\left\{2,n|F(q)-c_k|-n|\delta_n(q)|\right\}\right) dF\,. \tag{10}$$

为此，我们首先定义一个设定，使项 $n|\delta_n(q)|$ 可以忽略，然后为剩余项 $n|F(q) - c_k|$ 推导一些集中性结果。最后，在证明的最后一部分，我们转而处理总误差 $\sum_{k=1}^K E_k$。

步骤 2： δ_n **较小的一个大测度集合**。现在，定义集合

$$A = \left\{q \in \mathcal{Q} \text{ such that } |\delta_n(q)| \leq \frac{1}{n}\right\}.$$

由于 $d\mu = dF$，根据引理 4(b)，存在 $X_1, \ldots, X_n$ 的一个实现，使得 $\|\delta_n\|_\mu \leq \frac{1}{\sqrt{6}\,n}$。在证明的其余部分，我们都基于这一实现。根据马尔可夫不等式：

$$\mu\left(|\delta_n| > \frac{1}{n}\right) \leq \frac{\|\delta_n\|_\mu}{\frac{1}{n}} \leq \frac{n}{\sqrt{6}\,n} \leq \frac{1}{2}\,,$$

这意味着 $\mu(A) \geq 1/2$。现在，对每个 $k = 1, \ldots\ldots, K$，定义集合 $J_k = I_k \cap A$，即区间 I_k 中满足 $|\delta_n| \leq \frac{1}{n}$ 的子集。从 (10) 出发，我们有

$$\begin{aligned} E_k &= \int_{I_k} \log_2 \left(\max \left\{ 2, n|F(q) - c_k| - n|\delta_n(q)| \right\} \right) dF \\ &\geq \int_{J_k} \log_2 \left(\max \left\{ 2, n|F(q) - c_k| - n|\delta_n(q)| \right\} \right) dF \\ &\geq \frac{1}{2} \int_{J_k} \log_2 \left(2 + n|F(q) - c_k| - n|\delta_n(q)| \right) dF \\ &\geq \frac{1}{2} \int_{J_k} \log_2 \left(1 + n|F(q) - c_k| \right) dF \,, \end{aligned} \tag{11}$$

其中我们利用了以下事实：$|\delta_n| \leq \frac{1}{n}$ 在每个集合 J_k 上。

步骤 3：关于近似误差的集中性结果。现在我们着重为表达式 (11) 推导下界。为此，我们定义 $|F(q) - c_k|$ 在集合 J_k 上的一阶矩和二阶矩：

$$m_k = \int_{J_k} |F(q) - c_k| \, dF \,, \qquad s_k = \int_{J_k} |F(q) - c_k|^2 dF \,.$$

接下来，记 J_k' 为满足 $|F(q) - c_k| \geq \frac{m_k}{2\mu(J_k)}$ 的集合，也就是说，这是 J_k 的子集，其中量 $|F(q) - c_k|$ 与其在 J_k 上的归一化平均值相比并没有小很多。由于 $\log_2(1 + n|F(q) - c_k|)$ 恒非负，因此有

$$\begin{aligned} E_k &\geq \frac{1}{2} \int_{J_k} \log_2 \left(1 + n|F(q) - c_k| \right) dF \\ &\geq \frac{1}{2} \int_{J_k'} \log_2 \left(1 + n|F(q) - c_k| \right) dF \\ &\geq \frac{\mu(J_k')}{2} \log_2 \left(1 + \frac{n m_k}{2\mu(J_k)} \right) \\ &\geq \frac{m_k^2}{8\, s_k} \cdot \log_2 \left(1 + \frac{n m_k}{2\mu(J_k)} \right) \,, \end{aligned} \tag{12}$$

最后一步由 Paley–Zygmund不等式（见引理8）得出。因此，剩下的唯一挑战是为量 m_k 推导下界，并为 s_k 推导上界。为此，我们利用 F 的双Lipschitz性质：由于 f 有界，即 $c_F \leq f \leq C_F$，这直接意味着

$$c_F|x - y| \leq |F(x) - F(y)| \leq C_F|x - y| \tag{13}$$

对于所有 x, $y \in \mathcal{Q}$。现在，回顾我们的假设：对于每个 k，都存在 $q_k \in I_k$，使得 $F(q_k) = c_k$。因此，我们可以利用 $y = q_k$ 得到不等式 (13)，对于所有 $q \in I_k$：

$$\begin{aligned} |F(q) - c_k| &= |F(q) - F(q_k)| \geq c_F|q - q_k| \,, \\ |F(q) - c_k| &= |F(q) - F(q_k)| \leq C_F|q - q_k| \,. \end{aligned}$$

现在，在满足上述最后这些条件时，我们可以应用引理 9，得到 m_k 和 s_k 的界：

$$m_k \geq \frac{c_F^2}{4C_F^2} \, \mu(J_k)^2 \,, \qquad s_k \leq \frac{C_F^3}{3c_F^3} \, \mu(I_k)^3 \,.$$

现在，我们将这些值代入式 (12) 中的帕利－齐格蒙德界，得到：

$$E_k \geq 3\left(\frac{c_F}{2C_F}\right)^7 \frac{\mu(J_k)^4}{\mu(I_k)^3}\log_2\left(1+\frac{1}{2}\left(\frac{c_F}{2C_F}\right)^2 n\,\mu(J_k)\right). \tag{14}$$

这为每个 E_k 给出了一个显式的、完全非渐近的下界，其形式由 c_F, C_F, n 以及测度 $\mu(J_k)$ 和 $\mu(I_k)$ 决定。在证明的最后一部分，我们纳入有关这些测度的更多信息，并对所有 $k = 1,\ldots,K$ 进行汇总。

步骤4：对 k 求和并进行最终汇总。 我们首先根据各集合的测度相对于对应原始 I_k的下降程度，将集合 $\{J_k\}$划分为两个不同的组。具体而言，我们将索引集 $\mathcal{K}$ 定义为

$$\mathcal{K} = \{k \in \{1,\ldots,K\} \text{ such that } \mu(J_k) \geq \mu(I_k)/3\}.$$

一个关键观察是，对 $\mathcal{K}$ 之外的所有 k 求和时，我们得到

$$\sum_{k\notin\mathcal{K}} \mu(J_k) < \frac{1}{3}\sum_{k\notin\mathcal{K}} \mu(I_k) \leq \frac{1}{3}\mu(\mathcal{Q}) = \frac{1}{3}.$$

由此，我们推出一个关键条件：

$$\sum_{k\in\mathcal{K}} \mu(J_k) = \sum_{k=1}^{K}\mu(J_k) - \sum_{k\notin\mathcal{K}}\mu(J_k) = \mu(A) - \sum_{k\notin\mathcal{K}}\mu(J_k) \geq \frac{1}{2} - \frac{1}{3} = \frac{1}{6}, \tag{15}$$

其中我们用到了 $\mu(A) \geq 1/2$ 这一事实，该事实此前由马尔可夫不等式推得。注意，这必然意味着 $|\mathcal{K}|$ 非空；因此，我们可以假设 $|\mathcal{K}| \geq 1$。现在，根据定义，对于任意 $k \in \mathcal{K}$，都有 $3\mu(J_k) \geq \mu(I_k)$。结合式 (14)，并记$\gamma = \frac{c_F}{2C_F}$：

$$\begin{aligned}
\sum_{k=1}^{K} E_k &\geq \sum_{k\in\mathcal{K}} E_k \\
&\geq \frac{\gamma^7}{9}\sum_{k\in\mathcal{K}} \mu(J_k)\log_2\left(1+\frac{\gamma^2 n}{2}\mu(J_k)\right) \\
&\geq \frac{\gamma^7}{9}\sum_{k\in\mathcal{K}} \mu(J_k)\log_2\left(\frac{\gamma^2 n}{2}\mu(J_k)\right) \\
&= \frac{\gamma^7}{9}\left[\log_2\left(\frac{\gamma^2 n}{2}\right)\sum_{k\in\mathcal{K}}\mu(J_k) + \sum_{k\in\mathcal{K}}\mu(J_k)\log_2\left(\mu(J_k)\right)\right].
\end{aligned}$$

记 S 为和式 $S = \sum_{k\in\mathcal{K}}\mu(J_k)$。于是，根据引理10，我们得到：

$$\begin{aligned}
\sum_{k=1}^{K} E_k &\geq \frac{\gamma^7}{9}\left[S\log_2\left(\frac{\gamma^2 n}{2}\right) + S\log_2\left(\frac{S}{|\mathcal{K}|}\right)\right] \\
&= \frac{\gamma^7}{9}\left[S\log_2\left(\frac{n}{4|\mathcal{K}|}\right) + S\log_2(2S) + S\log_2\left(\gamma^2\right)\right].
\end{aligned}$$

现在，对于 $S > 0$，函数 $S \mapsto S\log_2(2S)$ 永远不小于 $-0.5 = \log_2(1/\sqrt{2})$。因此：

$$\geq \frac{\gamma^7}{9}\left[S\log_2\left(\frac{n}{4|\mathcal{K}|}\right) + \log_2\left(\frac{1}{\sqrt{2}}\right) + S\log_2\left(\gamma^2\right)\right].$$

另一方面，我们知道 $\gamma^2 < 1$，因此 $\log_2(\gamma^2) < 0$。由于 $S \le \mu(\mathcal{Q}) = 1$:

$$\geq \frac{\gamma^7}{9}\left[S\log_2\left(\frac{n}{4|\mathcal{K}|}\right) + \log_2\left(\frac{\gamma^2}{\sqrt{2}}\right)\right].$$

最后，考虑这一事实：$|\mathcal{K}| \le K$。此外，根据第5.1节，我们知道当$d\mu = dF$时，有$R_{K,P_0} = \frac{1}{4K}$（见公式 (6)）。代入上式：

$$\sum_{k=1}^{K} E_k \geq \frac{\gamma^7}{9}\left[S\log_2(nR_{K,P_0}) + \log_2\left(\frac{\gamma^2}{\sqrt{2}}\right)\right].$$

现在，根据假设，我们有$nR_{K,P_0} \geq 1$。因此，$\log_2(nR_{K,P_0}) \geq 0$，由 (15) 可得：

$$\begin{aligned}\sum_{k=1}^{K} E_k &\geq \frac{\gamma^7}{9}\left[\frac{1}{6}\log_2(nR_{K,P_0}) + \log_2\left(\frac{\gamma^2}{\sqrt{2}}\right)\right] \\ &\geq \frac{\gamma^7}{54}\left[\log_2(nR_{K,P_0}) - \log_2\left(\frac{8}{\gamma^{12}}\right)\right].\end{aligned}$$

证明通过令 $C_1 = \frac{\gamma^7}{54}$ 和 $C_2 = \log_2\left(\frac{8}{\gamma^{12}}\right)$得以完成。 ◀

C 附录 B 中的辅助结果

▶ **引理8** (Paley–Zygmund不等式)。 考虑任意测度 μ 于 $\mathbb{R}$上。设 $J \subseteq \mathbb{R}$ 是一个可测集，且 $\mu(J) > 0$，并设 $G: J \to [0,\infty)$ 是一个可测的非负函数。定义它在 J 上的一阶矩和二阶矩为

$$m_J = \int_J G(q)\,d\mu, \qquad s_J = \int_J G(q)^2 d\mu.$$

那么，对于任意 $\theta \in (0, 1)$，子集

$$J'_\theta = \left\{q \in J \ \ such\ that\ \ G(q) \geq \theta\frac{m_J}{\mu(J)}\right\}$$

满足下界

$$\mu(J'_\theta) \geq (1-\theta)^2\frac{m_J^2}{s_J}.$$

证明。 注意，一阶矩可以写成

$$m_J = \int_J G(q)\,d\mu = \int_J G(q)\,\mathbf{1}_{J'_\theta}(q)\,d\mu + \int_J G(q)\,\mathbf{1}_{J\setminus J'_\theta}(q)\,d\mu, \tag{16}$$

其中 $\mathbf{1}_U$ 表示集合 U的示性函数。右侧第一项可通过柯西–施瓦茨不等式进行估计：

$$\int_J G(q)\,\mathbf{1}_{J'_\theta}(q)\,d\mu \leq \left(\int_J \mathbf{1}_{J'_\theta}(q)\,d\mu\right)^{\frac{1}{2}}\left(\int_J G^2(q)d\mu\right)^{\frac{1}{2}} = \sqrt{\mu(J'_\theta)\cdot s_J}.$$

另一方面，式 (16) 中的最后一项可由以下上界限定：

$$\int_J G(q)\,\mathbf{1}_{J\setminus J'_\theta}(q)\,d\mu = \int_{J\setminus J'_\theta} G(q)\,d\mu \leq \int_{J\setminus J'_\theta}\theta\frac{m_J}{\mu(J)}\,d\mu \leq \theta\frac{m_J}{\mu(J)}\mu(J) = \theta\cdot m_J.$$

将这些界代入式 (16)，可得

$$m_J \le \sqrt{\mu(J'_\theta)\cdot s_J} + \theta \cdot m_J\,.$$

两边减去 $\theta\cdot m_J$，平方后再除以 s_J，即可得到该结果。◀

▶**引理 9.** 设 $I \subseteq \mathbb{R}$ 为一个有限区间，$J \subseteq I$ 为 I 的一个正测度可测子集 $\mu(J) > 0$，其中 μ 由密度 f 定义，且该密度满足$0 < c_F \le f(q) \le C_F < \infty$，对所有 $q \in I$均成立。假设 $F\colon J \to \mathbb{R}$ 是一个可测函数，且存在一点 $q_0 \in I$ 满足双 *Lipschitz* 条件

$$c_F|q-q_0| \le |F(q)-F(q_0)| \le C_F|q-q_0| \tag{17}$$

对所有 $q \in I$均成立，其中 c_F、$C_F > 0$ 是与上文相同的常数。然后，定义关于 μ、在 J 上的一阶矩和二阶矩：

$$m_J = \int_J |F(q)-F(q_0)|\,d\mu\,, \qquad s_J = \int_J |F(q)-F(q_0)|^2\,d\mu\,,$$

我们有以下界：

$$m_J \ge \frac{c_F^2}{4C_F^2}\mu(J)^2\,, \qquad s_J \le \frac{C_F^3\,\mu(I)^3}{3c_F^3}\,.$$

证明。 由双 Lipschitz 条件 (17)，可得一阶矩的以下界：

$$m_J = \int_J |F(q)-F(q_0)|\,d\mu \ge c_F\int_J |q-q_0|\,d\mu\,.$$

由于 $d\mu = f\,dq$ 和 $f \ge c_F$，我们有：

$$m_J \ge c_F\int_J |q-q_0|\,f(q)dq \ge c_F^2\int_J |q-q_0|\,dq\,.$$

当 J 是一个以 q_0为中心、长度为 $\lambda(J)$的对称区间时，右侧达到最小值，其中 λ 表示勒贝格测度。在这种情况下，积分变为

$$m_J \ge c_F^2\int_{-\lambda(J)/2}^{\lambda(J)/2} |q|\,dq = 2\,c_F^2\int_0^{\lambda(J)/2} q\,dq = \frac{c_F^2}{4}\lambda(J)^2\,.$$

最后，注意到 $\lambda(J) = \int_J 1dq \ge \frac{1}{C_F}\int_J f(q)dq = \frac{\mu(J)}{C_F}$，从而得到所需的界：

$$m_J \ge \frac{c_F^2}{4C_F^2}\mu(J)^2\,.$$

对于二阶矩 s_J，我们首先使用（17）中的上界，这意味着

$$s_J = \int_J |F(q)-F(q_0)|^2 d\mu \le C_F^2\int_J |q-q_0|^2\,d\mu\,.$$

由于 $d\mu = f\,dq$ 和 $f \le C_F$，我们有：

$$s_J \le C_F^2\int_J |q-q_0|^2\,f(q)dq \le C_F^3\int_J |q-q_0|^2\,dq\,.$$

当 $J = I$ 和 q_0 位于其某个端点时，这个最后的量达到最大值。这对应于以下积分：

$$s_J \leq C_F^3 \int_0^{\lambda(I)} q^2 dq = \frac{C_F^3}{3} \lambda(I)^3 ,$$

其中，λ 表示勒贝格测度。最后，注意到 $\lambda(I) = \int_I 1 dq \leq \frac{1}{c_F} \int_I f(q) dq = \mu(I)\, c_F$，由此得到所需的界：

$$s_J \leq \frac{C_F^3}{3 c_F^3} \mu(I)^3 .$$

证明至此完成。 ◀

▶ **引理10**（部分质量下的熵界）。设 K 为正整数，且 $p_1, \ldots, p_K$ 为非负实数，使得 $\sum_{k=1}^K p_k = S > 0$。则有

$$\sum_{k=1}^K p_k \log_2(p_k) \geq S \log_2 \left(\frac{S}{K} \right) .$$

证明。定义归一化后的值 $q_k = \frac{p_k}{S}$，对所有 $k = 1, \ldots, K$。于是有 $q_k \geq 0$ 且 $\sum_{k=1}^K q_k = 1$，即 $\{q_k\}_{k=1}^K$ 构成一个离散概率分布。我们计算：

$$\begin{aligned} \sum_{k=1}^K p_k \log_2(p_k) &= \sum_{k=1}^K S\, q_k \log_2(S\, q_k) \\ &= S \sum_{k=1}^K q_k \log_2(S) + S \sum_{k=1}^K q_k \log_2(q_k) \\ &= S \log_2(S) + S \sum_{k=1}^K q_k \log_2(q_k) . \end{aligned} \tag{18}$$

术语 $\sum_{k=1}^K q_k \log_2(q_k)$ 可视为（负的）香农熵，它在均匀分布下达到最大值（即 $q_k = 1/K$）。这意味着：

$$-\sum_{k=1}^K q_k \log_2(q_k) \leq \log_2(K) \implies \sum_{k=1}^K q_k \log_2(q_k) \geq \log_2 \left(\frac{1}{K} \right) .$$

将这一最后的不等式应用于公式（18），可得：

$$\begin{aligned} \sum_{k=1}^K p_k \log_2(p_k) &\geq S \log_2(S) + S \log_2 \left(\frac{1}{K} \right) \\ &= S \log_2 \left(\frac{S}{K} \right) , \end{aligned}$$

如所要证明。 ◀

D 第4.2节的证明

研究经验过程的一项重要结果是 Dvoretzky–Kiefer–Wolfowitz（DKW）不等式 [24]，我们在此重述该不等式。若 $X_1, \ldots, X_n$ 独立同分布，且 $\delta_n = F - F_n$ 是真实累积分布函数 F 与经验累积分布函数 F_n 之差，则对于任意 $t > 0$，均有

$$\mathbb{P}\left(\|\delta_n\|_\infty > t \right) \leq 2 e^{-2nt^2} .$$

23:26 学习索引的算法复杂度下界

我们在引理 3 的证明中使用这一结果。

引理 3 的证明。 （a）部分由期望值的尾积分公式和 DKW 不等式 [24]得出：

$$\mathbb{E}_X\left[\|\delta_n\|_\infty\right] = \int_0^\infty \mathbb{P}\left(\|\delta_n\|_\infty > t\right) dt \le \int_0^\infty 2e^{-2nt^2} dt\,.$$

代入 $u = 2t\sqrt{n},\quad du = 2\sqrt{n}\,dt$ 可得：

$$\mathbb{E}_X\left[\|\delta_n\|_\infty\right] \le \frac{1}{\sqrt{n}} \int_0^\infty e^{-\frac{u^2}{2}} du = \sqrt{\frac{2\pi}{n}} \underbrace{\int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du}_{=\frac{1}{2}} = \sqrt{\frac{\pi}{2n}}\,,$$

其中，最后一个积分等于 $1/2$，因为被积函数是标准高斯密度。对于（b）部分，我们利用 μ 是概率测度这一事实（即 $\mu(\mathcal{Q}) = 1$）：

$$\|\delta_n\|_\mu = \int_{\mathcal{Q}} |\delta_n(q)| d\mu \le \sup_{q\in\mathcal{Q}} |\delta_n(q)| \int_{\mathcal{Q}} 1\, d\mu \le \|\delta_n\|_\infty \cdot \mu(\mathcal{Q}) = \|\delta_n\|_\infty\,,$$

因此，对于 $X_1, \ldots, X_n$ 的任意实现，均有 $\|\delta_n\|_\mu \le \|\delta_n\|_\infty$。应用期望的单调性：

$$\mathbb{E}_X\left[\|\delta_n\|_\mu\right] \le \mathbb{E}_X\left[\|\delta_n\|_\infty\right] \le \sqrt{\frac{\pi}{2n}}\,.$$

◀

引理 4 的证明。 对于（a）部分，我们使用引理 3，它表明 $\mathbb{E}_X\left[\|\delta_n\|_\mu\right] \le \sqrt{\frac{\pi}{2n}}$。随机变量的期望不可能严格小于每一个单独的取值。因此，至少存在一个 $\{X_i\}_{i=1}^n$ 的实现，使得 $\|\delta_n\|_\mu \le \sqrt{\frac{\pi}{2n}}$。对于（b）部分，我们使用 Csörgő 和 Faraway 的一个结果 [4]。将 Cramér–冯·米塞斯统计量定义为

$$\omega_n^2 = n\int (F(q) - F_n(q))^2 dF\,.$$

那么，对于任意 $t \in \left[\frac{1}{12n}, \frac{n+3}{12n^2}\right]$，都有（[4, 公式 2.4]）：

$$\mathbb{P}\left(\omega_n^2 \le t\right) = \frac{n!\pi^{n/2}}{\Gamma(n/2+1)} \left(t - \frac{1}{12n}\right)^{n/2} > 0\,. \tag{19}$$

现在，选取 $t = \frac{1}{12n} + \frac{n}{12n^2} = \frac{1}{6n}$。注意：

$$\omega_n^2 = n\int (F(q) - F_n(q))^2 dF = n\|\delta_n^2\|_\mu \ge n\|\delta_n\|_\mu^2\,,$$

其中最后一步依据詹森不等式。因此，我们可以写成：

$$\mathbb{P}\left(\|\delta_n\|_\mu \le \frac{1}{\sqrt{6}\,n}\right) = \mathbb{P}\left(n\|\delta_n\|_\mu^2 \le \frac{1}{6n}\right) \ge \mathbb{P}\left(\omega_n^2 \le \frac{1}{6n}\right) = \frac{n!\pi^{n/2}}{\Gamma(n/2+1)} \left(\frac{1}{12n}\right)^{n/2} > 0\,,$$

最后两个步骤应用了方程 (19)。由于 $\left\{\|\delta_n\|_\mu \le \frac{1}{\sqrt{6}\,n}\right\}$ 的概率大于 0，因此必然存在一个发生该情况的 $\{X_i\}_{i=1}^n$ 的实现。这就完成了证明。◀

E 查询时间下界：形式化结果与证明

现在，我们针对第 4.3 节所述的三种搜索策略，推导查询时间 $T(q)$ 的具体下界。为此，我们将命题 1 和命题 2 中的逼近理论结果与引理 3 和引理 4 中的偏差界结合起来。我们首先讨论线性搜索的情况，此时 $T(q) \geq \varepsilon(q)$。由此不等式可得：

▶ **推论 11** （线性搜索的下界）。 设 X_1、…、X_n 是来自具有累积分布函数 F 的分布的独立同分布样本，并设 $q \sim \mu$ 独立采样自 $\{X_i\}$。设 $I(A_n)$ 是类别 $I_{K,\mathcal{L}}$ 中的学习型索引，则在线性搜索下，期望查询时间满足：

$$\mathbb{E}_{X,q}\left[T(q)\right] \geq n\left[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\right].$$

此外，如果 $d\mu = dF$，则存在一个实现 X_1、…、X_n，使得

$$\mathbb{E}_{q}\left[T(q)\right] \geq n\left[R_{K,\mathcal{L}}(F) - \tfrac{1}{\sqrt{6}\,n}\right].$$

证明。 根据命题 5，我们知道 $T(q) \geq \varepsilon(q)$。由于 q 与 $\{X_i\}$ 独立：

$$\mathbb{E}_{X,q}\left[T(q)\right] \geq \mathbb{E}_{X,q}\left[\varepsilon(q)\right] = \mathbb{E}_{X}\left[\mathbb{E}_{q}\left[\varepsilon(q)\right]\right].$$

现在，通过命题 1，我们给出 $\mathbb{E}_q[\varepsilon(q)]$ 的下界：

$$\begin{aligned}\mathbb{E}_{X,q}\left[T(q)\right] &\geq \mathbb{E}_{X}\left[n\left[R_{K,\mathcal{L}}(F) - \|\delta_n\|_\mu\right]\right] \\ &= n\left[R_{K,\mathcal{L}}(F) - \mathbb{E}_{X}\left[\|\delta_n\|_\mu\right]\right] \\ &\geq n\left[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\right],\end{aligned}$$

最后一步来自引理 3(b)。这证明了推论的第一部分。第二部分直接由命题 1 和引理 4(b) 得出。◀

请注意，推论的第一部分提供了更强类型的保证，因为它对数据和查询分布两者的完整期望均成立（$\mathbb{E}_{X,\,q}[\cdot]$），并且适用于任意查询分布。相比之下，第二部分假设 $d\mu = dF$，且仅对数据样本提供最坏情况保证。然而，这一更强的假设带来了更快的收敛速率：偏差项按 $1/n$ 的量级缩放，而不是 $1/\sqrt{n}$。对于其他搜索算法，也存在类似的权衡。

~对于指数搜索，命题 5 给出 $T(q) \geq T(q) = \log_2(\max\{2,\ \varepsilon(q)\})$。为简化记号，我们将写作 $T(q) \geq \log_2 \varepsilon(q)$，其中假设 $\log_2(x)$ 对 $x \leq 2$ 输出 1。在这一情形下，我们可以推导出以下下界。

▶ **推论 12** (指数搜索的下界)。 设 $I(A_n)$ 是类别 $I_{K,\mathcal{L}}$中的一个学习型索引。那么，存在一个 $\{X_i\}_{i=1}^n$ 和 $q \in Q$ 的实现，使得

$$T(q) \geq \begin{cases} \log_2\left(n\left[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\right]\right) & \textit{for arbitrary } \mu\,, \\ \log_2\left(n\left[R_{K,\mathcal{L}}(F) - \tfrac{1}{\sqrt{6}\,n}\right]\right) & \textit{if } d\mu = dF\,. \end{cases}$$

此外，若满足命题 2 中的假设，则存在一个 $\{X_i\}_{i=1}^n$ 的实现，使得指数搜索下的期望查询时间满足：

$$\mathbb{E}_q\left[T(q)\right] \geq C_1\left[\log_2\left(nR_{K,\mathcal{L}}(F)\right) - C_2\right],$$

其中，C_1、$C_2 > 0$ 是与 n 和 K无关的常数。

证明。我们知道 $T(q) \geq T(q)$（命题 5）。因此，推论的第二部分可直接应用命题 2 得出。对于第一部分，首先考虑任意一个 μ。根据引理 4(a)，存在一个 $\{X_i\}_{i=1}^n$ 的实现，使得 $\|\delta_n\|_\mu \leq \sqrt{\frac{\pi}{2n}}$。同时，根据命题 1：

$$\mathbb{E}_q\left[\varepsilon(q)\right] \geq n\left[R_{K,\mathcal{L}}(F) - \|\delta_n\|_\mu\right] \geq n\left[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\right],$$

因此，至少存在一个 q，使得 $\varepsilon(q) \geq n\left[R_{K,\mathcal{L}}(F) - \sqrt{\frac{\pi}{2n}}\right]$。结合 $T(q) \geq \log_2 \varepsilon(q)$ 即可证明结论。$d\mu = dF$ 的情形可用同样的方法证明，并根据引理 4 的情形 (b) 说明，存在一个 $\{X_i\}_{i=1}^n$ 的实现，使得 $\|\delta_n\|_\mu \leq \frac{1}{\sqrt{6}\,n}$。◀

最后，对于二分搜索，我们考虑对数据和查询值 $q \in \mathcal{Q}$ 取最坏情况时的查询时间。根据命题 5，代价由最坏情况预测误差决定。

▶ **推论 13** (二分搜索的下界)。设 $I(A_n)$ 是类别 $I_{K,\mathcal{L}}$ 中的一个学习型索引。那么，存在 $\{X_i\}_{i=1}^n$ 的一个实现以及 $q \in \mathcal{Q}$，使得

$$\sup_q T(q) \geq \begin{cases} \log_2\left(n\left[R_{K,\mathcal{L}}(F) - \sqrt{\frac{\pi}{2n}}\right]\right) & \text{for arbitrary } \mu\,, \\ \log_2\left(n\left[R_{K,\mathcal{L}}(F) - \frac{1}{\sqrt{6}\,n}\right]\right) & \text{if } d\mu = dF\,. \end{cases}$$

证明。首先考虑任意 μ的情形。根据引理 4(a)，存在$\{X_i\}_{i=1}^n$ 的一种实现，使得 $\|\delta_n\|_\mu \leq \sqrt{\frac{\pi}{2n}}$。同时，根据命题 1，我们知道

$$\mathbb{E}_q\left[\varepsilon(q)\right] \geq n\left[R_{K,\mathcal{L}}(F) - \|\delta_n\|_\mu\right] \geq n\left[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\right].$$

因此，至少存在一个 q，使得 $\varepsilon(q) \geq n\left[R_{K,\mathcal{L}}(F) - \sqrt{\frac{\pi}{2n}}\right]$，这意味着

$$\sup_q \varepsilon(q) \geq n\left[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\right].$$

结合 $\sup_q T(q) \geq \log_2(\sup_q \varepsilon(q))$（命题 5）即可证明结论。情形$d\mu = dF$ 的证明方法相同，可利用引理 4 的情形 (b) 说明存在 $\{X_i\}_{i=1}^n$ 的一种实现，使得 $\|\delta_n\|_\mu \leq \frac{1}{\sqrt{6}\,n}$。◀

这些结果进一步阐明了第 4 节提出的通用框架，并根据 n（键的数量）以及 $R_{K,\mathcal{L}}(F)$（给定模型类所能达到的最佳近似误差），给出了具体的查询时间下界。

F 分段常数逼近：不匹配的数据集分布与查询分布

与第5.1节中考虑的数据集分布与查询分布相匹配的情况 $f = g$ 相比，$f \neq g$ 的情形相对更为复杂；不过在这里，将 F进行逼近的问题同样可以归约为最优量化。唯一的区别在于，量化随机变量不再服从均匀分布，并且我们的结果对于 $K \to \infty$是渐近成立的，而不像第5.1节那样适用于任意给定的 K。

本节中的结果假设 F 可逆。这是一个相对宽松的条件，对大多数标准连续分布示例都成立。此外，这一假设可以放宽：如果 F 仅在子区间 $[a, b] \subseteq \mathcal{Q}$上可逆，分析仍然

适用，只需将查询分布条件化到 $[a,b]$。这只会通过改变条件查询分布影响缩放常数，但会保留 K中的渐近行为。为使论述清晰，我们重点讨论 F 在整个$\mathcal{Q}$上均可逆的情况。

▶ **命题14**（归约到量化）。设 $h \in \mathcal{I}_{K,\ P_0}$ 和 $Y = F(Q)$ 满足 $Q \sim \mu$且 $d\mu = g\,dq$。那么

$$\|F-h\|_{\mu,r} \overset{\text{def}}{=\!=} \int_{\mathcal{Q}} |F(q)-h(q)|^r g(q)dq = \mathbb{E}[|Y-Q_K(Y)|^r]\,.$$

证明。 与式（4）中一样，我们将$\|F-h\|_{\mu,r}$ 展开为

$$\|F-h\|_{\mu,r} = \sum_{k=1}^{K} \int_{I_k} |F(q)-c_k|^r g(q)dq\,.$$

引入 $y = F(q)$ $dy = f(q)dq,$ $q = F^{-1}(y),$ 以及 $dq = 1/f(q)\ dy,$ 我们将其改写为

$$\sum_{k=1}^{K} \int_{J_k} |y-c_k|^r \cdot \frac{g\left(F^{-1}(y)\right)}{f\left(F^{-1}(y)\right)} dy = \mathbb{E}[|Y-Q_K(Y)|^r]\,,$$

最后一个等号成立，是因为我们注意到，积分中的第二项是当 Q 服从概率密度函数 g 时，$Y = F(Q)$ 的概率密度函数（PDF）：

$$f_Y(y) = g\left(F^{-1}(y)\right) \cdot \left|\frac{d}{dy}F^{-1}(y)\right| = g\left(F^{-1}(y)\right) \cdot \left|\frac{1}{F'\left(F^{-1}(y)\right)}\right| = \frac{g\left(F^{-1}(y)\right)}{f\left(F^{-1}(y)\right)}\,; \tag{20}$$

第一个等式展开了随机变量函数的密度，第二个等式则展开了反函数的导数。需要注意的是，这要求累积分布函数 F 可逆。◀

命题14意味着，对 F 的近似可归结为对一个服从式 (20) 所示分布、定义在 $[0,1]$上的随机变量$Y = F(Q)$进行量化。当 $f \neq g$时，该分布不再是均匀分布，因此均匀量化也不再是最优的。继续借用图5中的直观类比，可以想象，在 F 下方的绿色虚线密度不再与 f成正比，因此 F 上方的虚线也不再均匀；在这种情况下，近似值$\{c_k\}$就必须相应地移动。

现在我们可以借助量化理论中的一个重要结果。已知渐近量化误差满足 [13, 定理 6。2]

$$\lim_{K\to\infty} K^r\ T_{K,r}(Y) = \frac{1}{(1+r)2^r}\left(\int f_Y(y)^{1/(1+r)}dy\right)^{1+r}\,.$$

等式右侧称为 r阶量化系数，是概率分布的一项基本性质，类似于其各阶矩；量化系数在量化理论中发挥着重要作用。因此，对于绝对误差

$$\lim_{K\to\infty} K\ R_{K,P_0}(F) = \lim_{K\to\infty} K\ T_{K,1}(Y) = \frac{1}{4}\left(\int \sqrt{f_Y(y)}dy\right)^2\,.$$

换句话说，渐近地，绝对误差随 K 线性减小，就像第5.1节中考察的匹配分布情形一样，从而得到 $R_{K,\ \mathcal{L}}(F) = \Omega\left(1/K\right)$，其中 $\mathcal{L} = P_0$。

G 第6节中的证明

引理6的证明。 $R_{K,\mathcal{L}}(F)$ 可写为

$$R_{K,\mathcal{L}}(F) = \inf_{h\in\mathcal{I}_{K,\mathcal{L}}} \|F-h\|_\mu = \inf_{\mathcal{P}\in\mathbf{P}_K(\mathcal{Q})} \inf_{h\in\mathcal{I}_{\mathcal{P},\mathcal{L}}} \|F-h\|_\mu ,$$

其中，下确界 $\inf_{\mathcal{P}\in\mathbf{P}_K(\mathcal{Q})}$ 取遍所有将 $\mathcal{Q}$ 划分为 K 个区间的划分 $\mathbf{P}_K(\mathcal{Q})$。根据下确界的定义，对于任意 $t>0$，存在 $\mathcal{P}_t \in \mathbf{P}_K(\mathcal{Q})$，使得

$$R_{K,\mathcal{L}}(F) \geq \inf_{h\in\mathcal{I}_{\mathcal{P}_t,\mathcal{L}}} \|F-h\|_\mu - t .$$

对 F 取上确界：

$$\sup_{F\in\mathcal{F}} R_{K,\mathcal{L}}(F) \geq \sup_{F\in\mathcal{F}} \inf_{h\in\mathcal{I}_{\mathcal{P}_t,\mathcal{L}}} \|F-h\|_\mu - t .$$

现在，$\mathcal{I}_{\mathcal{P}_t,\mathcal{L}}$ 是一个维数至多为 $\kappa = K\cdot\dim(\mathcal{L})$ 的线性空间。因此，可以通过对所有此类空间取下确界，得到最后一个表达式的下界：

$$\sup_{F\in\mathcal{F}} R_{K,\mathcal{L}}(F) \geq \inf_{H\in\mathcal{H}_\kappa} \sup_{F\in\mathcal{F}} \inf_{h\in H} \|F-h\|_\mu - t = d_\kappa(\mathcal{F}) - t .$$

请注意，现在下确界-上确界-下确界这一序列不再依赖于 t。因此，由于该不等式对所有 $t>0$ 都成立，根据实数的标准性质，它对 $t=0$ 也成立。◀

G.1 分段线性逼近的上界

我们简要概述在两种常见的目标函数 F正则性假设下，$R_{K,\mathcal{L}}(F)$（其中$\mathcal{L}=P_1$为一次多项式类）的近似误差上界。全文中，我们假设定义域为 $[0,1],$，测度 μ为勒贝格测度，且区间被划分为 K 个等长子区间 $[a_i,a_{i+1}]$，其中 $a_i=\frac{i-1}{K}$ 且 $a_{i+1}=\frac{i}{K}$。我们用 $\ell=1/K$表示每个子区间的长度。

情况1：F 是二阶连续可微的。我们使用基于一阶泰勒展开的线性函数，在每个区间 $[a_i,a_{i+1}]$ 上逼近 F：

$$L_i(x) = F(a_i) + (x-a_i)F'(a_i) .$$

由于 F 是 $C^2([0,1])$，我们可以利用泰勒定理对逐点误差进行界定：

$$|F(x)-L_i(x)| \leq \frac{1}{2}M\cdot(x-a_i)^2 ,$$

其中 $M=\max_{x\in[0,1]}|F''(x)|$。我们知道 M存在且有界，因为 F''连续，且 $[0,1]$ 是一个紧区间。对每个区间上的函数积分可得：

$$\int_{a_i}^{a_{i+1}} |F(x)-L_i(x)|dx \leq \frac{1}{2}M\int_{a_i}^{a_{i+1}} (x-a_i)^2dx = \frac{1}{6}M\ell^3 .$$

最后，对所有 K 区间求和可得：

$$\|F-h\|_\mu \leq K\cdot\frac{1}{6}M\ell^3 = O(1/K^2) .$$

情况 2： F **是** M**-Lipschitz**。现在，我们用端点之间的线性插值来近似 F：

$$L_i(x) = F(a_i) + \frac{F(a_{i+1}) - F(a_i)}{a_{i+1} - a_i} \cdot (x - a_i)\,.$$

根据利普希茨性质，对于任意 $x \in [a_i, a_{i+1}]$，我们有：

$$\begin{aligned}
|F(x) - L_i(x)| &= \left|F(x) - F(a_i) - \frac{F(a_{i+1}) - F(a_i)}{a_{i+1} - a_i}(x - a_i)\right| \\
&= \left|\frac{F(x) - F(a_i)}{x - a_i} - \frac{F(a_{i+1}) - F(a_i)}{a_{i+1} - a_i}\right| \cdot |x - a_i| \\
&\leq \left(\left|\frac{F(x) - F(a_i)}{x - a_i}\right| + \left|\frac{F(a_{i+1}) - F(a_i)}{a_{i+1} - a_i}\right|\right) \cdot |x - a_i| \\
&\leq (M + M)|x - a_i|\,.
\end{aligned}$$

因此，每个区间上的绝对误差至多为 $2M|x - a_i|$。对每个区间积分：

$$\int_{a_i}^{a_{i+1}} |F(x) - L_i(x)|dx \leq 2M \int_{a_i}^{a_{i+1}} |x - a_i|dx = M\ell^2\,.$$

最后，对所有 K 区间求和可得：

$$\|F - h\|_\mu \leq K \cdot M\ell^2 = O(1/K)\,.$$

这些关于 $R_{K,\ P_1}(F)$ 的渐近速率——即，在利普希茨连续性条件下为 $O(1/K)$，在 C^2 假设下为$O(1/K^2)$——更普遍地适用于定义在任意紧区间上的函数，以及任何具有有界密度的测度 μ。具体常数可能取决于定义域和 μ 的密度，但收敛速率不受影响。因此，分段线性函数在特定正则性假设下可以提供不同的逼近保证。

G.2 相对于 K 构造对抗性累积分布函数

我们将以 $\mathcal{Q} = [0, 1]$ 的情形为例，其中 μ 为均匀分布。这样我们便能进行一些精确计算；不过，本节推导出的渐近性质以及用于证明这些性质的论证，适用于非常一般的 $\mathcal{Q}$ 和 μ 选择。考虑函数 $F(x) = x^2$，以及用单个线性函数进行逼近时的最优近似误差 $R_{1,\ P_1}(F)$。注意，x^2 在 $[0, 1]$ 上是一个有效累积分布函数。根据 [26, 推论 3.4.1], ，很容易推导出

$$R_{1,P_1}(F) = 1/16\,, \tag{21}$$

其由 $h(x) = x - 3/16$ 达到。现在，取一个正整数 M。直观而言，我们可以构造一个阶梯型函数 F_M，共有 M 个阶梯，每个阶梯 F_M^i 都等于 $F(x) = x^2$ 经适当平移和缩放后的结果，使得整个 F_M 构成一个难以近似的有效累积分布函数。对于任意 $i = 0, \ldots\ldots, M - 1$，定义 F_M^i 为：

$$F_M^i(x) = \frac{1}{M}\left[M\left(x - \frac{i}{M}\right)\right]^2 + \frac{i}{M} \quad \text{for} \quad x \in I_i = \left[\frac{i}{M}, \frac{i+1}{M}\right].$$

现在，整个 F_M 定义为

$$F_M(x) = \sum_{i=0}^{M-1} F_M^i(x)\mathbf{1}_{I_i}(x) \quad \text{for} \quad x \in [0,1]\,.$$

容易验证， F_M 连续且递增，并满足 $F(0) = 0$、 $F(1) = 1$，因此它是定义在 $[0,1]$上的有效累积分布函数。现在考虑函数F_M^i 在区间 I_i上的任意线性近似 $L_i \in P_1$ 。近似误差为:

$$\begin{aligned} E_i &= \int_{\frac{i}{M}}^{\frac{i+1}{M}} |F_M^i(x) - L_i(x)|dx \\ &= \int_{\frac{i}{M}}^{\frac{i+1}{M}} \left| \frac{1}{M}\left[M\left(x - \frac{i}{M}\right)\right]^2 + \frac{i}{M} - L_i(x) \right| dx\,. \end{aligned}$$

代入 $u = M(x - i/M)$ $du = Mdx$:

$$\begin{aligned} &= \frac{1}{M}\int_0^1 \left| \frac{u^2}{M} + \frac{i}{M} - L_i\left(\frac{u+i}{M}\right) \right| du \\ &= \frac{1}{M^2}\int_0^1 \left| u^2 + i - ML_i\left(\frac{u+i}{M}\right) \right| du\,. \end{aligned}$$

现在我们来关注积分项。由于 $i - ML_i((u+1)/M) \in P_1$，它在逼近 u^2 时不可能比最优逼近函数 $h(u) = u - 3/16$产生更小的误差，因此由公式 (21) 可得:

$$E_i \geq \frac{1}{16M^2}\,. \tag{22}$$

现在我们用它来推导 $R_{K的P_1}(F_M)$的下界。由于 F_M 由 M 个互不相交的二次阶梯组成，每个阶梯对应一个长度为 $1/M$的区间，因此一个 K分段线性函数在所有 M 条子抛物线上最多只能分配 K 个线性片段。因此，至少有 $M - K + 1$条子抛物线必须各自仅用一个线性片段覆盖，而在每个对应区间上，界 (22) 均适用。对这 $M - K + 1$ 个区间求和，得到

$$R_{K,P_1}(F_M) \geq \frac{M - K + 1}{16M^2}\,.$$

当 $M = 2(K-1)$ 时，此表达式达到最大值，从而得到

$$R_{K,P_1}(F_{2(K-1)}) \geq \frac{1}{64(K-1)} = \Omega\left(\frac{1}{K}\right).$$

这一构造表明，即使 F 是有效累积分布函数，使用 K 段分段线性函数进行逼近时，仍可能存在 $\Omega(1/K)$ 的下界。具体而言，$R_{K,\ P_1}(F)$ 的最坏情况表现并不优于分段常数模型，尽管线性片段具有更强的表达能力。这凸显了对 F 作出正则性假设的关键作用：若缺乏足够的光滑性，更具表达能力的模型未必能带来更高的近似速率。

G.3 Lipschitz累积分布函数的柯尔莫哥洛夫宽度

不失一般性，我们考虑 $[a,b] = [0,1]$，并将注意力限制在 Lipschitz常数为 1 的 Lipschitz 连续函数类别 $\mathcal{F}$上，即:

$$\mathcal{F} = \left\{F \in L^1(\mu) : |F(x) - F(y)| \leq |x - y|\right\}.$$

令 $d_\kappa(\mathcal{F})$ 表示$\mathcal{F}$ 在 $L^1(\mu)$中的 Kolmogorov 宽度，则：

$$d_\kappa(\mathcal{F}) = \inf_{H\in\mathcal{H}_\kappa} \sup_{F\in\mathcal{F}} \inf_{h\in H} \|F-h\|_\mu = \Omega(1/\kappa)\,, \tag{23}$$

其中 $d_\kappa(\mathcal{F}) = \Omega(1/\kappa)$ 来自第6.2节中引用的经典结果。然而，$\mathcal{F}$ 中的并非所有函数都是有效累积分布函数（累积分布函数），因为它们不一定非负或单调递增。现在我们证明，对于任意 $F\in\mathcal{F}$，都可以构造出一个有效累积分布函数 $\widehat{F}$，并且其近似难度在常数因子意义下保持不变。

为简化记号，将 F 到集合 H 的距离记为 $\|F-H\|_\mu$，即$\|F-H\|_\mu = \inf_{h\in H}\|F-h\|_\mu$。注意，方程（23）中的最外层下确界是在所有维数至多为 κ 的线性子空间 $H\subseteq L^1(\mu)$ 上取得的。因此，该界适用于任何此类空间。于是，考虑任意 H^*，其至少包含所有常数函数和线性函数（只要 $\kappa\geq 2$，总可以做到这一点，因为 $\mathrm{span}\{1,x\}$ 就是这样一个空间的例子）。现在，固定 $t>0$。根据方程（23）和上确界的定义，存在 F 使得

$$\|F-H^*\|_\mu \geq \frac{C}{\kappa} - t\,,$$

其中，$C>0$ 是一个与 t无关的正常数。

步骤 1：构造一个严格递增函数。众所周知，$\mathcal{F}$ 与弱可微函数的 Sobolev 球 $W^{1,\infty}([0,1])$ 重合，这些函数的弱导数几乎处处有界于 1 [7, 第 5.8 章，定理 4]。这意味着$F'\geq -1$ 除一个零测集外。现在，定义

$$\widetilde{F}(x) = F(x) + 2x\,.$$

于是，$\widetilde{F}$在 $[0,1]$, 上单调，因为它的弱导数 F' 为正：

$$\widetilde{F}'(x) = F'(x) + 2 \geq -1 + 2 = 1\,.$$

因此，$\widetilde{F}$ 严格递增；并且由于 F 和 $2x$ 都是 Lipschitz 连续的，它也是连续的。

步骤 2：缩放为有效累积分布函数。现在，我们对 $\widetilde{F}$ 进行归一化，得到一个满足以及 $\widehat{F}(0)=0$和 $\widehat{F}(1)=1$的函数 $\widehat{F}$：

$$\widehat{F}(x) = \frac{\widetilde{F}(x)-\widetilde{F}(0)}{\widetilde{F}(1)-\widetilde{F}(0)} = \frac{F(x)+2x-F(0)}{F(1)+2-F(0)}\,.$$

$\widehat{F}$ 在 $[0,1]$, 上严格递增且连续，因为 $\widetilde{F}$ 如此。因此，它是一个有效累积分布函数。此外，我们可以给缩放常数确定上界和下界：

$$(\widetilde{F}(1)-\widetilde{F}(0)) = (F(1)+2-F(0)) \in [1,3]\,, \tag{24}$$

因为根据 1-Lipschitz 性质，$(F(1)-F(0))\in[-1,1]$。

步骤 3：保留逼近难度。回想一下

$$\|F-H^*\|_\mu \geq \frac{C}{\kappa} - t\,.$$

23:34 学习索引的算法复杂度下界

现在，记 $M = F(1) + 2 - F(0)$，并考虑

$$\|\widehat{F} - H^*\|_\mu = \left\| \frac{F + 2x - F(0)}{M} - H^* \right\|_\mu = \frac{1}{M} \|F + 2x - F(0) - M \cdot H^*\|_\mu .$$

由于 H^* 是一个线性子空间，因此它在标量乘法下保持不变，所以当 $M \neq 0$ 时，$M \cdot H^* = H^*$。此外，由于我们假设 H^* 张成常数函数和线性函数构成的集合，因此 $2x - F(0) - M \cdot H^* = H^*$ 也成立。根据等式 (24)，我们知道 $M \leq 3$，综上可得：

$$\|\widehat{F} - H^*\|_\mu = \frac{1}{M} \|F - H^*\|_\mu \geq \frac{1}{3} \|F - H^*\|_\mu \geq \frac{C}{3\kappa} - \frac{t}{3} . \tag{25}$$

设 $\widehat{\mathcal{F}} = \{F$ 是按照步骤 1 和 2 中的描述，由 $F \in \mathcal{F}$ 构造的$\}$。然后，在等式 (25) 中对该类别取上确界，可得：

$$\sup_{\widehat{F} \in \widehat{\mathcal{F}}} \|\widehat{F} - H^*\|_\mu \geq \frac{C}{3\kappa} - \frac{t}{3} ,$$

其中 C 不依赖于 t。由于这对任意 $t > 0$ 都成立，因此对 $t = 0$ 也成立。现在，对所有维数为 κ 的线性空间取下确界，可得：

$$d_\kappa(\widehat{\mathcal{F}}) \geq \frac{C}{3\kappa} = \Omega\left(\frac{1}{\kappa}\right) .$$

这一构造表明，Lipschitz 连续函数逼近的任何下界，都可以扩展为有效累积分布函数逼近的下界，二者至多相差一个常数乘法因子。因此，即使将范围特别限定为累积分布函数，第 6.2 节中的所有结论仍然成立。