Transformer作为测度论关联记忆：统计学视角与极小极大最优性

Ryotaro Kawata[1, 2, *], **Taiji Suzuki**[1, 2, §]

[1]东京大学数学信息学系，日本[2]理化学研究所先进智能项目中心，日本

[*]kawata-ryotaro725@g.ecc.u-tokyo.ac.jp[§]
taiji@mist.u-tokyo.ac.jp

arXiv:2602.01863v1 [stat.ML] 2 Feb 2026

摘要

Transformer 擅长通过内容寻址检索，并能利用原则上长度无界的上下文。我们在概率测度层面重新表述联想记忆，将上下文视为 token 上的分布，并将注意力视为作用于测度的积分算子。具体而言，对于混合上下文 $\nu = I^{-1}\sum_{i=1}^{I}\mu^{(i^*)}$ 和查询 $x_{\mathrm{q}}(i^*)$，任务可分解为：(i) 召回相关分量 $\mu^{(i^*)}$；(ii) 基于 $(\mu_{i^*}, x_{\mathrm{q}})$ 进行预测。我们研究通过经验风险最小化训练的学习型 softmax 注意力（而非固定核），并证明，在输入密度满足谱假设的条件下，由多层感知机与浅层测度论Transformer组成的模型能够学习召回与预测映射。我们还建立了速率指数相同的匹配极小极大下界（相差至多一个乘法常数），从而证明该收敛阶是尖锐的。该框架为设计和分析能够从任意长的分布式上下文中进行召回、且具有可证明泛化保证的 Transformer，提供了一套有原则的方法。

1 引言

Transformer（Vaswani 等人，2017）在自然语言（Brown 等人，2020）、视觉（Dosovitskiy 等人，2021）和语音/音频（Dong 等人，2018）领域均取得了出色的实证表现。以下两个特性促使我们开展这项研究：(i) 对关联信息的内容寻址检索 ——一种注意力机制的联想记忆视角——以及 (ii) 利用长度可变且原则上无界的上下文的能力。

在本文中，我们在<样式 id='1'>概率测度层面刻画联想记忆，将上下文视为词元上的分布，并在这一测度论框架下，对学习得到的 softmax 注意力 Transformer 展开严格的统计分析。

联想记忆为理解神经系统如何根据部分线索进行存储和检索提供了统一视角：从早期的自组织记忆和相关记忆，到 Hopfield吸引子（Amari，1972；Kohonen，1972；Nakano，1972；Hopfield，1982；1984）。Transformer 通过内容寻址注意力重新诠释了联想记忆或召回，在形式上等价于 Hopfield式联想更新（Vaswani 等人，2017；Ramsauer 等人，2021）。近期研究量化了记忆涌现及其容量（Bietti 等人，2023；Cabannes 等人，2024；Mahdavi 等人，2024；Kim 等人，2023；Jiang 等人，2024；Nichani 等人，2025）。

由于 Transformer 的设计目标是处理海量文本语料和长上下文，研究人员将这种“上下文”形式化为词元上的概率测度，从而以测度论的方式处理可变大小的输入。根据大数定律，将文本数据概括为一个测度，有助于他们证明结果与文本长度无关。Transformer 的测度论视角将注意力形式化为分布上的映射，从而能够分析稳定性和涌现结构（Vuckovic 等，2020；Sander 等，2022；Geshkovski 等，2025；Burger 等，2025）。在表达能力方面，Transformer 可以在输入/输出测度之间进行插值（Geshkovski

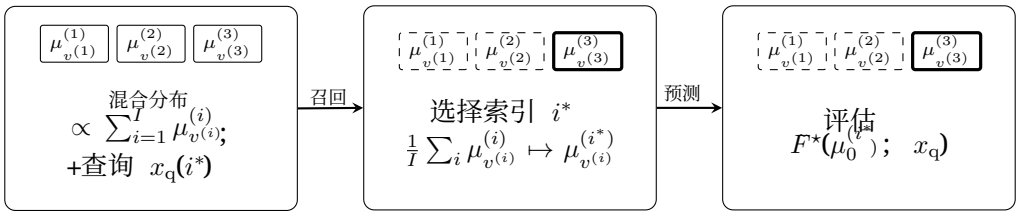

图1：测度层面的联想回忆（非正式式）：查询 $x_q(i^*)$ 从混合分布 $\nu \propto \sum_{i=1}^{I} \mu_{v^{(i)}}^{(i)}$ 中选出相关的分量测度 $\mu^{(i^*)}{}_{v^{(i^*)}}$，随后根据 $(\mu_0^{(i^*)}, x_q)$ 进行预测。注意，每个 $\mu_{v^{(i)}}^{(i)}$ 均由 $v^{(i)} \in \mathbb{S}^{d_1}$ 和定义在 $\mathbb{R}^{d_2}$ 上的测度 $\mu_0^{(i)}$ 构造而成。

等人 等，2024），甚至能够一致逼近连续的上下文映射，其中上下文 i s 本身一个概率分布（Furuya 等人，2025）。

近期研究对具有无限维输入的 Transformer 开展了统计分析。然而，与联想记忆的联系——可以说这是注意力的一项标志性特征——仍未得到充分阐明。以往关于分布回归的泛化结果通常假设注意力核是冻结的（不可学习的）（Liu 和 Zhou，2025），因此尚不清楚学习得到的注意力如何召回相关测度。同样，对于具有无限维输入的序列式上下文学习设定（Kim 等，2024），相关分析是在使用线性注意力的条件下进行的；线性注意力的表达能力有限，难以实现 softmax 注意力所能达到的尖锐、集中的权重分布（Han 等，2024；Fan 等，2025）。这些考量引出了核心问题：

问： 学习得到的 *softmax* 注意力 *Transformer* 能否召回无限维（取值为测度的）上下文，并据此进行预测，同时具备可证明的泛化保证？

测度层面的"联想记忆"（非正式表述）。考虑一个由 I 个文档组成的文本语料库。我们将每个词元建模为一个向量 $x = (v,\ z) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$，其中 v 编码文档级特征（例如主题），z 编码词元级内容。对于文档 i，文档特征是一个固定向量 $v^{(i)} \in \mathbb{S}^{d_1}$，而内容部分 z 则从某个定义在 $\mathbb{R}^{d_2}$ 上的分布$\mu_0^{(i)}$ 中采样。当文档长度趋于无穷时，文档 i 的经验词元分布收敛到定义在 $\mathbb{R}^{d_1+d_2}$ 上的概率测度 $\mu_{v^{(i)}}$，即 $x = (v^{(i)},\ Z)$ 的分布，其中$Z \sim \mu_0^{(i)}$。于是，模型所见的上下文便是如下混合分布

$$\nu \ = \ \sum_{i=1}^{I} w_i \, \mu_{v^{(i)}}^{(i)}, \qquad w_i \geq 0, \ \sum_i w_i = 1,$$

表示一个包含许多文档的完整数据集。给定这样的混合分布 ν 和一个查询 $x_q \in \mathbb{R}^{d_1+d_2}$，其前 d_1 个坐标与某个文档特征 $v^{(i^\star)}$ 对齐，所期望的"联想记忆"行为是：

- 首先，从混合分布 ν 中召回 由 $i^\star$ 索引的分量，然后
- 然后预测 一个仅取决于相关内容分布 $\mu_{(i^\star)\,0}$（也可能取决于x_q）的标量值。

我们将 $F^\star$：$(\nu,\ x_q) \mapsto F^\star(\nu,\ x_q) \in \mathbb{R}$ 称为这个真实值回忆与预测映射：根据构造，其输出仅通过查询所选取的单一分量 $\nu\ \mu_0^{(i^\star)}$ 依赖于查询（图 1）。我们研究通过经验风险最小化（ERM）训练的学习型 softmax 注意力 Transformer（采用注意力的积分／经验测度视角），以实现这一回忆与预测流程。在统计方面，我们关注一个非常光滑的情形：我们赋予上下文测度空间一个再生核希尔伯特空间（RKHS），其 Mercer 特征值满足 $\lambda_j \asymp \exp(-c\, j^{\alpha})$，其中某个 $\alpha > 0$（高斯型核即属此类（Sch¨olkopf & Smola，2002））。这种谱衰减体现了底层密度的强光滑性，并导出一个有效维数，它将决定我们的学习率。

贡献。现在我们概述本文的主要贡献：

1. **测度层面的联想记忆**。我们为测度上的联想回忆建立了一个一般且数学上严谨的框架：给定一个测度值上下文和一个查询，回忆算子会选出相关联的测度，而预测器则将回忆出的测度与查询一同映射为输出。我们将针对查询从任意长的上下文中进行选择这一过程形式化：模型回忆出捕捉相关内容的概率测度，并根据其统计量进行预测。

2. **泛化**。我们证明，由多层感知机组成的浅层（深度为 2）测度论Transformer能够在测度层面学习回忆与预测的映射（定理 1）。与线性注意力（Kim 等，2024）或冻结核（Zhou等，2024）不同，softmax 注意力能够对相关测度进行稀疏且自适应的回忆。对于有界参数假设类上的经验风险最小化——只要回忆候选项的数量不过大——我们建立了次多项式总体风险上界 $\exp\{-\Theta((\log n)^{\alpha/(\alpha+1)})\}$，表明统计难度由核的 Mercereigen 衰减 α 所决定。

3.**极小极大最优性**。 我们证明了一个具有相同速率指数（$\log n$）$^{\alpha/(\alpha+1)}$ 的极小极大下界，从而确立了该收敛阶的紧确性（定理 2）。因此，在我们的谱假设和混合增长假设下，所提出的测度论Transformer在指数阶上是极小极大最优的，尽管乘法常数可能有所不同。

本文其余部分安排如下。第 2 节回顾相关工作。第 3 节介绍问题设定和学生模型。第 4 节呈现我们的主要理论结果。第 5 节通过讨论作结。技术细节和证明见附录。

2 相关工作

联想记忆与召回。 联想记忆的概念起源于早期的神经科学模型（Hopfield，1982；Amari，1972；Kohonen，1972；Nakano，1972；Hopfield，1982；1984），之后又有 Graves 等人（2014）；Weston 等人（2014）；Ramsauer 等人（2021）；Millidge 等人（2022）的研究。Transformer架构通过采用自注意力作为内容寻址机制，与联想记忆密切相关（Vaswani 等人，2017）。近期研究日益关注联想记忆如何在Transformer架构中涌现并扩展（Bietti 等人，2023；Cabannes 等人，2024；Mahdavi 等人，2024；Kim 等人，2023；Jiang 等人，2024；Nichani 等人，2025）。

面向无限维输入的 Transformer。 测度论视角使对 Transformer架构的深入分析成为可能。Vuckovic 等人（2020）；Sander 等人（2022）将自注意力形式化为概率测度上的映射。其 Lipschitz 连续性在 Castin 等人（2024）中得到探讨。在该框架基础上，Geshkovski 等人（2023；2025）；Burger 等人（2025）将自注意力建模为相互作用粒子系统。Geshkovski 等人（2024）证明了一项普适性结果，表明 Transformer 可以插值任意输入—输出测度对，随后由 Furuya 等人（2025）加强为对分布与查询上连续映射的一致逼近。在泛化方面，Liu 与 Zhou（2025）研究了分布回归，但仅限于冻结的注意力核。在序列型无限维输入的上下文中，Kim 等人（2024）研究了基于线性注意力的上下文学习，其本质上归结为平均行为；因此他们的分析假设召回候选项具有宽松稀疏性和正交归一性，这反映了与 softmax 注意力相比，实现尖峰式独热回忆的困难（Han 等，2024；Fan 等，2025）。

函数型映射的多层感知机逼近。 在统计学习中，Mhaskar 与 Hahm（1997）奠定了基础，他们证明多层感知机（MLP）能够以最优速率逼近函数空间上的连续非线性泛函；Stinchcombe（1999）将这一结果推广了。Rossi 等人（2005）提出了一种适用于函数型数据的新型函数型多层感知机，之后又出现了多种变体（Yao 等人，2021；Song 等人，2023；Zhou 等，2024）。在优化方面，Suzuki（2020）和 Nishikawa 等（2022）为在无限维情形下运行的两层网络建立了全局优化保证。

3 问题设定

记号。 对于整数 $N_1 \le N_2$和 $\boldsymbol{v} \in \mathbb{R}^N$，我们记 $\boldsymbol{v}_{N_1:N_2} = (v_{N_1}, \dots, v_{N_2})^\top$。对于矩阵 A，$\|A\|_0$ 表示非零元素的个数，$\|A\|_\infty = \max_{i,j} |A_{i,j}|$。我们用 λ 表示 $\mathbb{R}^N$上的勒贝格测度，用 $f_\sharp \mu$表示 μ在 f下的推前测度。对于可测空间 $\mathcal{X}$，$\mathcal{P}(\mathcal{X})$表示概率测度的集合，$\mathcal{M}_+(\mathcal{X})$表示非负测度的集合。我们用 $(\Omega\ \mathcal{F}, \mathbb{P})$表示一个概率空间，$\|f\|_{L^p}$和 $\|f\|_\infty$分别表示通常的范数和本质上确界范数，并用 $\mathbb{P}_X$ 表示随机变量 X的分布。期望记作 $\mathbb{E}[\cdot]$ 或 $\mathbb{E}_X[\cdot]$，其中分布为 X。

Ou回归问题。 现在，我们将引言中的非正式回忆并预测场景形式化 e intr。

定义1。 设 $\mathcal{X}_0 \subset \mathbb{R}^{d_2}$ 为有界词元内容空间，令 $\mathcal{X} \subset \mathbb{R}^{d_1+d_2}$ 表示词元空间，其分解为 $x = (v, z)$，其中 $v \in \mathbb{R}^{d_1}$ 编码文档级特征，$z \in \mathcal{X}_0$ 编码词元级内容。

1. 混合上下文与查询。 每个文档 $i \in [I]$ 都与一个文档特征$v^{(i)} \in \mathbb{S}^{d_1-1}$和一个内容分布 $\mu_0^{(i)} \in \mathcal{M}_+(\mathcal{X}_0)$相关联。非正式地说，$\mu_0^{(i)}$ 表示文档 i中出现的词元内容（例如，词语或嵌入）的分布。相应的 $\mathcal{X}$上的词元分布是乘积测度

$$\mu_{v^{(i)}}^{(i)} \coloneqq \delta_{v^{(i)}} \otimes \mu_0^{(i)} = (\mathrm{Emb}_{v^{(i)}})_\sharp \mu_0^{(i)},$$

其中 $\mathrm{Emb}_{v^{(i)}}(z) \coloneqq (v^{(i)}, z)$，且 $f_\sharp \mu$表示 μ经 f的推前测度，从而 $\mu_{v^{(i)}}^{(i)}$是文档特征 $v^{(i)}$与文档 i中词元内容的联合分布。一个上下文 是这些分量测度的混合分布，

$$\nu \coloneqq \frac{1}{I} \sum_{i=1}^{I} \mu_{v^{(i)}}^{(i)} \in \mathcal{P}(\mathcal{X}), \tag{1}$$

它表示包含 I 个文档的整个数据集的词元分布。具体而言，ν是先随机采样一个文档 $i \in [I]$，再从该文档中采样一个词元 $x = (v^{(i)}, z)$ 所得到的分布。给定这样的混合分布，查询 $x_{\mathrm{q}} \in \mathcal{X}$ 被构造为指示一个特定的 索引 $i^\star \in [I]$。为具体起见，我们取

$$x_{\mathrm{q}} \coloneqq \begin{bmatrix} v^{(i^\star)} \\ 0_{d_2} \end{bmatrix} = \mathrm{Emb}_{v^{(i^\star)}}(0_{d_2}) \in \mathbb{R}^{d_1+d_2},$$

也就是说，文档特征 $v^{(i^\star)}$ 在最后 d_2 个坐标上用零填充。[1]

2. 真实值回忆与预测映射。 该学习任务是预测一个实值响应

$$y = F^\star(\nu, x_{\mathrm{q}}) + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2),$$

根据配对 (ν, x_{q})。关键结构假设是，$F^\star$对上下文 ν的依赖仅通过查询所选定的索引 $i^\star$所关联的单个分量实现。等价地，˜存在一个（隐藏的）泛函 $F^\star$，使得

$$F^\star(\nu, x_{\mathrm{q}}) = \tilde{F}^\star\big(\mu_0^{(i^\star)}, x_{\mathrm{q}}\big), \qquad \nu = \tfrac{1}{I}\textstyle\sum_{i=1}^{I} \mu_{v^{(i)}}^{(i)},$$

从而回归映射 $(\nu, x_{\mathrm{q}}) \mapsto F^\star(\nu, x_{\mathrm{q}})$ 分解为两个概念阶段

$$\big(\nu, x_{\mathrm{q}}\big) \xrightarrow{\text{recall } i^\star} \mu_0^{(i^\star)} \xrightarrow{\text{predict}} \tilde{F}^\star(\mu_0^{(i^\star)}, x_{\mathrm{q}}).$$

例如，$F^\star(\nu, x_{\mathrm{q}})$ 可以是文档 $i^\star$ 的情感分数，或文档 $i^\star$ 提及查询所指定实体的概率。

特别地，$F^\star$ 必须先将查询与混合上下文中相关的分量关联起来，然后再根据召回的分量预测一个标量（例如，上下文学习（Brown 等人，2020））。这将概率测度层面的“联想记忆”任务形式化。

[1]更一般的查询，例如带有非零内容部分的查询，同样可以处理；为记号简洁性起见，我们在此固定零填充。

统计估计问题。受上述回忆与预测回归任务的启发，我们现在阐述与之相关的统计估计问题。我们观测到 n 个独立同分布样本

$$\mathcal{S}_n \coloneqq \{(\nu_t, x_{\mathrm{q}_t}, y_t)\}_{t=1}^n$$

从 $(\nu,\ x_{\mathrm{q}},\ y)$ 的联合分布中抽取。令 $\mathcal{F}$ 表示由可测函数 $F\colon \mathcal{P}(\mathcal{X}_0) \times \mathbb{R}^{d_1+d_2} \to \mathbb{R}$ 组成的假设类。给定训练数据后，我们定义经验风险最小化器

$$\hat{F} \ :=\ \arg\min_{F\in\mathcal{F}} \hat{\mathbb{E}}_n\big[\big(y - F(\nu, x_{\mathrm{q}})\big)^2\big],$$

^^ 其中 $\mathbb{E}_n$ 表示对 n 个样本的经验期望。给定一个假设（回归器） F，我们的目标是学习 $F^\star$，以最小化平方 L^2 损失

$$R(F^\star, \hat{F}) \ :=\ \mathbb{E}_{\mathcal{S}_n}\big[\big\|F^\star(\nu, x_{\mathrm{q}}) - \hat{F}(\nu, x_{\mathrm{q}})\big\|^2_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})}\big].$$

内容测度的 RKHS 视角。 在陈述假设之前，我们先简要回顾核 K 如何诱导出一个用于对词元分布建模的函数空间。给定有界词元内容域 $\mathcal{X}_0 \subset \mathbb{R}^{d_2}$ 上的一个正定核 K：$\mathcal{X}_0 \times \mathcal{X}_0 \to \mathbb{R}$，存在唯一的再生核希尔伯特空间（RKHS）（例如，(Sch¨olkopf & Smola, 2002)） $\mathcal{H}_0$，由函数 $f:\mathcal{X}_0 \to \mathbb{R}$ 构成，使得

$$K(\cdot, x) \in \mathcal{H}_0 \quad \text{and} \quad f(x) = \langle f, K(\cdot, x)\rangle_{\mathcal{H}_0} \quad \text{for all } x \in \mathcal{X}_0.$$

直观而言， $\mathcal{H}_0$ 是与 K 兼容的“平滑”函数类。在 Mercer 基中

$$K(x, x') = \textstyle\sum_{j\ge1} \lambda_j e_j(x) e_j(x'),$$

每个 $f \in \mathcal{H}_0$都可以写成 $f(x) = \sum_{j\ge1} b_j e_j(x)$，且具有有限的 RKHS 范数 $\|f\|^2_{\mathcal{H}_0} = \sum_{j\ge1} b_j^2/\lambda_j$. 高频方向（较小的 λ_j）上的大系数 b_j 会受到严重惩罚，因此$\mathcal{H}_0$ 偏好能量集中于低阶特征分量上的函数。在我们的设定中，我们用 $\mathcal{X}_0$上的密度$^{p_{\mu_0^{(i)}}}$ 来表示每个内容测度 $\mu_0^{(i)}$，并要求这些密度位于 $\mathcal{H}_0$的一个固定球内。快速的特征值衰减 $\lambda_j \asymp \exp(-cj^\alpha)$对应于一个非常光滑的（高斯核型（Sch¨olkopf & Smola, 2002））情形，其中该函数类的有效维数较小；这一有效维数将驱动我们的统计速率。

3.1 假设与技术设定

现在陈述上界和下界分析所用的高层假设。完整的技术版本见第 A.1 节。在本小节中，上下文和查询均按照第 3 节开头所述的方式生成。

假设 1 （光滑核与正则内容测度）。 设$\mathcal{X}_0 \subset \mathbb{R}^{d_2}$ 为有界的词元内容定义域，并设 K：$\mathcal{X}_0 \times \mathcal{X}_0 \to \mathbb{R}$ 为一个具有 Mercer 展开（Sch¨olkopf 与 Smola，2002）的正定核

$$K(x, x') = \sum_{j\ge1} \lambda_j e_j(x) e_j(x').$$

我们假设：(i) 特征值呈指数衰减，即$\lambda_j \asymp \exp(-cj^\alpha)$，其中 $c,\ \alpha > 0$ （高斯型光滑情形）；(ii) 每个内容分布 $\mu_0^{(i)}$ 都是 $\mathcal{X}_0$上的有限测度，其密度属于由 K诱导的再生核希尔伯特空间 $\mathcal{H}_0$ 中的一个固定球。直观而言，词元分布非常光滑，且其有效维数较小，因为大部分质量都集中在低阶特征分量上。我们首先聚焦于这一指数衰减情形，以使分析更清晰。

例 1 (热核再生核希尔伯特空间（Grigor'yan，2006）)。 考虑 $\mathcal{X}_0 = [0,1]$ 及带有狄利克雷边界条件的拉普拉斯算子$-\Delta = \mathrm{d}^2/\mathrm{d}x^2$。其特征函数和特征值分别为$e_k(x) = \sqrt{2}\sin(k\pi x)$ 和 $\zeta_k = (k\pi)^2$，其中 $k \ge 1$。热核是热方程的基本解，用于描述在时刻0 将热量置于 y 后，热量如何在时刻 t之前扩散至 x，其表达式为

$$K_t(x, y) = \textstyle\sum_{k=1}^\infty e^{-\zeta_k t} e_k(x) e_k(y) = \sum_{k=1}^\infty e^{-\pi^2 k^2 t} e_k(x) e_k(y),$$

因此 *Mercer* 特征值满足 $\lambda_k = e^{-\pi^2 k^2 t} \asymp \exp\big(-ck^2\big)$，且假设 *1* 在 $\alpha = 2$ 下成立。紧流形上更一般的构造见第 *A.1* 节。

假设 2（分离的上下文向量）**.** 用于构造混合分布 (1) 的上下文向量 $v^{(i)} \in \mathbb{S}^{d_1-1}$ 充分分离： $\langle v^{(i)}, v^{(i')} \rangle \leq 0$, 对所有 $1 \leq i < i' \leq I$, 并且我们假设 $I \leq d_1$。这保证不同文档对于回忆步骤[2]而言足够可区分。

假设 3 (Lipschitz真实函数)。 存在一个度量 $\mathrm{d}_{\mathrm{prod}}$ ，定义在成对对象 $(\mu_0,\ x)$ 的空间上，并由上述再生核希尔伯特空间结构诱导，使得隐藏泛函 $\tilde{F}^\star$: $\{\mu_0^{(i)}\} \times \mathbb{R}^{d_1+d_2} \to \mathbb{R}$是 L-利普希茨的:

$$\left|\tilde{F}^\star(\mu_0, x) - \tilde{F}^\star(\mu_0', x')\right| \ \leq \ L\, \mathrm{d}_{\mathrm{prod}}\big((\mu_0, x), (\mu_0', x')\big)$$

对于所有可容许输入 $(\mu_0,\ x)$ 和 $(\mu_0',\ x')$。在证明中，$\mathrm{d}_{\mathrm{prod}}$ 将是密度之间由再生核希尔伯特空间诱导的距离与查询之间的欧几里得距离之和；其精确定义见附录 A.1。

假设 4. 每个内容分布 $\mu_0^{(i)}$ 都是 $\mathcal{X}_0$ 上的概率测度。

设定 1 (上界的概率设定（设定 1）)。 上下文和查询按照第 3 节中的方式生成。假设 1、2、3 和 4 均成立。换言之，我们考虑再生核希尔伯特空间球中的光滑（高斯型）内容概率分布、彼此充分分离的上下文向量，以及关于再生核希尔伯特空间诱导度量满足 L-利普希茨条件的真实函数。

下界的结构化模型。 对于极小极大下界，我们按照 Lanthaler (2024) 的做法，采用一个简化的内容密度随机模型： μ_0 的密度由 K的 Mercer 展开中的随机系数生成。

假设5 （密度的非正式结构模型）。 令 (λ_j, e_j) 为 K 的谱，如假设 1 所述。我们假设 $\frac{\mathrm{d}\mu_0}{\mathrm{d}\lambda}(x) \ = \ \sum_{j\geq 1} \lambda_j^{\Theta(1)} Z_j e_j(x)$，其中系数 Z_j 相互独立、有界且具有单位方差。完整的结构条件见第 A.1 节中的假设 8。

设定2 （下界的结构化设定（设定2））。 上下文和查询按照第 3 节中的方式生成。假设 1、2 和 3 成立，且内容测度的密度遵循假设5中的结构模型。在此设定下，我们在随机 Mercer 系数条件下推导极小极大下界。

3.2 学生模型：测度论 Transformer

我们遵循 Furuya 等人（2025）的工作，将学生模型定义为一类测度论 Transformer 架构。

测度论注意力。 给定一个词元集合 $X = (x_\ell)_{\ell=1}^w \in \mathbb{R}^{d_{\mathrm{attn}} \times w}$ 以及一个查询$x \in \mathbb{R}^{d_{\mathrm{attn}}}$ （其编码了有关部分词元的信息），一个参数为 $\theta^{(h)} = (K^h, Q^h, V^h)$ 的单个无掩码注意力头，以“上下文内”形式（Furuya 等，2025）计算

$$\mathrm{SAttn}_{\theta^{(h)}}(X, x) = \sum_j \mathrm{Softmax}\big(\langle Q^h x, K^h X\rangle\big)\, V^h x_j,$$

其中，$\mathrm{Softmax}((z_1,\ \ldots\ldots,\ z_N)) \coloneqq (\exp(z_i)/\sum_j \exp(z_j))_{i=1}^N$。随后，具有$H$个头的标准多头注意力表示为 $\mathrm{MSAttn}_\theta(X) \ = \ \sum_{h=1}^H W^h \mathrm{SAttn}_{\theta^{(h)}}(X,\ x)$，其中 $W^h \in \mathbb{R}^{d_{\mathrm{attn}} \times d_{\mathrm{attn}}}$。在无掩码情形下，注意力对词元索引具有置换等变性。这使我们能够用经验测度表示输入集，特别是表示为混合测度

$$\nu_X = \tfrac{1}{I}\sum_{i=1}^I \left(\tfrac{1}{w_i}\sum_{\ell_i=1}^{w_i} \delta_{v^{(i)}} \otimes \delta_{u_\ell^{(i)}}\right) \in \mathcal{P}(\mathbb{R}^{d_{\mathrm{attn}}}) \quad \text{in the limit as} \quad w \to \infty,$$

其中 $v^{(i)} \in \mathbb{R}^{d_1}$ （组共享，可能由查询指示）， $u_\ell^{(i)} \in \mathbb{R}^{d_2}$ （词元特定）对于$i \in [1:\ I]$, $\ell \in [1:\ w_i]$, $\sum_i w_i = w$ 并将注意力改写为测度论形式，在极限情况下，积分取代离散求和。

[2]在 Transformer 的事实提取理论分析中也采用了类似的分离/正交性结构，例如 Ghosal 等人（2024）

定义 2（测度论注意力层（Furuya 等，2025））。设 d_{attn} 为注意力的嵌入维度。测度论注意力层Attn $_\theta \quad \mathcal{P}\ \mathbb{R}^{d_{\text{attn}}} \times \mathbb{R}^{d_{\text{attn}}} \to :\ (\,)$ 定义为 $\mathbb{R}^{d_{\text{attn}}}$

$$\text{Attn}_\theta(\nu, x) = Ax + \sum_{h=1}^{H} W^h \int \text{Softmax}(\langle Q^h x, K^h y\rangle)\, V^h y \, \mathrm{d}\nu(y),$$

其中，Softmax(⟨Qx，　Ky⟩) ≔ exp (⟨Qx，　Ky⟩) / ∫exp (⟨Qx，　Kz⟩) dν(z)。此处，$A \in \mathbb{R}^{d_{\text{attn}} \times d_{\text{attn}}}$对跳跃连接应用一个学习得到的线性变换。

当 ν 是经验混合测度 ν_X时，这就简化为标准的离散注意力层（本文不探讨离散情形）。我们预期，查询 x 会根据组共享向量 $v^{(i^*)}$ 指示 ν_X 中i^*分量里的词元，模型便可将其召回。我们通过以下有界参数假设类来约束这些层。

定义 3（注意力假设类）。对于常数 B_a，　B'_a，　S_a，　S'_a，　$F > 0$，定义H头测度论注意力层类

$$\mathcal{A}(d_{\text{attn}}, H, B_a, B'_a, S_a, S'_a) \coloneqq \big\{\text{Attn}_\theta \mid \max_h \big\{\|W^h\|_\infty, \|Q^h\|_\infty, \|K^h\|_\infty, \|V^h\|_\infty\big\} \le B_a,$$
$$\max_h \big\{\|W^h\|_0, \|Q^h\|_0, \|K^h\|_0, \|V^h\|_0\big\} \le S_a, \|A\|_\infty \le B'_a,\ \|A\|_0 \le S'_a,\ \|\text{Attn}_\theta\|_\infty \le F\big\},$$

其中，　$\|M\|_\infty \coloneqq \max_{i,\ j} |M_{ij}|$, $\|M\|_0$表示矩阵 M中非零项的数量。为简便起见，我们假设 W^h, Q^h，　K^h，　V^h 均为方阵。当$B'_a = B_a$ 且 $S'_a = S_a$时，我们记作 $\mathcal{A}$(d_{attn}，　H，　B_a，　S_a)。

MLP层。除了注意力之外，Transformer块还包含一个前馈组件。由于这一部分不依赖于底层测度 μ，我们将其简单地建模为标准多层感知机（MLP），定义如下。

定义 4（MLP假设类）。设 $\ell \in \mathbb{N}$ 为深度，　$\boldsymbol{p} = (p_0,\ p_1,\ \ldots\ldots,\ p_{L+1}) \in \mathbb{N}^{L+2}$ 为层宽。架构为（ℓ，　$\boldsymbol{p}$）的神经网络是具有以下形式的任意函数

$$f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{\ell+1}}, \quad \mathbf{x} \mapsto f(\mathbf{x}) = W_\ell \sigma_{\mathbf{v}_\ell} W_{\ell-1} \sigma_{\mathbf{v}_{\ell-1}} \cdots W_1 \sigma_{\mathbf{v}_1} W_0 \mathbf{x},$$

其中，　$W_i \in \mathbb{R}^{p_{i+1} \times p_i}$ 是第 i 层的权重矩阵，$\mathbf{v}_i \in \mathbb{R}^{p_i}$ 是通过激活函数 $\sigma_{\mathbf{v}_i}(\mathbf{z}) : = \sigma(\mathbf{z} \ -\mathbf{v}_i)$ 施加的平移（偏差）向量，ReLU 激活函数为 σ：　$\mathbb{R} \to \mathbb{R}$。架构为（$\ell$，　$\mathbf{p}$）的多层感知机的假设类记作

$$\mathcal{F}(\ell, \boldsymbol{p}, s, F) \coloneqq \big\{f \text{ of the form above } \mid$$
$$\max_j\{\|W_j\|_\infty, |v_j|_\infty\} \le 1, \textstyle\sum_j \|W_j\|_0 + |v_j|_0 \le s,\ \|f\|_\infty \le F\big\}.$$

通过复合定义 Transformer 假设类。为了在我们的测度论框架中正式描述 Transformer，我们引入测度论映射的复合概念，该概念遵循 Furuya 等人（2025）的做法。这样，我们便可以将多层Transformer视为注意力层与前馈层的逐次复合，这与标准架构完全对应。

核心思想是，这样的映射 Γ 同时作用于一个词元 x 及其生成分布 μ。给定 $z \sim \mu$，应用 Γ 会生成变换后的词元 $\Gamma(\mu, z)$，并在变换后的词元上诱导出一个新分布 $\Gamma(\mu, z)$，即 ($\Gamma(\mu,$ ·))$\sharp\mu$。因此，复合两个映射意味着在样本层面和分布层面依次应用这些联合变换。

定义 5（测度论映射的复合）。设 Γ_1：$\mathcal{P}(\mathbb{R}^{d_1^{(1)}}) \times \mathbb{R}^{d_1^{(1)}} \to \mathbb{R}^{d_1^{(2)}}$ 和 Γ_2：$\mathcal{P}(\mathbb{R}^{d_2^{(1)}}) \times \mathbb{R}^{d_2^{(1)}} \to \mathbb{R}^{d_2^{(2)}}$，其中 $d_1^{(2)} = d_2^{(1)}$。它们的复合定义为

$$(\Gamma_2 \diamond \Gamma_1)(\nu, x) \coloneqq \Gamma_2(\mu_1, \Gamma_1(\nu, x)), \quad \text{where} \quad \mu_1 \coloneqq (\Gamma_1(\nu, \cdot))_\sharp \nu.$$

注 1. 如果我们将 ν 理解为经验测度 $\nu_X = \lim_{w\to\infty} \frac{1}{w} \sum_{\ell=1}^{w} \delta_{x_\ell}$的极限，那么复合 $\Gamma_2 \diamond \Gamma_1$ 的作用就是同时更新单个词元及其经验分布。在这种情况下，这一构造与 Transformer 中标准的逐层组合一致：每一层都将词元序列（x_1，$\ldots$，　x_w）映射为新的序列，同时相应地更新其经验分布。

基于上文引入的测度论复合概念，我们现在可以将 Transformer 正式描述为注意力层与前馈层的连续复合。以下定义给出了相应的假设类。

定义 6 (Transformer假设类)。 对于参数 $(d_j, H_j, B_{a,j}, B'_{a,j}, S_{a,j}, S'_{a,j}, \ell_j, \boldsymbol{p}_j, s_j)_{j=1}^{L}$，定义 TF 为如下形式的映射集合

$$\mathrm{Attn}_{\theta_L} \diamond \mathrm{MLP}_{\xi_L} \diamond \cdots \diamond \mathrm{Attn}_{\theta_1} \diamond \mathrm{MLP}_{\xi_1},$$

其中 $\mathrm{Attn}_{\theta_j} \in \mathcal{A}(d_j, H_j, B_{a,j}, B'_{a,j}, S_{a,j}, S'_{a,j})$ 和 $\mathrm{MLP}_{\xi_j} \in \mathcal{F}(\ell_j, p_j, s_j)$。MLP层与 μ 无关（即 $\mathrm{MLP}_{\xi_1}(\mu, x) \coloneqq \mathrm{MLP}_{\xi_1}(x)$），并假设所有中间输出均一致有界。

4 主要结果

4.1 测度论Transformer的估计误差

我们首先考察测度论 Transformer 在召回与预测任务中的泛化性能。在本小节中，我们均采用概率设定（设定 1），其中每个内容测度 $\mu_0^{(i)}$ 在 $\mathcal{X}_0$ 上具有平滑的再生核希尔伯特空间密度，且混合分量的数量不太大。

定理 1 （亚多项式收敛；定理3的非正式版本）。 设 $F^\star(\nu, x_{\mathrm{q}}) = \tilde{F}^\star(\mu_0^{(i^\star)}, x_{\mathrm{q}})$ 是第 *3* 节中的利普希茨回忆与预测映射，并假设底层核的特征值满足 $\lambda_j \asymp \exp(-cj^\alpha)$，其中某个 $\alpha > 0$。假设以下条件之一成立：*(i)* 混合分量的数量满足 $I \le d_1 \lesssim (\log\, n)^{1/(\alpha+1)}$，或 *(ii)* $\tilde{F}^\star$ 不依赖于x_{q} 和 $I \le d_1 = n^{o(1)}$。那么，对于适当选择的架构参数（定义一个深度为2^的测度论 *Transformer* 类 TF_n），任意经验风险最小化器 F_n 在 TF_n 上均满足

$$R(F^\star, \hat{F}_n) \;\lesssim\; \exp\big\{ -\Omega\big((\log n)^{\alpha/(\alpha+1)}\big)\big\}$$

在设定 *1* 下。

在统计上统一联想回忆与无限 token 区域。 以往研究分别探讨了 (i)*Transformer* 中的联想回忆（Ramsauer 等人，2021）和 (ii)以测度为模型的无限 *token*／无限维输入（Vuckovic 等，2020）。定理 1 将这两条研究脉络结合起来，给出了测度层面联想回忆的统计 理论：具有学习型 softmax 注意力的测度论 Transformer 能够在测度层面进行回忆与预测——稀疏地分离出 ν 中与查询相关的分量，并基于回忆出的测度进行预测；这不同于普适性或逼近结果（Geshkovski 等人，2024；Furuya 等，2025）。

非正式解释：有效维数。谱衰减 $\lambda_j \asymp \exp(-cj^\alpha)$意味着只有前几个 Mercer 模态承载着显著信号。经过回忆步骤后，我们的 Transformer 实际上聚合了前 D 个 Mercer系数

$$b_j \;\approx\; \int e_j \,\mathrm{d}\mu_0^{(i^\star)}, \quad j = 1, \ldots, D,$$

因此，无限维测度被压缩成 D 维向量 $b = (b_1, \ldots, b_D)$。基于 n 个样本学习关于 b 的利普希茨函数，其表现如同一个 D 维问题（参见 Schmidt-Hieber (2020)），其估计误差大致为

$$\text{Error of } D\text{-dim. problem} \;\approx\; n^{-\Theta(1/D)} \;\simeq\; \exp\big(-\Theta((\log n)/D)\big),$$

而在 D 个模态后截断 Mercer 展开会带来阶为 $\exp(-cD^\alpha)$ 的偏差。平衡这些项可得有效维数 $D_{\mathrm{eff}}(n) \asymp (\log\, n)^{1/(\alpha+1)}$ 以及次多项式收敛速率

$$R(F^\star, \hat{F}_n) \;\approx\; \exp\big(-\Theta((\log n)^{\alpha/(\alpha+1)})\big)$$

如定理 1 所述。从这个意义上说，尽管每个分量都是无限维测度，估计量的行为却仿佛只拟合了 $D_{\mathrm{eff}}(n)$ 个自由度。作为最基本的合理性检验，D 节给出了一项合成实验，其收敛速率与此一致。

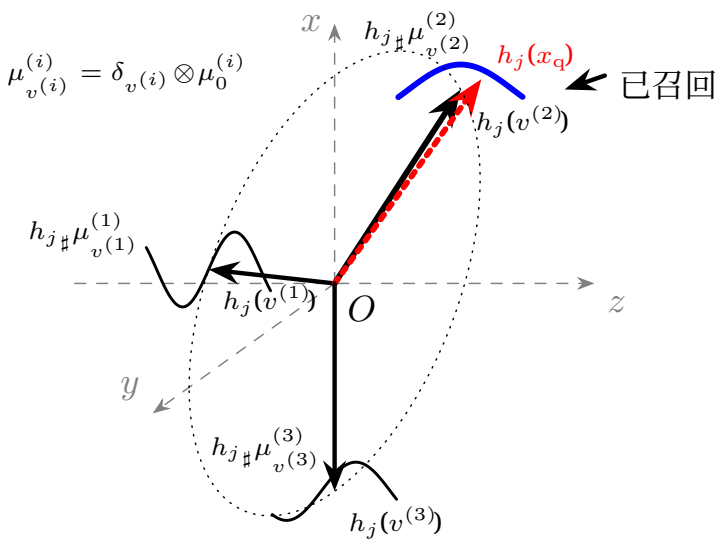

图 2：查询 x_q 和分量 $\mu^{(i)}_{v^{(i)}}$ 如何经由（简化的）第一层 $h_j((v,\ z)) = (v,\ e_j(z)) \in \mathbb{R}^{d_1+d_2}$ 进行映射的几何示意图，其中 e_j 是第 j个 Mercer 特征函数。 $h_j(x_q)$ 和 $h_j(y)$ 的前 d_1 个指标的乘积 $\sim h_j\sharp\mu^{(i)}_{v^{(i)}}$ 用于判断 $y = (v,\ z)$ 是否从 $\mu^{(i^*)}{}_{v^{(i^*)}}$ 中采样。

机制：softmax 注意力作为测度值联想记忆。 给定一个混合分布上下文 $\nu = I^{-1}\sum_{i=1}^{I}\mu^{(i)}_{v^{(i)}}$ 和一个查询 $x_q = (v^{(i^\star)},0)$，我们的二层Transformer按如下方式工作。首先，多层感知机将每个词元 $y = (v^{(i)}, z)$嵌入为一个特征向量 $h(y)$，其中包含前 D 个Mercer特征 $(e_j(z))_{j=1}^{D}$。随后，softmax 注意力计算分数 $\langle Qh(x_q)\ Kh(y)\rangle$，并通过参数化使得仅当文档标签 $v^{(i)}$ 匹配 $v^{(i^\star)}$ 时，这些分数才主要较大；否则较小（图 2）。

随后，softmax 注意力计算分数 $\langle Qh(x_q)\ Kh(y)\rangle$，使得归一化后，权重集中于来自 $\mu^{(i^\star)}{}_{v^{(i^\star)}}$的样本。记 $w_{x_q}(y)$为词元 y对应的注意力权重，值路径对每个 j计算：

$$\hat{b}_j \approx \int e_j(z)\, w_{x_q}(y)\, \mathrm{d}\nu(y) \approx \int e_j(z)\, \mathrm{d}\mu_0^{(i^\star)}(z), \qquad j = 1,\dots,D,$$

$\hat{}$ 从而得到一个 D维的描述符 $b = (b_1,\ \dots,\ b_D)$，用于表征召回的分量测度。最后，多层感知机将 $(b,\ x_q)$映射为标量预测 $\tilde{F}^\star(\mu_0^{(i^\star)},\ x_q)$。这样，softmax 注意力过滤混合分布，仅保留相关分量，并整合其 Mercer 特征，因此，从无限维测度进行预测可归结为学习一个关于 D 个汇总统计量的函数。这种近乎独热、依赖查询的过滤超出了冻结核（Zhou 等，2024）和线性注意力（Kim 等，2024）的能力范围。

4.2 极小极大最优性（下界）

接下来，我们证明结构化设定（设定2）中的下界。我们表明，所得上界的指数基本上是紧的，尽管设定2的技术假设与设定 1 不同。

定理 2（极小极大下界）。 在设定*2*中，任何 $L^2(\mathbb{P}_{\nu,\ x_q}(\mathcal{H}_0))$ -估计量 $\hat{}F$ 均满足

$$\sup_{\tilde{F}^\star\in\mathcal{F}^\star} R(\hat{F}, F^\star) \gtrsim \exp\Big(-O((\ln n)^{\frac{\alpha}{\alpha+1}})\Big)$$

其中 $F^\star(\nu, x_q) := \tilde{F}^\star(\mu_0^{(i^*)}, x_q)$， α 是 $\mathcal{H}_0$ 的核的特征值衰减率。

Transformer 的最优性。结合定理 1，我们在定理 2 中给出的信息论下界表明，基于 Transformer 的经验风险最小化的统计收敛率α达到了极小极大速率 $(\ln\ n)^{\alpha+1}$，指数上仅相差乘法常数。换言之，任何方法都无法在 n 依赖性方面超越所述指数（至多相差普适常数），因此，学习型 softmax Transformer 在这一问题类中实现了最优样本复杂度。即使混合分量数 I 固定或缓慢增长，这一最优性仍然成立，这证实了学习型 softmax 注意力为测度级召回提供了恰当的归纳偏置。softmax Transformer 的这一极小极大最优性，与简单无限维回归的统计结果（Takakura & Suzuki, 2023）以及需要检索的有限维上下文学习场景（Nishikawa 等，2025）相一致。

技术贡献——极小极大下界。我们首先指出，从混合输入 ν 进行估计并不比从纯测度 $\mu_0^{(i^*)}$ 更容易，因此可将联想回忆简化为无限维利普希茨回归。通过截断 Mercer 系数并进行各向异性重标度——将先前的重标度论证（Lanthaler, 2024）推广到具有指数衰减的更一般几何结构——使诱导几何基本呈各向同性，从而能够嵌入经典的d维利普希茨类，并采用标准打包界。将这些界与经典结果（Yang & Barron, 1999）结合，可得与我们的上界相匹配的速率。

5 结论与讨论

我们提出了测度论联想记忆（召回）的概念，并证明学习型 softmax 注意力即使在无限维、测度值的集合场景中也能实现精确召回——这是冻结核无法实现、线性注意力也难以做到的。我们的分析进一步表明，Transformer 的统计效率并不局限于有限维上下文，并为其召回能力提供了有理论依据的解释。虽然目前的结果聚焦于特征函数光滑且谱指数衰减的情形，但也为拓展到更广泛的情形打开了大门：将速率推广到多项式衰减，并在分析中纳入特征函数光滑性，是构建更完整理论的自然后续方向。

大语言模型使用声明

大语言模型用于两个目的：校对并润色英文写作，以及帮助我们查找相关工作。我们未使用任何大语言模型助手来设计问题设定和构建证明。

5.0.1 致谢

RK 与 TS 部分获得 JSPS KAKENHI（24K02905）和 JST CREST（JPMJCR2115）的资助。本研究得到新加坡国家研究基金会、资讯通信媒体发展局在其可信技术资助计划下的支持，以及数字发展与信息部在人工智能访问教授计划（资助编号 AIVP-2024-004）下的支持。本材料中所表达的任何意见、发现和结论或建议均属于作者本人，并不反映新加坡国家研究基金会、资讯通信媒体发展局以及数字发展与信息部的观点。RK 获得东京大学“智能世界创新”国际研究生项目（IIW）的 2024 财年自主研究活动资助支持。

参考文献

Shun-ichi Amari. 通过阈值元件自组织网络学习模式和模式序列. *IEEE*计算机汇刊，C-21(11):1197–1206，1972.

Alberto Bietti、Vivien Cabannes、Diane Bouchacourt、Herv´e J´egou 和 L´eon Bottou。Transformer的诞生：记忆视角。载于神经信息处理系统进展（*NeurIPS*），2023。

Emmanuel Boissard。经验测度和占据测度在1-瓦瑟斯坦距离下收敛的简单界。概率电子期刊，16：2296–2333，2011。Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习器。神经信息处理系统进展，33：1877–1901，2020。Martin Burger、Samira Kabri、Yury Korolev、Tim Roith 和 Lukas Weigand。层归一化Transformer架构中自注意力动力学所产生的平均场模型分析。哲学汇刊*A*辑，383(2298)：20240233，2025。Vivien Cabannes、Elvis Dohmatob 和 Alberto Bietti。联想记忆的缩放定律。载于国际学习表征会议（*ICLR*），2024。Val´erie Castin、Pierre Ablin 和 Gabriel Peyr´e。注意力有多平滑？载于国际机器学习会议，第5817–5840页。PMLR，2024。Ronald R. Coifman 和 St´ephane Lafon。扩散映射。应用与计算调和分析，21(1)：5–30，2006。Linhao Dong、Shuang Xu 和 Bo Xu。语音Transformer：一种用于语音识别的无循环序列到序列模型。载于*2018*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），第5884–5888页。IEEE，2018。Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszko-reit 和 Neil Houlsby。一张图像胜过16×16个词：面向大规模图像识别的Transformer。载于国际学习表征会议，2021。鄂维南和Qingcan Wang。解析函数的深度神经网络逼近的指数收敛。中国科学：数学，61(10)：1733–1740，2018。ISSN 1869-1862。Qihang Fan、Huaibo Huang、Yuang Ai 和 Ran He。纠正线性注意力中的幅值忽略问题，2025。Takashi Furuya、Maarten V. de Hoop 和 Gabriel Peyr´e。Transformer是通用的上下文学习器。载于第十三届国际学习表征会议，2025。Borjan Geshkovski、Cyril Letrouit、Yury Polyanskiy 和 Philippe Rigollet。自注意力动力学中簇的涌现。神经信息处理系统进展，36：57026–57037，2023。Borjan Geshkovski、Philippe Rigollet 和 Dom`enec Ruiz-Balet。使用Transformer进行测度到测度插值。*arXiv*预印本 *arXiv:2411.04551*，2024。Borjan Geshkovski、Cyril Letrouit、Yury Polyanskiy 和 Philippe Rigollet。关于Transformer的数学视角。美国数学会通报，62(3)：427–479，2025。Gaurav Ghosal、Tatsunori Hashimoto 和 Aditi Raghunathan。理解用于事实知识抽取的微调，2024。Alex Graves、Greg Wayne 和 Ivo Danihelka。神经图灵机。*arXiv*预印本*arXiv:1410.5401*，2014。Alexander Grigor'yan。加权流形上的热核及其应用。*Cont.Math*，398 (2006)：93–191，2006。Dongchen Han、Yifan Pu、Zhuofan Xia、Yizeng Han、Xuran Pan、Xiu Li、Jiwen Lu、Shiji Song 和 Gao Huang。弥合差异：重新审视softmax与线性注意力。神经信息处理系统进展，37：79221–79245，2024。John J. Hopfield。具有涌现集体计算能力的神经网络与物理系统。美国国家科学院院刊，79(8)：2554–2558，1982。

John J. Hopfield。分级响应神经元具有与二态神经元类似的集体计算特性。美国国家科学院院刊，81(10):3088–3092，1984。Y. Jiang、G. Rajendran、P. Ravikumar和B. Aragam。大语言模型会梦见大象吗（当被告知不要这样做时）？ Transformer中的潜在概念关联与联想记忆。*arXiv*预印本 *arXiv:2406.18400*，2024。J. Kim、M. Kim和B. Mozafari。Transformer的可证明记忆容量。见国际学习表征会议（*ICLR*），2023。Juno Kim、Tai Nakamaki和Taiji Suzuki。Transformer是极小极大最优的非参数上下文学习器。载于A. Globerson、L. Mackey、D. Belgrave、A. Fan、U. Paquet、J. Tomczak和C. Zhang（编），神经信息处理系统进展，第37卷，第106667–106713页。Curran Associates, Inc.，2024。Teuvo Kohonen。相关矩阵记忆。*IEEE*计算机汇刊，C-21(4):353–362，1972。Samuel Lanthaler。利普希茨算子的算子学习：一种信息论视角，2024。Peilin Liu和Ding-Xuan Zhou。Transformer在分布回归中的泛化分析。神经计算，37(2):260–293，2025年1月。ISSN 0899-7667。S. Mahdavi、R. Liao和C. Thrampoulidis。Transformer中多头注意力的记忆容量。见国际学习表征会议（*ICLR*），2024。H. N. Mhaskar和Nahmwoo Hahm。用于函数逼近与系统辨识的神经网络。神经计算，9(1):143–159，1997年1月。ISSN 0899-7667。Beren Millidge、Tommaso Salvatori、Yuhang Song、Thomas Lukasiewicz和Rafal Bogacz。通用Hopfield网络：单次联想记忆模型的一般框架。见国际机器学习会议，第15561–15583页。PMLR，2022。Boaz Nadler、Stephane Lafon、Ioannis Kevrekidis和Ronald Coifman。扩散映射、谱聚类与福克–普朗克算子的特征函数。载于Y. Weiss、B. Sch¨olkopf和J. Platt（编），神经信息处理系统进展，第18卷。麻省理工学院出版社，2005。Kenji Nakano。联想器——一种联想记忆模型。*IEEE*系统、人和控制论汇刊，SMC-2(3):380–388，1972。Eshaan Nichani、Jason D. Lee和Alberto Bietti。通过联想记忆理解Transformer中的事实回忆。见国际学习表征会议（*ICLR*），2025。Spotlight报告，预印本 arXiv:2412.06538。Naoki Nishikawa、Taiji Suzuki、Atsushi Nitanda和Denny Wu。无限维数据上的双层神经网络：平均场机制下的全局优化保证。载于S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho和A. Oh（编），神经信息处理系统进展，第35卷，第32612–32623页。Curran Associates, Inc.，2022。Naoki Nishikawa、Yujin Song、Kazusato Oko、Denny Wu和Taiji Suzuki。非线性Transformer能够进行推理时特征学习。见第四十二届国际机器学习会议，2025。Hubert Ramsauer、Bernhard Sch¨afl、Anna Hopkins、Michael Widrich、Thomas Adler、Lukas Gruber、Markus Holzleitner、Zoltan Pavlovi´c、Geir Kjetil Sandve、Vidar Greiff、Johannes Brandstetter和Sepp Hochreiter。你只需要Hopfield网络。见国际学习表征会议（*ICLR*），2021。Fabrice Rossi、Nicolas Delannay、Brieuc Conan-Guez和Michel Verleysen。神经网络中的函数数据表示。神经计算，64:183–210，2005。ISSN 0925-2312。神经计算趋势：2004年第12届欧洲人工神经网络研讨会。

Michael E. Sander、Pierre Ablin、Mathieu Blondel和Gabriel Peyr´e。Sinkformers：具有双随机注意力的Transformer。载于Gustau Camps-Valls、Francisco J. R. Ruiz和Isabel Valera（编），第*25*届人工智能与统计国际会议论文集，《机器学习研究论文集》第151卷，第3515–3530页。PMLR，2022年3月28日至30日。

Johannes Schmidt-Hieber. 使用深度神经网络和 ReLU 激活函数进行非参数回归。《统计学年鉴》，48(4):1875–1897，2020。

BernhardSch¨olkopf 和 Alexander J Smola。 核学习：支持向量机、正则化、优化及其他。麻省理工学院出版社，2002。

Linhao Song、Jun Fan、Di-Rong Chen 和 Ding-Xuan Zhou。更正：使用深度 ReLU 网络逼近非线性泛函。《傅里叶分析与应用》，29(5): 57，2023。ISSN 1531-5851。

M.B. Stinchcombe。神经网络对连续泛函以及紧化上的连续函数的逼近。神经网络，12(3):467–477，1999。ISSN 0893-6080。

Taiji Suzuki。全局最优非凸神经网络训练的泛化界：利用无限维朗之万动力学进行输运映射估计。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin（编），神经信息处理系统进展，第 33 卷，第 19224–19237 页。Curran Associates, Inc.，2020。

Shokichi Takakura 和 Taiji Suzuki。无限维输入下 Transformer 对序列到序列函数的逼近与估计能力。载于国际机器学习会议，第 33416–33447 页。PMLR，2023。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力机制就是一切。载于神经信息处理系统进展 *(NeurIPS)*，2017。

James Vuckovic、Aristide Baratin 和 Remi Tachet des Combes。注意力的数学理论 n. arXiv预印本 *arXiv:2007.02876*，2020。

Jason Weston、Sumit Chopra 和 Antoine Bordes。记忆网络。载于神经信息处理系统进展 *(NeurIPS)*，2014。

Weichun Xia 和 Lei Shi. 通过扩散在流形上应用谱算法。*arXiv*预印本，202 4.

Yuhong Yang 和 Andrew Barron. 极小极大收敛速率的信息论确定。《统计学年鉴》，27(5):1564–1599，1999。ISSN 00905364, 21688966.

Junwen Yao、Jonas Mueller 和 Jane-Ling Wang. 基于自适应基函数层的函数型数据分析深度学习。载于 Marina Meila 和 Tong Zhang（编），第*38*届国际机器学习会议论文集，第139卷，《机器学习研究论文集》，第11898–11908页。PMLR，2021年7月18–24日。

Tian-Yi Zhou、NamjoonSuh、Guang Cheng 和 Xiaoming Huo. 神经网络对再生核希尔伯特空间泛函的逼近。*CoRR*，abs/2403.12187，2024。

A 预备知识与记号说明

我们首先确定本文通篇使用的记号和约定，包括基本的向量与矩阵运算、测度论记号，以及测度与其密度之间的标准对应关系。

对于整数 $N_1 \le N_2$ 和向量 $\boldsymbol{v} \in \mathbb{R}^N$，我们定义 $\boldsymbol{v}_{N_1:N_2} := (v_{N_1}, \dots, v_{N_2})^\top$。对于矩阵 A，我们定义 $\|A\|_0$ 为非零元素的个数，并定义 $\|A\|_\infty := \max_{i,j} |A_{i,j}|$。我们将 λ 记为 $\mathbb{R}^N$ 上的勒贝格测度，其中 N 为整数。 $f_\sharp \mu(\cdot) := \mu(f^{-1}(\cdot))$ 表示测度 μ 在映射 f 下的推前。对于可测空间 $\mathcal{X}$，我们记 $\mathcal{P}(\mathcal{X}) := \{\mu \mid \mu \text{ 是 } \mathcal{X}\}$ 上的概率测度， $\mathcal{M}_+(\mathcal{X}) := \{\mu \mid \mu \text{ 是 } \mathcal{X}\}$ 上的非负测度。

除非另有说明，我们用 $(\Omega, \mathcal{F}, \mathbb{P})$ 表示底层概率空间。对于可测函数 $f: \Omega \to \mathbb{R}$ 和 $1 \le p < \infty$，$L^p(\mathbb{P})$ 范数定义为 $\|f\|_{L^p(\mathbb{P})} := \left(\int_\Omega |f|^p \, d\mathbb{P} \right)^{1/p}$，而 $L^\infty(\mathbb{P})$ 范数定义为 $\|f\|_{L^\infty(\mathbb{P})} := \operatorname{ess\,sup}_{\omega \in \Omega} |f(\omega)|$。当底层测度可由上下文明确确定时，我们就简记为 $\|f\|_{L^p}$ 和 $\|f\|_\infty$。

对于随机变量 $X: \Omega \to \mathcal{X}$，我们将其分布（$\mathbb{P}$ 在 X 下的推前测度）记为 $\mathbb{P}_X(\cdot) := \mathbb{P}(X \in \cdot)$。关于 $\mathbb{P}$ 的期望记为 $\mathbb{E}[\cdot]$；如果 X 是一个服从分布 $\mathbb{P}_X$ 的随机变量，我们也将 $\mathbb{E}_X[f(X)]$ 记作 $\int f(x) \, d\mathbb{P}_X(x)$。

备注 2（测度与密度的识别）。令 λ 为 X 上的参考测度（例如勒贝格测度）。给定 $f \in \mathcal{H}$ 和一个常数 $c \in \mathbb{R}$，我们定义概率测度 μ 为

$$\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x) := f(x) + c,$$

其中选择 c 使得 $f + c \ge 0$ λ-几乎处处成立，并且 $\mu(X) = \int_X \big(f(x) + c\big)\, dx = 1$。在这种情况下，我们通过将 μ 与其密度 $f + c$ 识别来记 $\mu \in \mathcal{H}$。在采样 $\mu_0 \in B(\mathcal{H}_0, \|\|_{\mathcal{H}_0}\gamma_b)$ 时，请注意我们选择 c 为无随机性的，并且为简便起见，在本文中始终令 $c = 0$。

定义 7（Mercer 展开）。设 $\mathcal{H}_0$ 是定义域 $\mathcal{X}$ 上的再生核希尔伯特空间（RKHS），其再生核为 $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$。根据 Mercer 定理，$K$ 可展开为

$$K(x, x') = \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(x'),$$

其中，$(\lambda_j)_{j \ge 1}$ 是（非负且非递增的）Mercer 特征值，$(e_j)_{j \ge 1}$ 是相应的 L^2-标准正交特征函数。对于任意 $\mu \in \mathcal{H}_0$，其在特征基中的表示为

$$p_\mu = \sum_{j=1}^{\infty} b_j e_j, \quad \text{with coefficients } b_j = \langle p_\mu, e_j \rangle_{L^2}.$$

定义 8（广义再生核希尔伯特空间范数）。设 $\mathcal{H}_0$ 是再生核希尔伯特空间，其在 L^2 中的标准正交基为 $\{e_j\}_{j \ge 1}$，对应特征值为 $\{\lambda_j\}_{j \ge 1}$。对于测度 μ，若其在 $\mathcal{H}_0$ 中对应的函数可展开为

$$p_\mu = \sum_{j \ge 1} b_j e_j, \quad \text{where} \quad \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x) = p_\mu(x) + \text{constant}.$$

我们定义，对于 $a \in \mathbb{R}$，其广义范数为

$$\|\mu\|_{\mathcal{H}_0^a}^2 \;:=\; \|p_\mu\|_{\mathcal{H}_0^a}^2 := \sum_{j \ge 1} \lambda_j^{-a} b_j^2.$$

特殊情况包括：$a = 0$ 时，$\|\cdot\|_{\mathcal{H}_0^0}$ 与 L^2 范数一致；$a = 1$ 时，$\|\cdot\|_{\mathcal{H}_0^1}$ 是标准再生核希尔伯特空间范数；$a = -1$ 时，$\|\cdot\|_{\mathcal{H}_0^{-1}}$ 是最大均值差异范数。

定义9（度量球）。设 (X, d) 为度量空间。对于 $x \in X$ 和 $\epsilon > 0$，以 x 为中心、半径为 ϵ 的（闭）度量球为

$$B(x, \mathrm{d}, \epsilon) := \{\, y \in X \mid \mathrm{d}(x, y) \le \epsilon \,\}.$$

当 $\epsilon = 1$ 时，我们简记为 $B(x, \mathrm{d})$，并称其为单位球。

定义10（Lipschitz 函数）。设 (X, d) 为度量空间，且 $A \subset X$ 为固定的定义域。若函数 $f: A \to \mathbb{R}$ 满足以下条件，则称其在 A 上关于 d 是 *L-Lipschitz* 的：

$$|f(x) - f(x')| \le L \, \mathrm{d}(x, x') \quad \forall x, x' \in A.$$

所有此类函数构成的集合记为 $\mathrm{Lip}_L(A, \mathrm{d})$。

度量熵预备知识

在后续证明中，我们将反复用到度量熵中的标准概念。具体而言，覆盖、打包及其相关数为量化假设类的复杂度提供了一种便捷的方法。因此，我们在此汇总相关定义和基本引理。

定义11 （ϵ-覆盖）。设 (X,d) 是一个度量空间，$A \subset X$，且 $\epsilon > 0$。有限集合$\{x_1,\dots,x_N\} \subset X$ 称为 A 的 ϵ-覆盖 （相对于 d），如果

$$A \subset \bigcup_{i=1}^{N} B(x_i, \mathrm{d}, \epsilon),$$

其中，$B(x_i, \mathrm{d}\ \epsilon)$ 表示以 x_i 为中心、半径为 ϵ 的闭度量球。

定义1 12 （ϵ-打包）。设（X，d）为一个度量空间，$A \subset X$，且 $\epsilon > 0$。有限集合 $\{x_1,\dots,x_N\} \subset A$ 称为 A 的 ϵ-打包 （相对于 d），如果

$$\mathrm{d}(x_i, x_j) > \epsilon \quad \text{for all } i \neq j.$$

等价地，度量球 $B(x_i, \mathrm{d}\ \epsilon/2)$ 两两不相交。

定义 13 （覆盖数）。相对于 d，A 在尺度 ϵ 下的覆盖数 为

$$\mathcal{N}(A,\epsilon)_{\mathrm{d}} \ := \ \min\big\{\, N \mid \exists \epsilon\text{-covering of } A \text{ of size } N \,\big\}.$$

定义14 (填充数**)**。相对于 d，A在尺度 ϵ 下的填充数 为

$$\mathcal{M}(A,\epsilon)_{\mathrm{d}} \ := \ \max\big\{\, M \mid \exists \epsilon\text{-packing of } A \text{ of size } M \,\big\}.$$

引理 1 （覆盖–填充等价）。对于任意度量空间 (X,d)、任意 $A \subset X$ 以及 $\epsilon > 0$，有

$$\mathcal{M}(A, 2\epsilon)_{\mathrm{d}} \ \le \ \mathcal{N}(A,\epsilon)_{\mathrm{d}} \ \le \ \mathcal{M}(A,\epsilon)_{\mathrm{d}}.$$

*In p*尤其，覆盖数与填充数在常数因子意义下等价 *i* *n* *the* 尺度参数。

Lemma 2 （度量支配下的单调性）。设d 与 d' 为定义在 X 上的两个度量，并设 $c >$ 0 *suc*使得 $\mathrm{d}(x, y) \le c\,\mathrm{d}'(x, y)$ 对所有 $x, y \in X$ 成立。则对任意 $A \subset X$ 与 $\epsilon > 0$,

$$\mathcal{N}(A,\epsilon)_{\mathrm{d}} \ \le \ \mathcal{N}(A,\epsilon)_{c\,\mathrm{d}'} = \mathcal{N}(A, c^{-1}\epsilon)_{\mathrm{d}'}.$$

换言之：若 d 被 $c\mathrm{d}'$ 所支配，则相对于 $c\mathrm{d}'$ 的 ϵ-覆盖也是相对于 d 的 ϵ-覆盖，因此在 d 下的覆盖并不更难。

A.1 第 3.1 节的技术版本

现在我们将介绍两种数据生成的技术设定。首先给出用于建立估计上界的假设，然后介绍用于推导下界的其他结构假设。

A.1.1 上界的概率设定。

假设 6 （共同假设：再生核希尔伯特空间结构、正则性和利普希茨性：假设 1、2、3 的技术版本）。固定一个整数 $i^* \in \{1,\dots,I\}$。查询向量 x_{q}、输入测度 $\nu = \sum_i \delta_{v^{(i)}} \otimes \mu_0^{(i)}$ 和输出 $y = F^\star(\nu, x_{\mathrm{q}}) + \xi$ 按如下方式生成：

- 再生核希尔伯特空间设定：我们假设 $\mu_0 \sim \mathbb{P}_{\mu_0}$ 的密度属于空间 $\mathcal{H}_0$，忽略一个常数。$\mathcal{H}_0$ 是定义在有界定义域 $\mathcal{X}_0 \subset [-M, M]^{d_2}$ 上的再生核希尔伯特空间，并具有 Mercer 分解（例如，(Sch¨olkopf & Smola, 2002)）

$$K(x, x') = \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(x'), \quad \lambda_j \simeq \exp(-cj^{\alpha}),$$

其中 c，$\alpha > 0$ 且特征函数 $(e_j)_{j\ge 1}$ 是 L^2-正交归一的。μ_0 非负，且 $\mathrm{d}\mu_0\mathrm{d}\lambda$ $- c'$ 位于度量球 $B(\mathcal{H}_0,\ \|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}})$ 内，其中 $\gamma_{\mathrm{b}} > 0$，λ 为勒贝格测度，c'为不含任何随机性的全局常数。为简便起见，本文始终令 $c' = 0$。该密度是无限维的，但其信息内容高度集中在低阶分量中，高频贡献呈指数衰减。类似的再生核希尔伯特空间结构可见 Suzuki（2020）；Nishikawa 等（2022）；Zhou 等（2024）；Liu 与 Zhou（2025）。

- 特征函数在 $\mathcal{H}_0$上的光滑性： (a) $(e_j)_{j\geq 1}$ 一致有界，且在$[-M,M]^{d_2}$上解析；(b) 对某个 $\delta > 0$， $\mathcal{X}_0 \subset [-M+\delta, M-\delta]^{d_2}$ ； (c) 每个 e_j 都可表示为在 $[-M,M]^{d_2}$ 上绝对收敛的幂级数 $e_j(x) = \sum_{\boldsymbol{k}\in\mathbb{N}^{d_2}} a_{\boldsymbol{k}} x^{\boldsymbol{k}}$。这三个条件旨在聚焦于衰减率 α相关的样本复杂度。

- 采样上下文向量的可区分性：向量 $(\mathbb{S}^{d_1-1})^{\otimes I} \ni (v_{(i)})_{i=1}^{I} \sim \mathbb{P}_v$ 满足$\langle v_{(i)} v_{(i')}\rangle \leq 0, \quad 1 \leq i < i' \leq I$。为满足这一条件，我们还要求 $I \leq d_1$。这确保上下文具有足够的可区分性以实现召回。

- 目标泛函的 *Lipschitz* 连续性：请记住，输出生成为 $y := \tilde{F}^{\star}\big(\mu_{(i^*)\,0},\ x_{\mathrm{q}}\big) + \xi, \quad \xi \sim \mathcal{N}(0, \sigma^2)$。隐藏泛函 $\tilde{F}^{\star}$： $B(\mathcal{H}_0,\ |||_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}}) \times \mathcal{X}_{\mathrm{q}} \to \mathbb{R}$ 假设满足 Lipschitz 连续性：设乘积集合 $B(\mathcal{H}_0,\ |||_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}}) \times \mathcal{X}_{\mathrm{q}}$ 上的度量为

$$\mathrm{d}_{\mathrm{prod}}\big((\mu_0,x),(\mu_0',x')\big) := \|\mu_0 - \mu_0'\|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}} + \|x - x'\|_2, \quad \gamma_{\mathrm{f}} < 0,$$

其中 $\mu_0,\ \mu_0' \in B(\mathcal{H}_0,\ |||_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}})$ 和 $x,\ x' \in \mathcal{X}_{\mathrm{q}}$。我们将 $\mu_0 \in \mathcal{H}_0$定义为：将 μ_0与其在 $\mathcal{H}_0$ 中的密度相对应，并忽略一个加性常数（见备注 2）。对于 $a \in \mathbb{R}$， $\mathcal{H}_0$ 上的 $\mathcal{H}_0^a$-范数定义为 $\|\mu_0\|_{\mathcal{H}_0^a}^2 := \big\| \frac{\mathrm{d}\mu_0}{\mathrm{d}\lambda} \big\|_{\mathcal{H}_0^a}^2 := \sum_{j\geq 1} \lambda_j^{-a} b_j^2$其中 $\frac{\mathrm{d}\mu_0}{\mathrm{d}\lambda} = \sum_j b_j e_j$。我们假设 $\tilde{F}^{\star}$属于 $\mathrm{Lip}_L(B(\mathcal{H}_0,\ |||_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}}) \times \mathcal{X}_{\mathrm{q}},\ \mathrm{d}_{\mathrm{prod}})$，这是关于 $\mathrm{d}_{\mathrm{prod}}$的 L-Lipschitz 泛函集合。

这些假设规定了再生核希尔伯特空间的解析与结构条件，以及上下文的可区分性；在整个分析过程中，我们都会采用这些假设。下面给出一个满足这些假设的例子：

示例 2 （假设 6 中的特征值快速衰减）。 已知（*Grigor'yan，2006*），在紧黎曼流形 $\mathcal{M}$ 上，拉普拉斯-贝尔特拉米算子具有满足韦尔定律的离散谱：热核展开为 $p_t(x,y) = \sum_{k=0}^{\infty} e^{-\Theta(k^{2/n})\, t} \varphi_k(x)\, \varphi_k(y)$，其中 $\{\varphi_k\}_k$ 为特征函数， $n =$ 维数（$\mathcal{M}$），且 $t > 0$。特征值快速衰减的设定已作为一种数据结构得到研究（*Nadler* 等，*2005*；*Coifman & Lafon，2006*；*Xia & Shi，2024*）。尽管高斯核的支撑集并非紧集，但它们也广泛用于机器学习任务（例如 *Sch¨olkopf & Smola*（*2002*））。

接下来，我们将输入测度限制为概率测度：

假设7 （设定 3 的概率假设：假设 4 的技术版本）。 令$\mathcal{P}(X)$表示可测集 X上的博雷尔概率测度集合。对于 $\mu_0 \sim \mathbb{P}_{\mu_0}$， $\mu_0 \in \mathcal{P}(\mathcal{X}_0)$ 几乎处处成立（概率为 1）。

基于这些假设，我们可以概述上界分析的概率设定：

设定 3 （上界的概率设定：设定 1 的技术版本）。 假设 6 和假设7均成立。简言之：$\mu_0 \sim \mathbb{P}_{\mu_0}$ 被约束为一个概率测度，其密度属于再生核希尔伯特空间球。

A.1.2 下界的结构化设定。

对于极小极大下界，我们放宽假设7中的概率约束，转而
摭弼 Mercer 展开的系数的结构条件（参照 Lanthaler (2024)） n:

假设 8 （设定 4 的结构假设：假设5的技术版本）。 设 $(\Omega\ \mathbb{P})$ 为一个概率空间。我们不再要求 $\mu_0 \sim \mathbb{P}_{\mu_0}$ 是概率测度，而是按定义1如下生成 μ_0：

- 随机再生核希尔伯特空间元素： 与 μ_0 相关的"密度"$p_{\mu_0} : \Omega \to \mathcal{H}_0$ 展开为

$$\frac{\mathrm{d}\mu_0}{\mathrm{d}\lambda}(\omega)(\cdot) := p_{\mu_0}(\omega)(\cdot) = \sum_{j=1}^{\infty} \lambda_j^{\gamma_{\mathrm{d}}} Z_j(\omega) e_j(\cdot), \quad \omega \in \Omega,$$

其中，$(e_j)_{j\geq 1}$是 $\mathcal{X}_0$上的 L^2-标准正交基， $\gamma_{\mathrm{d}} > 0$， 且$\lambda_1^{\gamma_{\mathrm{d}}} \geq \lambda_2^{\gamma_{\mathrm{d}}} \geq \cdots \geq 0$可求和。

• 随机系数：变量 $Z_j : \Omega \to \mathbb{R}$相互独立，满足 $\mathbb{E}[|Z_j|^2] = 1$，　$Z_j \sim \rho_j(z)\,dz$，并且满足一致界 $\sup_j \|\rho_j\|_\infty \le R$，　$\lambda_1^{\gamma_d/2} \le R$，其中$R > 0$。

例 3. 假设 *6* 和假设 *8* 相容：取 $\rho_j = \frac{1}{2}\mathbb{1}_{\{-1,1\}}$ 和 $\lambda_j \le A\exp(-cj^\alpha)$。那么，只要 $\gamma_d > \gamma_b$ 且 $\|\mu_0\|_{\mathcal{H}_0^{\gamma_b}} = O(A^{\gamma_d-\gamma_b})$，$\|\mu_0\|_{\mathcal{H}_0^{\gamma_b}} \le \sum_j \lambda_j^{\gamma_d-\gamma_b}$就总是收敛。因此，存在某个 $A > 0$ 使得$\|\mu_0\|_{\mathcal{H}_0^{\gamma_b}} \le 1$ 几乎处处成立。

最后，我们总结相应的设定：

设定 4（下界结构化设定：设定2 的技术版本）。 假设 6 和假设 8 均成立。简而言之：密度以 Mercer 展开的形式进行抽样 $\mathrm{d}\mu_0/\mathrm{d}\lambda := \sum_j \lambda_j^{\gamma_d} Z_j e_j$ 其中 Z_j 是相互独立的随机变量，且 μ_0 可能不是概率测度。

B 估计误差分析（上界）

本节的总体目标是为设定 3 中基于Transformer的估计器推导统计上界。具体而言，我们证明经验风险最小化器能够达到如下形式的收敛速率

$$R(\hat{F}, F^\star) \lesssim \exp\big(-\Omega((\log n)^{\alpha/(\alpha+1)})\big).$$

该速率可视为经典 $n^{-\Theta(1/d)}$ 风险界在 d 维回归中的无限维类比，其中有效维数按 $d \sim (\log n)^{1/(\alpha+1)}$ 的速度增长。特别是，尽管联想回忆任务需要处理测度值分量，我们的分析表明，其统计复杂度与纯无限维回归问题相同；该问题的极小极大下界将在 C 节中给出。接下来的各小节将通过依次推导网络各层的逼近界，逐步证明这一结果。

B.1 估计上界的证明概要

本部分将说明如何证明以下定理：

定理3（亚多项式收敛，对应定理 1，即定理5与定理6的简化版本）。设 $F^\star \in \mathrm{Lip}_1(B(\mathcal{H}_0, \|\|_{\mathcal{H}_0{}^{\gamma_b}}) \times \mathcal{X}_q, \mathrm{d}_{\mathrm{prod}})$ 并假设以下情况之一成立：

(i) 混合分量的数量有界，即 $I \le d_1 \lesssim (\ln n)^{\frac{1}{\alpha+1}}$；

(ii) 隐藏目标函数 $\tilde{F}^\star$ 与 x_q 和 $I \le d_1 \simeq n^{o(1)}$相互独立。

设 $\hat{}F$ 为以 *Transformer* 类 *TF(ϵ)* 为映射集合的经验风险最小化器，其中映射为 $Attn_{\theta_2} \diamond MLP_{\xi_2} \diamond Attn_{\theta_1} \diamond MLP_{\xi_1}$，满足 $H_1, B_{a,1}, \ell_2 = (\log \epsilon^{-1})^{O(1)}$，$\|p_2\|_\infty, s_2 \lesssim \exp\Big(O((\log\epsilon^{-1})^{1+\alpha^{-1}})\Big)$, $d_{\mathrm{attn}_2}, S'_{a,2}, H_2, B'_{a,2} = 1$, $S_{a,2}, B_{a,2} = 0$，在两种情况下均成立，且在 *(i)* 中：　$\ell_1, \|p_1\|_\infty, s_1, d_{\mathrm{attn}_1}, S_{a,1} = (\log \epsilon^{-1})^{O(1)}$；在 *(ii)* 中：　$\ell_1, \|p_1\|_\infty, s_1, d_{\mathrm{attn}_1}, S_{a,1} \lesssim \exp\Big(O((\log\epsilon^{-1})^{1+\alpha^{-1}})\Big)$。那么在设定 *3* 中，

$$R(F^\star, \hat{F}) \lesssim \exp\big(-\Omega((\ln n)^{\frac{\alpha}{\alpha+1}})\big),$$

其中，α 是 $\mathcal{H}_0$的底层核的特征值衰减率。

我们将给出第一种情况的证明概要：(i) 混合分量的数量有界为 $I \le d_1 \lesssim (\ln n)^{\frac{1}{\alpha+1}}$。

为推导统计收敛率，我们必须校准测度论 Transformer 假设类的大小。具体而言，我们选择一种架构，使其能够以 ϵ 逼近 $F^\star$，同时令覆盖熵 $V(\delta) := \log \mathcal{N}(\mathrm{TF}(\epsilon); \delta)_\infty$ 最小；随后，通过平衡逼近误差 ϵ 与由 $V(\cdot)$ 决定的估计项，即可得到风险界。简言之，一般理论

Schmidt-Hieber (2020) 指出，超额风险可以由逼近项与复杂度项之和界定：对于任意 $\delta > 0$， L^2-风险 R 满足

$$R(F^\star, \hat{F}) \lesssim \inf_{\hat{F}} \|\hat{F} - F^\star\|^2_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})} + \delta + \frac{V(\delta)}{n}$$

这里，第一项衡量架构对 $F^\star$ 的逼近效果，而第二项反映了在大小为 $V(\delta)$ 的类中进行搜索所付出的统计代价。因此，我们的证明首先逐层控制 $V(\cdot)$，然后选择 ϵ，以实现最优权衡。

我们的估计界依赖于两个主要要素：(i) 一种通过深度为 $L = 2$ 的 Transformer 表示目标泛函 $F^\star(\nu, x_{\mathrm{q}}) = \tilde{F}^\star(\mu_0^{(i^\star)}, x_{\mathrm{q}})$ 的逼近策略；(ii) 架构各分量的覆盖熵界，并通过测度论映射的复合引理将其结合起来。

步骤 1：复合引理（B.2节）。我们首先给出一个关于测度论映射复合的覆盖数的一般结果。

引理 3（复合引理）。设 $\mathcal{G}_i$、$i = 1$、2 是映射集合 $\Gamma_i \colon \mathcal{P}(\mathcal{X}_{(1)i}) \times \mathcal{X}_{(1} \underset{i}{\to} \mathcal{X}_i^{(2)})$，满足 $\mathcal{X}_1^{(2)} \subset \mathcal{X}_{(1)2}$、$\mathcal{N}(\mathcal{G}_i; \epsilon)_\infty \lesssim N_i$，且任意 $\Gamma_2 \in \mathcal{G}_2$ 关于 $(L_{2,1}$、$L_{2,2})$，相对于 1-*Wasserstein* 度量和欧氏度量均为利普希茨连续。那么，

$$\mathcal{N}(\{\Gamma_2 \diamond \Gamma_1; \Gamma_i \in \mathcal{G}_i, i = 1, 2\}; \epsilon)_\infty \lesssim \mathcal{N}(\mathcal{G}_2; \tfrac{\epsilon}{2})_\infty \cdot \mathcal{N}(\mathcal{G}_1; \tfrac{\epsilon}{2(L_{2,1}+L_{2,2})})_\infty.$$

证明遵循标准的有限维复合论证：用最近的覆盖元素逼近每个映射 Γ_i，并利用 Lipschitz 常数界定复合映射之间的差异。

步骤 2：用深度为2 的 Transformer。我们使用以下架构逼近 $F^\star(\nu, x) = \tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}})$，重点关注前 $D \gtrsim (\log \epsilon^{-1})^{\alpha^{-1}}$ 个 Mercer 系数：

1. **第一 MLP 层（B.3 节）**。构造 MLP_{ξ_1}，将输入 (x_1, x_2) 扩充为解析基 $\{e_i\}$ 的取值 $(e_i(x_2))$，误差不超过 $O(\epsilon_1)$。解析性意味着 $\log \mathcal{N}(\mathcal{F}_1; \epsilon_1)_\infty \lesssim \mathrm{poly} \log \epsilon_1^{-1} \cdot \mathrm{poly} D$。
2. **第一注意力层（B.4 节）**应用 Attn_{θ_1} 于 $\mathrm{MLP}_{\xi_1}(\mu)$，计算经验均值 $\int e_j(y_2)\, \mathrm{d}\mu_0(y_2)$，误差不超过 $O(\epsilon_1 + \epsilon_2)$。我们逼近测度的独热选择，并以非正式的方式计算：

$$\int \underbrace{\mathrm{Softmax}(\mathrm{MLP}_1(x_{\mathrm{q}})^\top Q^\top K\ \mathrm{MLP}_1(y))}_{\mathbb{1}[y \sim \mu^{(i^*)}_{v^{(i^*)}}]} \mathrm{MLP}_1(y) \underbrace{\mathrm{d}\nu(y)}_{\propto \mathrm{d}\sum_i \mu^{(i)}_{v^{(i)}}} \simeq \int \mathrm{MLP}_1(y) \mathrm{d}\mu^{(i^*)}_{v^{(i^*)}}(y).$$

根据第一个 MLP_1 层的构造以及 $\mu_{v^{(i^*)}}$ 的乘积分解，右侧近似于 $\int e_j(y_2) \mathrm{d}\mu_0(y_2^{(i^*)})$。在注意力矩阵满足稀疏性约束的情况下，

$$\log \mathcal{N}(\mathcal{A}_1(d_{\mathrm{attn}}, H, B_a, S_a); \epsilon_2)_\infty \lesssim \mathrm{poly} D \cdot \mathrm{poly} \log\big(I \epsilon_2^{-1}\big).$$

3. **第二个 MLP 层（B.5 节）**。近似定义在 $\mathbb{R}^{D+d_1+d_2}$ 上的利普希茨映射，其输入为保留的 Mercer 系数（$\int e_j(y_2) \mathrm{d}\mu_0(y_2))_{j=1}^D$ 和查询 x_{q}。我们还证明，$D \gtrsim (\log \epsilon^{-1})^{\alpha^{-1}}$（这是 ϵ，而非 ϵ_1）足以提取特征。对于 $c_\epsilon \lesssim \mathrm{poly} \underline{\log} \log\ \epsilon^{-1}$，该层的 Lipschitz 常数 $L_{\mathrm{MLP},\ 2} \lesssim e^{-O(c_\epsilon (\log\ \epsilon^{-1})^{\ 2+2\alpha}_{\alpha})}$ 可能很大；通过引理 5 中讨论的测度论复合，这会要求 $\epsilon_1 \vee \epsilon_2$ 小到超多项式级。4. **第二个注意力层（B.6 节）**。该层以（一维）跳跃连接的形式实现，并加入 Lipschitz 常数为 $O(1)$ 的项，以符合形式假设集的定义。

步骤 3：对覆盖熵进行上界估计（B.7 节）。对四层递归应用复合引理，可得

$$\log \mathcal{N}(\mathrm{TF}; \epsilon)_\infty \lesssim \log \mathcal{N}(\mathcal{A}(d + D, 1, 0, 1, 0, d + D); \epsilon)_\infty + \log \mathcal{N}(\mathcal{F}(\ell_2, p_2, s_2); \tilde{\Omega}(\epsilon))_\infty$$

$$+\log\mathcal{N}(\mathcal{A}(d_{\text{attn}},H,B_a,S_a);\tilde{\Omega}(L_{\text{MLP},2}^{-1}\epsilon))_\infty$$
$$+\log\mathcal{N}(\mathcal{F}(\ell_1,p_1,s_1);\tilde{\Omega}(L_{\text{MLP},2}^{-1}(L_{\text{Attn},1,W^1}+L_{\text{Attn},1,\|\|_2})^{-1}\epsilon))_\infty.$$

由此可得（同时仔细地对上文所省略的关于 d_1 的项作上界估计）：

引理 4. *Transformer* 类的覆盖熵满足

$$\log\mathcal{N}(\text{TF};\epsilon)_\infty\lesssim\exp\left(O((\log\epsilon^{-1})^{(1+\min(\alpha,\beta))/\min(\alpha,\beta)})\right)$$

其中 α 是再生核希尔伯特空间 $\mathcal{H}_0$ 的衰减率，且 $d_1\simeq(\ln\epsilon^{-1})^{\beta^{-1}}$。

步骤 4：从覆盖熵到风险界（B.7 节）。 应用 Schmidt-Hieber (2020) 中给出的回归界，其中 $V(\delta)=\log\mathcal{N}(\text{TF};\delta)_\infty$，并选取

$$\epsilon\simeq\exp\left(-\Theta(\log n)^{\frac{\min(\alpha,\beta)}{\min(\alpha,\beta)+1}}\right),$$

其中 $d_1\simeq(\ln\epsilon^{-1})^{\beta^{-1}}$，我们得到次多项式收敛速率：

定理 4 （亚多项式收敛，定理 1 的推广版本）。 在概率设定（设定 *3*）中，

$$\sup_{F^\star\in\mathcal{F}^\star}R(\hat{F},F^\star)\lesssim\exp\left(-\Omega\big((\log n)^{\frac{\min(\alpha,\beta)}{\min(\alpha,\beta)+1}}\big)\right),$$

其中，特征值为 $\lambda_j\simeq\exp(-cj^\alpha)$，混合分量的数量为 $I\le d_1\lesssim(\ln n)^{\beta^{-1}\cdot\min(\alpha,\beta)/(\min(\alpha,\beta)+1)}$。

B.2 步骤 1：复合引理。

引理 5 (复合引理，重述)。 设 $\mathcal{G}_i$, $i=1,2$ 为 Γ_i 的集合，这些函数是从$\mathcal{P}(\mathcal{X}_{(1)i})\times\mathcal{X}_{(1)i}$到 $\mathcal{X}_{(2)i}$的映射，且满足 $\mathcal{X}_{(2}{}_1\subset\mathcal{X}_2^{(1)})$, $\mathcal{N}(\mathcal{G}_i;\ \epsilon)_\infty\lesssim N_i$。此外，任意 $\Gamma_2\in\mathcal{G}_2$ 关于$(L_{2,1},L_{2,2})$，相对于 1-*Wasserstein* 距离和欧几里得距离都是利普希茨连续的。于是，我们有

$$\mathcal{N}(\{\Gamma_2\diamond\Gamma_1\mid\Gamma_i\in\mathcal{G}_i\};\epsilon)_\infty\lesssim\mathcal{N}(\mathcal{G}_2;\frac{\epsilon}{2})_\infty\cdot\mathcal{N}(\mathcal{G}_1;\frac{\epsilon}{2(L_{2,1}+L_{2,2})})_\infty.$$

*Pro*关于。 首先请记住，对于标准的一维函数类 $\mathcal{F}$，$\mathcal{G}$，其覆盖数 -bers N_f, N_g,

$$\mathcal{N}(\{f\circ g\mid f\in\mathcal{F},g\in\mathcal{G}\};\epsilon)\lesssim\mathcal{N}(\mathcal{F};\epsilon)\cdot\mathcal{N}(\mathcal{G};\epsilon/L_f)$$

其中， L_f 是 $\forall f\in\mathcal{F}$的 Lipschitz 常数上界。对于测度论映射 $\Gamma_1\in\mathcal{G}_1$和 $\Gamma_2\in\mathcal{G}_2$，取 Γ_1^i和 Γ_2^j为（ϵ_1和 ϵ_2的最近覆盖元素（即 $\sup_{\mu,\ x\in\mathcal{P}(\mathcal{X}_i^{(1)})\times\mathcal{X}_i^{(1}}|\Gamma_1)$（$\mu$，$x$）$-\Gamma_1^i$（$\mu$，$x$）$|\le\epsilon$）。然后，我们对复合之差进行界定：

$$\begin{aligned}&|(\Gamma_2\diamond\Gamma_1)(\mu,x)-(\Gamma_2^j\diamond\Gamma_1^i)(\mu,x)|\\&=|(\Gamma_2\diamond\Gamma_1)(\mu,x)-(\Gamma_2\diamond\Gamma_1^i)(\mu,x)|+|(\Gamma_2\diamond\Gamma_1^i)(\mu,x)-(\Gamma_2^j\diamond\Gamma_1^i)(\mu,x)|\\&\le L_{2,1}W_1(\Gamma_1(\mu)_\sharp\mu,\Gamma_1^i(\mu)_\sharp\mu)+L_{2,2}\|\Gamma_1^i(\mu,x)-\Gamma_1(\mu,x)\|_2+\epsilon_2\\&\le(L_{2,1}+L_{2,2})\epsilon_1+\epsilon_2.\end{aligned}$$

□

B.3 步骤 2-1：第一 MLP 层

我们首先形式化第一层多层感知机的逼近性质。该层负责将输入词元和辅助解析特征嵌入到更高维的表示中。此层对于确保后续的注意力层和多层感知机层能够在具有足够表达能力的特征空间中运作，起着至关重要的作用。

引理 6 2018 f $[-M,M]^{d_2}$ $\mathcal{X}_0 \subset$(E& $[-M+\delta, M-\delta]^{d_2}$ $\delta > 0$ $f\ \boldsymbol{x} = \sum_{\boldsymbol{k}\in\mathbb{N}^{d_2}} a_{\boldsymbol{k}}\boldsymbol{x}^{\boldsymbol{k}}$ Wang())。 设 是定义在 上的解析函数，使得对于某个，幂级数 () 在 $[-M,M]^{d_2}$ 上绝对收敛。那么，一个深层 *ReLU* 网络 $\hat{}\, f$，其深度为 $O((\log \epsilon^{-1})^{2d_2})$，宽度为$d_2+4$ （与 ϵ 无关），满足

$$\sup_{x\in\mathcal{X}_0} |f(x) - \hat{f}(x)| \lesssim \epsilon.$$

这里，记号 $\boldsymbol{x}^{\boldsymbol{k}}$ 表示多元单项式 $\prod_{i=1}^{d_2} x_i^{k_i}$，而绝对收敛条件保证幂级数在 $[-M,M]^{d_2}$上一致收敛，从而能够在内部区域 $\mathcal{X}_0$上实现一致逼近。

作为引理 6 的一个推论，我们得到如下结果：

推论 1. 设 $d = d_1 + d_2$。对于函数 f，使得

$$f : [-M,M]^{d_1+d_2} \to \mathbb{R}^{d+D}, \quad f(x_1,x_2) = \begin{bmatrix} x_1 \\ x_2 \\ e_1(x_2) \\ \vdots \\ e_D(x_2) \end{bmatrix},$$

其中 e_j 在假设 *6* 中定义。也就是说， f保留前 d_1+d_2 个坐标(x_1, x_2)，并用仅依赖于 x_2的 D 个解析特征函数 e_j 对其进行增广。存在一个网络$\hat{}f \in \mathcal{F}(\ell_1, \boldsymbol{p_1}, s_1, \infty)$，其中 $\ell_1 \lesssim (\log\ \epsilon_1^{-1})^{2d_2}$, $p_{1,j} \lesssim D + d$, $s_1 \lesssim d + (\log\ \epsilon_1^{-1})^{2d_2} \cdot d_2^2 \cdot D$，使得

$$\|f - \hat{f}\|_\infty \lesssim \epsilon_1.$$

证明。 对于前 d_1 个指标，我们模拟 $x_{1,\ i} = -$ReLU$(-\boldsymbol{e}_i\boldsymbol{e}_i^\top \cdot x_1)$ +ReLU$(\boldsymbol{e}_i\boldsymbol{e}_i^\top \cdot x_1)$。这需要 O(1) 深度、 $O(d_1)$ 宽度和 $O(d_1)$ 个参数。第 $i+d_1$ 个（$i = 1,$ ....., d_2）指标需要 O(1) 深度、 $O(d_2)$ 宽度和 $O(d_2)$ 个参数。其余指标参见引理 6。 □

换言之，每个解析分量 e_j 都可以用参数数量呈对数级的 ReLU 网络进行一致逼近，而拼接映射 f 可以由一个具有上述参数界的分块结构网络表示。

为计算覆盖数，我们使用以下引理：

引理 7 （Schmidt-Hieber (2020)）。设 $V = \prod_{i=0}^{\ell}(p_i+1)$。则，

$$\log \mathcal{N}(\mathcal{F}(\ell,p,s,F);\delta)_\infty \lesssim (s+1)\log\big(\delta^{-1}\ell V\big).$$

覆盖数是

$$\log\ \mathcal{N}(\mathcal{F}(\ell_1,p_1,s_1);\ \epsilon_1)_\infty \lesssim (d + (\log\ \epsilon_1^{-1})^{2d_2} \cdot D)\ \cdot (\log\ \epsilon_1^{-1} + \log\log\ \epsilon_1^{-1} + (\log\ \epsilon^{-1})^{2d_2}\log(d+D)))$$

$$\lesssim C_{D,\epsilon_1}(d + D(\log\epsilon_1^{-1})^{4d_2}),$$

其中 $C_{D,\epsilon_1} \lesssim$多对数 $D+$多对数 $d+$多对数对数 ϵ_1^{-1}。

B.4 步骤 2-2：第一注意力层

在上一节中，我们构造了一个能够高精度逼近基函数 e_j （假设 6）的 MLP 层。现在，第一注意力层的作用是处理输入混合测度 ν_f，识别与给定查询分量 y_f相关的测度 $_{,\ i}$，然后输出原始查询坐标与相对于该相关测度的 e_j 积分的拼接。形式上，这一操作由下文定义的映射ϕ_2 实现。

现在，我们分析架构中第一注意力层的逼近性质和复杂度。回顾一下，注意力算子 Attn_θ 已经以测度论形式定义为

$$\mathrm{Attn}_\theta : \mathcal{P}(\mathbb{R}^{d_{\mathrm{attn}}}) \times \mathbb{R}^{d_{\mathrm{attn}}} \to \mathbb{R}^{d_{\mathrm{attn}}},$$

其中，第一个参数是在词元表示上的概率测度，第二个参数是查询向量。以下引理表明，在对输入测度和函数作出适当结构假设的条件下：

1. 目标映射 ϕ_2 可以由注意力类$\mathcal{A}(d_{\mathrm{attn}},\ H,\ B_a,\ S_a)$中的一个成员以精度 ϵ_2 实现（引理8）；
2. 此类注意力映射是 Lipschitz 连续的，其上界可用模型参数显式表示（引理 9）；
3. $\mathcal{A}$ 的 ϵ_2-覆盖数在上述参数范围内具有一个上界（引理 10）。

我们依次给出这些结果。

在介绍引理8之前，我们先说明映射 ϕ_2 在映射复合视角下所扮演的角色（参见定义 5）。在我们的构造中，第一个 MLP层 ϕ_1 近似 Mercer特征：

$$\phi_1(\mu, x) \ \simeq \ \begin{bmatrix} x_1 \\ \vdots \\ x_d \\ e_1(x_{d_1+1:d}) \\ \vdots \\ e_D(x_{d_1+1:d}) \end{bmatrix},$$

其中，$\{e_j\}_{j\geq 1}$ 是$\mathcal{X}_0 \subset \mathbb{R}^{d_2}$上的 Mercer（再生核希尔伯特空间）特征基。因此，经过 ϕ_1 后的推前测度为

$$\mu_1 := \big(\phi_1(\mu, \cdot)\big)_\sharp \mu \ \in \ \mathcal{P}(\mathbb{R}^{d+D}),$$

而复合规则给出

$$(\phi_2 \diamond \phi_1)(\mu, x) \ = \ \phi_2\big(\mu_1,\, \phi_1(\mu, x)\big).$$

在当前设定中，ϕ_2 保留前 d 个坐标，并将最后 D 个坐标替换为相关测度对 Mercer 基的积分（逐分量计算）：

$$\phi_2\big(\mu_1,\, \phi_1(\mu, x)\big) \ = \ \begin{bmatrix} x_1 \\ \vdots \\ x_d \\ \displaystyle\int e_1 \,\mathrm{d}\mu_0^{(i^*)} \\ \vdots \\ \displaystyle\int e_D \,\mathrm{d}\mu_0^{(i^*)} \end{bmatrix},$$

其中，i^* 表示查询所选取的相关分量的索引。此外，若 $\mu_0^{(i^*)}$ 相对于参考测度 λ具有（博雷尔）密度，记作 $\frac{\mathrm{d}\mu_0^{(i^*)}}{\mathrm{d}\lambda} = f_{\mu_0^{(i^*)}}$，其中 $f_{\mu_0^{(i^*)}} \in L^2(\lambda)$（参见备注 2），那么，写出 Mercer 展开 $f_{\mu_0^{(i^*)}} = \sum_{j\geq 1} b_j e_j$, 后，忽略无关紧要的常数项，我们有：

$$\int e_j \,\mathrm{d}\mu_0^{(i^*)} \ = \ \int e_j \, f_{\mu_0^{(i^*)}} \,\mathrm{d}\lambda \ = \ \langle e_j, f_{\mu_0^{(i^*)}} \rangle_{L^2(\lambda)} \ = \ b_j.$$

因此，ϕ_2 在其最后 D 个坐标中给出相关密度的（截断）Mercer 系数。

引理8表明，在我们对输入测度和变换f的结构假设下，目标映射 ϕ_2 可由具有有界参数的注意力机制以精度 ϵ_2 一致逼近。

引理8的直观理解。关键在于第一注意力层的结构：对于固定的 j，为了提取关联的Mercer系数 $b_j = \int e_j \, \mathrm{d}\mu^{(i^*)}{}_0$，我们构造QK矩阵 W_{QK}^j，使得在词元由（简化的）第一MLP层映射为$(x_1, x_2) \mapsto \psi_j(x_1, x_2) = (x_1, e_j(x_2))$时，

$$\psi_j(x_\mathrm{q})^\top W_{QK}^j \psi_j(y) = \begin{cases} \gg 1 & \text{if} \quad i = i^*; \\ \leq 0 & \text{if} \quad i \neq i^*, \end{cases} \quad \text{for} \quad y \sim \mu_{v^{(i)}}^{(i)} = \delta_{v^{(i)}} \otimes \mu_0^{(i)}.$$

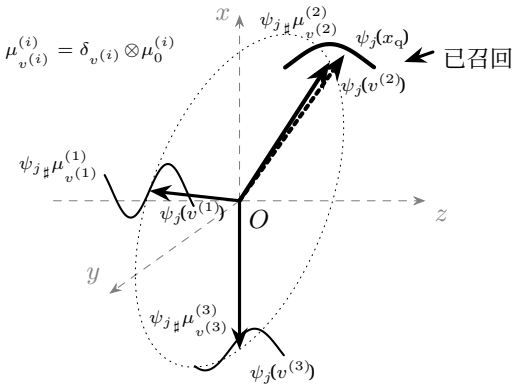

图 3：联想回忆的几何示意图。分量 $\mu_{v^{(i)}}^{(i)}$ 沿特征轴通过第一个 MLP 层 f_1分离；查询映射到 $\psi_j(x_q)$，并与锚点 $\psi_j(v^{(i^*)})$ 对齐：$=\psi_j((v^{(i^*)\top}, \mathbf{0}^\top)^\top)$，从而召回 $\mu_0^{(i^*)}$。推前测度 $\psi_{j\sharp}\mu^{(i^*)}{}_{v^{(i^*)}}$ 提供了$F^\star(\mu^{(i^*)}\, x_q)$所用的特征。

那么，softmax 值为

$$\mathrm{Softmax}(\psi_j(x_\mathrm{q})^\top W_{QK}^j \psi_j(y)) \simeq \mathbb{1}[i = i^*].$$

构造很简单：取 $W_{QK}^j \propto \sum_{k=1}^{d_1} \boldsymbol{e}_k \boldsymbol{e}_k^\top$，使得 $\psi_j(x_\mathrm{q})^\top W_{QK}^j \psi_j(y) \propto_{\langle v^{(i^*)} v^{(i)} \rangle}$，再乘以一个较大的标量，其中 $\boldsymbol{e}_k$ 是第 k个独热向量，因为 ψ 保留了前 d 个坐标：$\psi(x_\mathrm{q}) = [v^{(i^*)\top}; *]^\top$, $\psi(y) = [v^{(i)\top}; *]^\top$，且 $(v^{(i)})_i$ 彼此可区分（即$\langle v^{(i)}, v^{(j)} \rangle \leq 0$ 对于 $i \neq j$），如图 3 所示。接着，经过第一个 MLP 层f_1映射的词元中，第 j个注意力头的第 $(d+j)$个输出为

$$\begin{aligned} & \boldsymbol{e}_{d+1}^\top \underbrace{W}_{\boldsymbol{e}_{d+1}\boldsymbol{e}_{d+1}^\top} \int \underbrace{\mathrm{Softmax}(\psi_j(x_\mathrm{q}^\top) W_{QK}^j \psi_j(y))}_{\simeq \mathbb{1}[i=i^*]} \underbrace{V}_{\boldsymbol{e}_{d+1}\boldsymbol{e}_{d+1}^\top} \psi_j(y) \mathrm{d}\nu(y) \\ \simeq & \int \mathbb{1}[i = i^*] e_j(y_{d_1+1:d}) \mathrm{d} \left(\sum_i \delta_{v^{(i)}} \otimes \mu_0^{(i)} \right)(y) \\ \simeq & \int e_j(y) \mathrm{d}\mu_0^{(i^*)}(y), \end{aligned}$$

其中我们将参数取为 $W = V = \boldsymbol{e}_{d+1}\boldsymbol{e}_{d+1}^\top$。最后，第二个 MLP 层将 $((\int e_j \, d\mu_0^{(i^*)})_{i=1}^D, v^{(i^*)})$ 映射为 y，而第二个注意力则充当跳跃连接。

引理8。令 $\mathcal{P}_f$ 表示如下形式的概率测度集合

$$\nu_f = \frac{1}{I} \sum_{i=1}^I f_\sharp \mu_i = f_\sharp \left(\frac{1}{I} \sum_{i=1}^I \mu_i \right),$$

其中，测度 μ_i 和映射 f 满足以下条件：

- μ_i 是一个支撑于 $\mathbb{R}^d$ 的有界子集上的概率测度，其中 $d = d_1 + d_2$，且具有乘积形式

$$\mu_i = \delta_{v_i} \otimes \tilde{\mu}_i,$$

其中 $v_i \in \mathbb{S}^{d_1} \subset [-B_x, B_x]^{d_1}$ 满足 $\langle v_i$，对所有 $i \neq j$均成立； $\tilde{\mu}_i$ 是一个支撑在 $\mathbb{R}^{d_2}$ 的有界子集上的概率测度。

- f： $\mathbb{R}^d \to \mathbb{R}^{d+D}$ 由下式给出

$$f(x) = \begin{bmatrix} x_1 \\ \vdots \\ x_d \\ \tilde{f}(x_{d_1+1}, \ldots, x_d) \end{bmatrix},$$

其中 $\tilde{f}$： $\mathbb{R}^{d_2} \to \mathbb{R}^D$ 是一个有界函数。

- $I \leq d_1$ *and* $\operatorname{supp}(f_\sharp \mu_i) \subset [-B_y, B_y]^{d+D}$ *for all* i.

此外，对每个 i，定义 $x_i \colon = [v_i^\top, 0^\top]^\top \in \mathbb{S}^{d_1} \times \{0_{d_2}\}$，并令 $y_{f,\ i} \colon = f(x_i)$。
定义映射 ϕ_2： $\mathcal{P}_f \times \mathbb{R}^{d+D} \to \mathbb{R}^{d+D}$ 如下

$$\phi_2(\nu_f, y_{f,i}) \coloneqq \begin{bmatrix} y_{f,i,1} \\ \vdots \\ y_{f,i,d} \\ \int \mathrm{d}\left(\tilde{f}_1\right)_\sharp \tilde{\mu}_i \\ \vdots \\ \int \mathrm{d}\left(\tilde{f}_D\right)_\sharp \tilde{\mu}_i \end{bmatrix},$$

其中 $\tilde{f}_j$ 表示 f的第 j个坐标函数。

那么，存在一个注意力算子 $\hat{}$注意力 $\in \mathcal{A}(d_{\text{attn}}, H, B_a, S_a)$ 使得

$$\sup_{\nu_f \in \mathcal{P}_f,\, y_{f,i} : v_i \in \mathbb{R}^d} \left\| \phi_2(\nu_f, y_{f,i}) - \hat{\mathrm{A}\mathrm{t}\mathrm{t}\mathrm{n}}(\nu_f, y_{f,i}) \right\|_\infty \leq \epsilon_2,$$

where $d_{\text{attn}} = d + D$, $H = D$, $B_a \lesssim \sqrt{\log\left(I \cdot \epsilon_2^{-1}\right)}$, *and* $S_a = d$.

证明。 固定任意的 $h \in \{1, \ldots, H = D\}$。我们首先将注意力权重矩阵指定如下:

$$W^h = V^h = \boldsymbol{e}_{d+h} \boldsymbol{e}_{d+h}^\top,$$

where $d = d_1 + d_2$和 $\boldsymbol{e}_{d+h}$表示 $\mathbb{R}^{d+D}$中的第（$d+h$）个标准基向量。类似地，定义 e

$$Q^h = K^h = c \begin{bmatrix} I_{d_1} & O \\ O & O_{d_2+D} \end{bmatrix},$$

对于一个足够大的常数 $c \gtrsim \sqrt{\log\left(I^3 \cdot \epsilon_2^{-1}\right)} \gtrsim \sqrt{\log\left(I \cdot \epsilon_2^{-1}\right)}$。

查询 x_i 与键 $f(y)$对应的注意力权重为

$$\begin{aligned} &\operatorname{Softmax}\left(\langle Q^h x_i,\, K^h f(y) \rangle\right) \\ &= \frac{\exp\left(c^2 \langle x_{i,1:d_1},\, y_{1:d_1} \rangle\right)}{\int \exp\left(c^2 \langle x_{i,1:d_1},\, y_{1:d_1} \rangle\right) \mathrm{d}\nu_f(y)} \\ &= \frac{\exp\left(c^2 \langle x_{i,1:d_1},\, y_{1:d_1} \rangle\right)}{I^{-1} \sum_{i'=1}^{I} \int \exp\left(c^2 \langle x_{i,1:d_1},\, y_{1:d_1} \rangle\right) \mathrm{d}(\delta_{v_{i'}} \otimes \tilde{\mu}_{i'})(y)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\exp\left(c^2\langle x_{i,1:d_1},\, y_{1:d_1}\rangle\right)}{I^{-1}\sum_{i'=1}^{I}\exp\left(c^2\langle v_{i'},\, y_{1:d_1}\rangle\right)} \\
&= \frac{\exp\left(c^2\right)}{\frac{1}{I}\left(\exp(c^2)+(I-1)\right)}\,\mathbb{1}_{\{y_{1:d_1}=v_i\}} \quad (v_i \perp v_j,\ i\neq j) \\
&= I\cdot\mathbb{1}_{\{y_{1:d_1}=v_i\}} + O(I^{-1}\epsilon_2),
\end{aligned}
$$

其中，指示函数 $\mathbb{1}_{\{y_{1:d_1}=v_i\}}$ 的出现是因为键 $y_{1:d_1}$ 取值于有限集合$\{v_1,\ldots,v_I\}$，且集合中任意两个元素的内积均为非正数。最后一个等式源于选择$c \gtrsim \sqrt{\log\left(I\cdot\epsilon_2^{-1}\right)}$，该选择确保正确的键与其他键之间实现指数级分离。

接下来，应用值投影矩阵和输出投影矩阵，我们得到

$$
\begin{aligned}
&W^h\int \mathrm{Softmax}\big(\langle Q^h x_i,\, K^h y\rangle\big)V^h y\,\mathrm{d}\nu_f(y) \\
&= \frac{1}{I}\sum_{i'=1}^{I} W^h\int \mathrm{Softmax}\big(\langle Q^h x_i,\, K^h f(y)\rangle\big)V^h f(y)\,\mathrm{d}\mu_{i'}(y) \\
&= \frac{1}{I}\sum_{i'=1}^{I}\int\left(I\cdot\mathbb{1}_{\{y_{1:d_1}=v_i\}}+O(I^{-1}\epsilon_2)\right)\boldsymbol{e}_{d+h}\left(\tilde{f}_h(y_{d_1+1:d})+O(\epsilon_1)\right)\mathrm{d}(\delta_{v_{i'}}\otimes\tilde{\mu}_{i'})(y) \\
&\quad \Big(\text{where } \boldsymbol{e}_{d+h}^{\top}\tilde{f}=\tilde{f}_h\Big) \\
&= \frac{1}{I}\sum_{i'=1}^{I}\left(\int\left(I\cdot\mathbb{1}_{\{y_{1:d_1}=v_i\}}+O(I^{-1}\epsilon_2)\right)\mathrm{d}\delta_{v_{i'}}(y_{1:d_1})\right)\left(\int\tilde{f}_h(y_{d_1+1:d})\,\mathrm{d}\tilde{\mu}_{i'}(y_{d_1+1:d})\right)\boldsymbol{e}_{d+h} \\
&= \left(\int\tilde{f}_h(y_{d_1+1:d})\,\mathrm{d}\tilde{\mu}_i(y_{d_1+1:d})\right)\boldsymbol{e}_{d+h}+O(\epsilon_2).
\end{aligned}
$$

最后，为了在前 d个坐标上纳入跳跃连接，令

$$
A=\begin{bmatrix} I_d & O \\ O & O\end{bmatrix}.
$$

将 A 应用于输入向量$y_{f,\ i}$ 得到

$$
Ay_{f,i}=\begin{bmatrix} y_{f,i,1} \\ \vdots \\ y_{f,i,d} \\ 0_D\end{bmatrix}.
$$

将每个注意力头 h 的注意力输出与该跳跃连接相结合，可在 ℓ_∞ 范数下以阶为 $O(\epsilon_2)$的误差复现目标映射 ϕ_2。这确立了所需的逼近性质。

□

注记3（为什么我们需要softmax 注意力？）。我们非正式地说明，在没有正交性的情况下，线性注意力难以实现对密度的独热选择。主要问题在于，上下文向量$(v^{(i)})_i$ 可能具有负相关。例如，我们考虑 $I=2$ 和

$$
v^{(1)}=-v^{(2)}.
$$

如果我们只能使用线性注意力，并且其 QK 矩阵与引理中的相同，

$$
\mathrm{LinAttn}\big(\langle Q^h f(x_i),\, K^h f(y)\rangle\big)\simeq\langle v^{(i)}, v^{(i^*)}\rangle=\begin{cases}1 & \text{if}\quad i=i^*;\\ -1 & \text{if}\quad i\neq i^*.\end{cases}
$$

这意味着，线性注意力很难通过积分 $\int\mathrm{LinAttn}(\langle Q^h x_i,\ \ K^h y\rangle)V^h y\mathrm{d}\nu_f(y)$ 提取仅第 i^* 个测度。关于实证讨论，参见 Han 等人（2024）；Fan 等人（2025）；另见 Kim 等人（2024），其中需要对召回候选项作出宽松稀疏性和正交性等强假设，才能绕过这一难题。

因此，第一注意力层具有足够的表达能力来实现“关联与提取”运算：给定一个混合分布，它可以选出相关分量，并计算下游处理所需的 e_j 积分。接下来，我们将讨论此类算子在测度和查询向量受到扰动时的稳定性。

下面的引理证明了注意力$_\theta$ 关于两个自变量都满足利普希茨性质。该引理的灵感来自 Vuckovic 等人（2020）。这一量化的稳定性对于后续的覆盖数分析至关重要。

注意力算子 注意力$_\theta$通过对内积 $\langle Q^h x,\ K^h y\rangle$应用 softmax，对值进行加权平均。为了界定 $(\mu,\ x)$ 变化时它的变化幅度，我们将 μ和 x的影响分开处理。

对于测度部分，康托罗维奇–鲁宾斯坦对偶将 1-瓦瑟斯坦距离 W_1 （标准的 Wasserstein 度量）表示为对所有 1-利普希茨函数的期望差取上确界，因此我们可以通过 softmax 核的 Lipschitz 常数来控制其变化。

对于查询部分，我们直接界定核对 x 的 Lipschitz 依赖性，并利用输出投影的稀疏性。将两者结合即可得到所述的 Lipschitz 界。

引理 9. 设注意力$_\theta$ $\in$ $\mathcal{A}(d_{\text{attn}}, H, B_a, S_a)$ 为第 *3* 节中定义的注意力算子。假设查询输入满足 $\|x_1\|_\infty, \|x_2\|_\infty \le B_x$，并且对于每个 $i \in \{1,2\}$，每个 $\mathrm{supp}(\mu_i)$ 都满足 $\|y\|_\infty \le B_y$。那么注意力$_\theta$ 关于联合变量(μ, x) 是利普希茨连续的，具体而言：

$$\begin{aligned}&\|\mathrm{Attn}_\theta(\mu_1, x_1) - \mathrm{Attn}_\theta(\mu_2, x_2)\|_\infty \\ &\le H \exp\big(O(S_a^2 B_a^2 B_x B_y)\big) \cdot \big(W_1(\mu_1,\mu_2) + \|x_1 - x_2\|_2\big).\end{aligned}$$

此外，若 $d_{\text{attn}} \lesssim D$, $H \lesssim D$, $B_a \lesssim \sqrt{\text{对数}(I\epsilon_2^{-1})}$，且 $S_a \lesssim d$, $B_x, B_y \lesssim 1$，则 ***Lipschitz*** 常数的上界为 D 指数$\big(O(d^2\text{对数}\big(I\epsilon_2^{-1}\big))\big)$。

证明。 **估计 μ中的差值**。 我们先在固定查询的情况下，估计 μ变量中的差值。根据 **(P-i)**，当 A 每行至多有 s 个非零元素，且每个元素的绝对值不超过 b 时，矩阵稀疏度界 $\|Av\|_\infty \le sb\|v\|_\infty$ 满足：

$$\begin{aligned}&\|\mathrm{Attn}_\theta(\mu_1, x_1) - \mathrm{Attn}_\theta(\mu_2, x_1)\|_\infty \\ \le &\sum_h S_a^2 B_a^2 \left\| \int \frac{y \exp\big(\langle Q^h x, K^h y\rangle\big)}{\int \exp\big(\langle Q^h x_1, K^h z\rangle\big) \mathrm{d}\mu_1(z)} \mathrm{d}\mu_1 - \int \frac{y \exp\big(\langle Q^h x, K^h y\rangle\big)}{\int \exp\big(\langle Q^h x_1, K^h z\rangle\big) \mathrm{d}\mu_2(z)} \mathrm{d}\mu_2 \right\|_\infty\end{aligned}$$

接下来，插入中间项以使分母和分子对齐，得到：

$$\begin{aligned}&\|\mathrm{Attn}_\theta(\mu_1, x_1) - \mathrm{Attn}_\theta(\mu_2, x_1)\|_\infty \\ \le &\sum_h S_a^2 B_a^2 \left\| \int \frac{y \exp\big(\langle Q^h x, K^h y\rangle\big)}{\int \exp\big(\langle Q^h x_1, K^h z\rangle\big) \mathrm{d}\mu_1(z)} \mathrm{d}\mu_1 - \int \frac{y \exp\big(\langle Q^h x, K^h y\rangle\big)}{\int \exp\big(\langle Q^h x_1, K^h z\rangle\big) \mathrm{d}\mu_2(z)} \mathrm{d}\mu_1 \right. \\ &\left. + \int \frac{y \exp\big(\langle Q^h x_1, K^h y\rangle\big)}{\int \exp\big(\langle Q^h x_1, K^h z\rangle\big) \mathrm{d}\mu_2(z)} \mathrm{d}\mu_1 - \int \frac{y \exp\big(\langle Q^h x, K^h y\rangle\big)}{\int \exp\big(\langle Q^h x_1, K^h z\rangle\big) \mathrm{d}\mu_2(z)} \mathrm{d}\mu_2 \right\|_\infty \\ \le &\underbrace{\sum_h S_a^2 B_a^2 \left\| \int \frac{y \exp\big(\langle Q^h x_1, K^h y\rangle\big)}{\int \exp\big(\langle Q^h x_1, K^h z\rangle\big) \mathrm{d}\mu_1(z)} \mathrm{d}\mu_1 - \int \frac{y \exp\big(\langle Q^h x, K^h y\rangle\big)}{\int \exp\big(\langle Q^h x_1, K^h z\rangle\big) \mathrm{d}\mu_2(z)} \mathrm{d}\mu_1 \right\|_\infty}_{(i)} \\ &+ \underbrace{\sum_h S_a^2 B_a^2 \left\| \int \frac{y \exp\big(\langle Q^h x_1, K^h y\rangle\big)}{\int \exp\big(\langle Q^h x_1, K^h z\rangle\big) \mathrm{d}\mu_2(z)} \mathrm{d}\mu_1 - \int \frac{y \exp\big(\langle Q^h x, K^h y\rangle\big)}{\int \exp\big(\langle Q^h x_1, K^h z\rangle\big) \mathrm{d}\mu_2(z)} \mathrm{d}\mu_2 \right\|_\infty}_{(ii)}.\end{aligned}$$

对第 (i) 项进行界定。我们有

$$(i) \le \sum_h S_a^2 B_a^2 \left\| \int y \exp\big(\langle Q^h x_1, K^h y\rangle\big) \mathrm{d}\mu_1 \right\|_\infty$$

$$
\begin{aligned}
&\times \left| \frac{1}{\int \exp\left(\langle Q^h x_1, K^h z\rangle\right) \mathrm{d}\mu_1(z)} - \frac{1}{\int \exp\left(\langle Q^h x_1, K^h z\rangle\right) \mathrm{d}\mu_2(z)} \right| \\
\leq & \sum_h S_a^2 B_a^2 \left\| \int y \exp\left(\langle Q^h x_1, K^h y\rangle\right) \mathrm{d}\mu_1 \right\|_\infty \\
&\times \left(\min\left\{ \int \exp\left(\langle Q^h x_1, K^h z\rangle\right) \mathrm{d}\mu_1(z), \int \exp\left(\langle Q^h x_1, K^h z\rangle\right) \mathrm{d}\mu_2(z) \right\} \right)^{-2} \\
&\times \left| \int \exp\left(\langle Q^h x_1, K^h z\rangle\right) \mathrm{d}\mu_1(z) - \int \exp\left(\langle Q^h x_1, K^h z\rangle\right) \mathrm{d}\mu_2(z) \right| \quad \text{(by (P-iii))} \\
\lesssim & \sum_h S_a^2 B_a^2 B_y \exp\left(3 S_a^2 B_a^2 B_x B_y\right) \left| \int \exp\left(\langle Q^h x_1, K^h z\rangle\right) \mathrm{d}(\mu_1 - \mu_2)(z) \right| \quad \text{(by (P-ii))},
\end{aligned}
$$

其中我们使用了：

(P-i) 矩阵稀疏度界 $\|Av\|_\infty$，$\|Av\|_1 \leq sb\|v\|_\infty$ 当 A 每行至多有 s 个非零项且每个元素的绝对值不超过 b 时；

(P-ii) $|\langle Q^h x_1,\ K^h y\rangle| \leq S_a B_a B_x \cdot S_a B_a B_y$。事实上，由于 $\|Q^h\|_0$，$\|K^h\|_0 \leq S_a$（非零项总数）且 $|Q^h_{jk}|$，$|K^h_{jk}| \leq B_a$，而 $\|x_1\|_\infty \leq B_x$ 且 $\|y\|_\infty \leq B_y$，我们有

$$
\begin{aligned}
&|\langle Q^h x_1, K^h y\rangle| \\
\leq & \|Q^h x_1\|_\infty \|K^h y\|_1 \\
\leq & \Big(\max_j \sum_k |Q^h_{jk}|\,|x_{1,k}| \Big) \sum_{j,k} |K^h_{jk}|\,|y_k| \\
\leq & (S_a B_a B_x)(S_a B_a B_y).
\end{aligned}
$$

这里，界 $\|Q^h x_1\|_\infty \leq S_a B_a B_x$ 成立是因为 $Q^h x_1$ 的每个坐标都是至多 S_a 项之和，且每项的绝对值至多为 $B_a B_x$；同理，$\|K^h y\|_1 \leq \sum_{j,\ k} |K^h_{jk}|\,|y_k| \leq S_a B_a B_y$ 因为矩阵中非零项总数至多为 S_a；

（P-iii）该界

$$
\left| \frac{1}{\alpha_1} - \frac{1}{\alpha_2} \right| \leq A^{-2} |\alpha_1 - \alpha_2|, \quad A < \min(\alpha_1, \alpha_2).
$$

右端有如下界

$$
\begin{aligned}
((i) \lesssim) & \sum_h S_a^2 B_a^2 B_y \exp\left(3 S_a^2 B_a^2 B_x B_y\right) \left| \int \exp\left(\langle Q^h x_1, K^h z\rangle\right) \mathrm{d}(\mu_1 - \mu_2)(z) \right| \\
\lesssim & H S_a^4 B_a^4 B_x B_y \exp\left(4 S_a^2 B_a^2 B_x B_y\right) W_1(\mu_1, \mu_2)
\end{aligned}
$$

利用康托罗维奇–鲁宾斯坦对偶

$$
W_1(\mu_1, \mu_2) = \sup_{\mathrm{Lip}(\phi) \leq 1} \int \phi \,\mathrm{d}(\mu_1 - \mu_2),
$$

并且由于 $y \mapsto \exp\left(\langle Q^h x_1, K^h y\rangle\right)$ 在 $[-B_y, B_y]^{d_{\mathrm{attn}}}$ 上是 $S_a^2 B_a^2 B_x \exp\left(S_a^2 B_a^2 B_x B_y\right)$-利普希茨的，这是因为

$$
\begin{aligned}
&\left| \exp\left(\langle Q^h x_1, K^h y_3\rangle\right) - \exp\left(\langle Q^h x_1, K^h y_4\rangle\right) \right| \\
\leq & \exp\left(S_a^2 B_a^2 B_x B_y\right) \left| \langle Q^h x_1, K^h (y_3 - y_4)\rangle \right| \\
\leq & S_a^2 B_a^2 B_x \exp\left(S_a^2 B_a^2 B_x B_y\right) \|y_3 - y_4\|_2.
\end{aligned}
$$

对于 $y_3, y_4 \in [-B_y, B_y]^{d_{\mathrm{attn}}}$。

对第 (ii) 项进行界定。我们有

$$
(ii) \lesssim \sum_h S_a^2 B_a^2 \left| \frac{1}{\int \exp\left(\langle Q^h x_1, K^h z\rangle\right) \mathrm{d}\mu_2(z)} \right| \left\| \int y \exp\left(\langle Q^h x_1, K^h y\rangle\right) \mathrm{d}(\mu_1 - \mu_2)(y) \right\|_\infty
$$

$$\lesssim \sum_h S_a^2 B_a^2 \exp\big(S_a^2 B_a^2 B_x B_y\big) \max_{i=1,\ldots,d_{\text{attn}}} \left| \int y_i \exp\left(\langle Q^h x_1, K^h y\rangle\right) \mathrm{d}(\mu_1 - \mu_2)\,(y) \right|$$
$$\lesssim H(1 + S_a^2 B_a^2 B_x B_y) S_a^2 B_a^2 \exp\big(2 S_a^2 B_a^2 B_x B_y\big) W_1(\mu_1, \mu_2)$$

利用康托罗维奇–鲁宾斯坦对偶

$$W_1(\mu_1, \mu_2) = \sup_{\mathrm{Lip}(\phi) \le 1} \int \phi \,\mathrm{d}(\mu_1 - \mu_2),$$

并且，由于 $y \mapsto y_i \exp\left(\langle Q^h x_1, K^h y\rangle\right)$ 在 $[-B_y, B_y]^{d_{\text{attn}}}$ 上是 $(1 + S_a^2 B_a^2 B_x B_y) \exp\big(S_a^2 B_a^2 B_x B_y\big)$-利普希茨的。

最后，我们有

$$\begin{aligned}
&\|\mathrm{Attn}_\theta(\mu_1, x_1) - \mathrm{Attn}_\theta(\mu_2, x_1)\|_\infty \\
\lesssim & H S_a^2 B_a^2 \left(S_a^2 B_a^2 B_x B_y \exp\big(4 S_a^2 B_a^2 B_x B_y\big) + (1 + S_a^2 B_a^2 B_x B_y) \exp\big(2 S_a^2 B_a^2 B_x B_y\big) \right) \\
& \times W_1(\mu_1, \mu_2).
\end{aligned}$$

估计 x 中的差异。接下来，我们估计查询 x 中的差异。对于固定的 μ_1，利用 x 中类似的插值和利普希茨估计，

$$\begin{aligned}
&\|\mathrm{Attn}_\theta(\mu_1, x_1) - \mathrm{Attn}_\theta(\mu_1, x_2)\|_\infty \\
\le & \sum_h S_a^2 B_a^2 \left\| \int \frac{y \exp\left(\langle Q^h x_1, K^h y\rangle\right)}{\int \exp\left(\langle Q^h x_1, K^h z\rangle\right) \mathrm{d}\mu_1(z)} - \frac{y \exp\left(\langle Q^h x_2, K^h y\rangle\right)}{\int \exp\left(\langle Q^h x_2, K^h z\rangle\right) \mathrm{d}\mu_1(z)} \mathrm{d}\mu_1(y) \right\|_\infty \\
& + S_a B_a \|x_1 - x_2\|_\infty \\
\le & \sum_h S_a^2 B_a^2 \left| \int \frac{y \exp\left(\langle Q^h x_1, K^h y\rangle\right)}{\int \exp\left(\langle Q^h x_1, K^h z\rangle\right) \mathrm{d}\mu_1(z)} \mathrm{d}\mu_1 - \int \frac{y \exp\left(\langle Q^h x_1, K^h y\rangle\right)}{\int \exp\left(\langle Q^h x_2, K^h z\rangle\right) \mathrm{d}\mu_1(z)} \mathrm{d}\mu_1(y) \right. \\
& \left. + \int \frac{y \exp\left(\langle Q^h x_1, K^h y\rangle\right)}{\int \exp\left(\langle Q^h x_2, K^h z\rangle\right) \mathrm{d}\mu_1(z)} \mathrm{d}\mu_1 - \int \frac{y \exp\left(\langle Q^h x_2, K^h y\rangle\right)}{\int \exp\left(\langle Q^h x_2, K^h z\rangle\right) \mathrm{d}\mu_1(z)} \mathrm{d}\mu_1(y) \right| \\
& + S_a B_a \|x_1 - x_2\|_2 \\
\lesssim & (S_a B_a \vee H S_a^4 B_a^4 B_x B_y \exp\big(4 S_a^2 B_a^2 B_x B_y\big)) \|x_1 - x_2\|_2.
\end{aligned}$$

□

结合逼近与稳定性，我们可以通过注意力类的覆盖数来控制其复杂度，如下所述。

现在，我们在关注的参数范围内估计 $\mathcal{A}(d_{\text{attn}}, H, B_a, S_a)$ 的 ϵ_2-覆盖数。**引理 10**。 ϵ_2 的覆盖数 $\mathcal{A}(d_{\text{attn}}, H, B_a, S_a)$ 有如下上界

$$\begin{aligned}
& \mathcal{N}(\mathcal{A}(d_{\text{attn}}, H, B_a, S_a); \epsilon_2)_\infty \\
\lesssim & \left(d_{\text{attn}}^2 \cdot \exp\left(O(\log(H) + S_a^2 B_a^2 B_x B_y) \right) \left(\epsilon_2^{-1} + 1 \right) \right)^{O(S_a H)}.
\end{aligned}$$

此外，若 $d_{\text{attn}} \lesssim d + D$, $H = O(D)$、 $B_a \lesssim \sqrt{\log(I \epsilon_2^{-1})}$, $S_a \lesssim d$, 且 $B_x, B_y = O(1)$, 则覆盖熵为

$$\log \mathcal{N}(\mathcal{A}(d_{\text{attn}}, H, B_a, S_a), \epsilon_2) \lesssim C_D \cdot D d^3 (\log I + \log \epsilon_2^{-1})^2 \cdot \log \epsilon_2^{-1}$$

其中 $C_{d,D} \lesssim \mathrm{poly}(\log D + \log d)$。

证明。 我们将 $\mathcal{A}(d_{\text{attn}}, H, B_a, S_a)$ 的 $\Omega(\bar{\epsilon})$-覆盖集定义为一组映射，其参数可按如下方式构造：

- For each matrix W^h, Q^h, K^h, V^h in each $h \in \{1, \ldots, H\}$,
  1. 从 $O(d_{\text{attn}}^2)$ 个矩阵条目中选择 S_a 个。

2. 对于每个矩阵条目，–将其值设为 $\{(j\cdot\tilde{\epsilon}-1)\cdot B_a \mid j=0,\ldots,\lceil 2\tilde{\epsilon}^{-1}\rceil\}$ 其中 $\tilde{\epsilon}\gtrsim\exp\left(-C_2(\log(H)+S_a^2B_a^2B_xB_y)\right)\bar{\epsilon}$，其中 C_2 是一个足够大的常数。

•将 A 中所选 S_a 个条目的值设为 $\{(j\cdot\tilde{\epsilon}-1)\cdot B_a \mid j=0,\ldots,\lceil 2\tilde{\epsilon}'\rceil\}$，其中$\tilde{\epsilon}'\simeq(S_aB_a)^{-1}\bar{\epsilon}$。

下面我们证明上述映射集合是一个$\Omega(\bar{\epsilon})$-覆盖。显然，对于 W^h $h=1,\ldots,H$，在 $\bar{\epsilon}$-覆盖集中存在矩阵$\hat{}W^h$，使得

$$\|W^h-\hat{W}^h\|_\infty\lesssim H^{-1}\exp\left(-C_3(S_a^2B_a^2B_xB_y)\right)\bar{\epsilon},\quad \|W^h-\hat{W}^h\|_0\le 2S_a$$

其中 C_3 是一个充分大的常数。对 Q^h，K^h，V^h和 A，类似的不等式同样成立。令$\hat{}\theta=(\hat{}A,(\hat{}W^h,\hat{}Q^h,\hat{}K^h,\hat{}V^h)_h)$。则对所有 $\mu\in\mathcal{P}([-B_y,B_y]^{d_{\text{attn}}})$和 $x\in[-B_x,B_x]^{d_{\text{attn}}}$，

$$\begin{aligned}
&\|\mathrm{Attn}_\theta(\mu,x)-\mathrm{Attn}_{\hat{\theta}}(\mu,x)\|_\infty\\
\le&\sum_h\left\|W^h\int\frac{V^hy\exp\left(\langle Q^hx,K^hy\rangle\right)}{\int\exp\left(\langle Q^hx,K^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu-\hat{W}^h\int\frac{\hat{V}^hy\exp\left(\langle\hat{Q}^hx,\hat{K}^hy\rangle\right)}{\int\exp\left(\langle\hat{Q}^hx,\hat{K}^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu\right\|_\infty\\
&+\|(A-\hat{A})x\|_\infty\\
\lesssim&\sum_h\underbrace{\left\|W^h\int\frac{V^hy\exp\left(\langle Q^hx,K^hy\rangle\right)}{\int\exp\left(\langle Q^hx,K^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu-\hat{W}^h\int\frac{V^hy\exp\left(\langle Q^hx,K^hy\rangle\right)}{\int\exp\left(\langle Q^hx,K^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu\right\|_\infty}_{(i)}\\
&+\sum_h\underbrace{S_aB_a\left\|\int\frac{V^hy\exp\left(\langle Q^hx,K^hy\rangle\right)}{\int\exp\left(\langle Q^hx,K^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu-\int\frac{V^hy\exp\left(\langle\hat{Q}^hx,K^hy\rangle\right)}{\int\exp\left(\langle Q^hx,K^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu\right\|_\infty}_{(ii)}\\
&+\sum_h\underbrace{S_aB_a\left\|\int\frac{V^hy\exp\left(\langle\hat{Q}^hx,K^hy\rangle\right)}{\int\exp\left(\langle Q^hx,K^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu-\int\frac{V^hy\exp\left(\langle\hat{Q}^hx,\hat{K}^hy\rangle\right)}{\int\exp\left(\langle Q^hx,K^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu\right\|_\infty}_{(iii)}\\
&+\sum_h\underbrace{S_aB_a\left\|\int\frac{V^hy\exp\left(\langle\hat{Q}^hx,\hat{K}^hy\rangle\right)}{\int\exp\left(\langle Q^hx,K^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu-\int\frac{V^hy\exp\left(\langle\hat{Q}^hx,\hat{K}^hy\rangle\right)}{\int\exp\left(\langle\hat{Q}^hx,K^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu\right\|_\infty}_{(iv)}\\
&+\sum_h\underbrace{S_aB_a\left\|\int\frac{V^hy\exp\left(\langle\hat{Q}^hx,\hat{K}^hy\rangle\right)}{\int\exp\left(\langle\hat{Q}^hx,K^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu-\int\frac{V^hy\exp\left(\langle\hat{Q}^hx,\hat{K}^hy\rangle\right)}{\int\exp\left(\langle\hat{Q}^hx,\hat{K}^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu\right\|_\infty}_{(v)}\\
&+\sum_h\underbrace{S_aB_a\left\|\int\frac{V^hy\exp\left(\langle\hat{Q}^hx,\hat{K}^hy\rangle\right)}{\int\exp\left(\langle\hat{Q}^hx,\hat{K}^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu-\int\frac{\hat{V}^hy\exp\left(\langle\hat{Q}^hx,\hat{K}^hy\rangle\right)}{\int\exp\left(\langle\hat{Q}^hx,\hat{K}^hz\rangle\right)\mathrm{d}\mu(z)}\mathrm{d}\mu\right\|_\infty}_{(vi)}\\
&+\underbrace{\|(A-\hat{A})x\|_\infty}_{(vii)}\\
\lesssim&\,\bar{\epsilon}.
\end{aligned}$$

每一项的界如下：

(i). 设 $\boldsymbol{w} \coloneqq \int \frac{V^h y}{\exp(\langle Q^h x, K^h y\rangle) \int \exp(\langle Q^h x, K^h z\rangle) \mathrm{d}\mu(z)} \mathrm{d}\mu$. 则，$(i) \leq \|(W^h - \hat{} W^h) \boldsymbol{w}\|_\infty \lesssim (2S_a) \cdot H^{-1} \exp\left(-C_3(S_a^2 B_a^2 B_x B_y)\right) \bar{\epsilon} \cdot \exp\left(O(S_a^2 B_a^2 B_x B_y)\right) \lesssim H^{-1}\bar{\epsilon}$ 由 (P-i,ii)；

(ii). 第二项有如下界

$$
\begin{aligned}
(ii) \lesssim & S_a B_a \left\| \sup_{\boldsymbol{y} \in \mathrm{supp}(\mu)} \left| \frac{V^h y_i}{\int \exp(\langle Q^h x, K^h z\rangle) \mathrm{d}\mu(z)} \right.\right. \\
& \left.\left. \times \left(\exp\left(\langle Q^h x, K^h y\rangle\right) - \exp\left(\langle \hat{Q}^h x, K^h y\rangle\right) \right) \right| \right\|_\infty \\
\lesssim & S_a B_a \exp\left(O(S_a^2 B_a^2 B_x B_y)\right) \sup_{\boldsymbol{y} \in \mathrm{supp}(\mu)} \left| \langle Q^h x, K^h y\rangle - \langle \hat{Q}^h x, K^h y\rangle \right| \\
\lesssim & S_a B_a \exp\left(O(S_a^2 B_a^2 B_x B_y)\right) \|(Q^h - \hat{Q^h})x\|_\infty \sup_{\boldsymbol{y} \in \mathrm{supp}(\mu)} \left| \|K^h y\|_1 \right| \\
\lesssim & \exp\left(O(S_a^2 B_a^2 B_x B_y)\right) \cdot H^{-1} \exp\left(-C_3(S_a^2 B_a^2 B_x B_y)\right) \bar{\epsilon} \\
\lesssim & H^{-1}\bar{\epsilon}.
\end{aligned}
$$

注意，第二个不等式由 (P-ii,P-iv) 推导得出，第四个不等式则由 (P-i) 支持。

(iii). 第三项可以用与 (ii) 相同的方式界定。

(iv). 第四项有如下界

$$
\begin{aligned}
(iv) \lesssim & S_a B_a \left\| \int V^h y \exp\left(\langle \hat{Q}^h x, \hat{K}^h y\rangle\right) \mathrm{d}\mu \right\|_\infty \\
& \times \left| \frac{1}{\int \exp(\langle Q^h x, K^h z\rangle) \mathrm{d}\mu(z)} - \frac{1}{\int \exp\left(\langle \hat{Q}^h x, K^h z\rangle\right) \mathrm{d}\mu(z)} \right| \\
\lesssim & S_a B_a \left\| \int V^h y \exp\left(\langle \hat{Q}^h x, \hat{K}^h y\rangle\right) \mathrm{d}\mu \right\|_\infty \\
& \times \left(\min\left\{ \int \exp\left(\langle Q^h x, K^h z\rangle\right) \mathrm{d}\mu(z)^{-2}, \int \exp\left(\langle \hat{Q}^h x, K^h z\rangle\right) \mathrm{d}\mu(z) \right\} \right)^{-2} \\
& \times \left| \int \exp\left(\langle \hat{Q}^h x, K^h z\rangle\right) - \exp\left(\langle \hat{Q}^h x, K^h z\rangle\right) \mathrm{d}\mu(z) \right| \quad \text{(by (P-iii))} \\
\lesssim & \exp\left(O(S_a^2 B_a^2 B_x B_y)\right) \cdot \sup_{z \in \mathrm{supp}(\mu)} \left| \langle Q^h x, K^h z\rangle - \langle \hat{Q}^h x, K^h z\rangle \right| \\
\lesssim & H^{-1}\bar{\epsilon};
\end{aligned}
$$

(v). 第五项可用与 (iv) 类似的方式界定。

(vi). 第六项的界与 (i) 相同。

(vii). 使用 (P-i) 即可轻松得到其界 $H^{-1}\bar{\epsilon}$。

请注意，(P-i) $\|Ax\|_\infty$，$\|Ax\|_1 \leq sb\|x\|_\infty$ 矩阵 A s $S_a B_a B_x \cdot S_a B_a B_y$ b $Q^h x$ S_a $|\langle Q^h x\ K^h y\rangle| \leq$ 中非零项的数量有界，且每个元素的绝对值有界，（P-ii），因为每个坐标都是至多项之和，且每项均以 $B_a B_x$ 为界；类似地，$K^h y$ 的每个坐标均以 $S_a B_a B_y$ 为界，（P-iii） $1/\alpha_1 - 1/\alpha_2 = (\alpha_1 - \alpha_2)/(\alpha_1 \alpha_2) \leq A^{-2}|\alpha_1 - \alpha_2|$ 当 $A < \alpha_1$，α_2 时，以及（ P-iv ） $y \mapsto \exp(y)$ 和 $y \mapsto y\exp(y)$ 在 $[-B, B]$ 上是 $\exp(O(B))$- 利普希茨。

根据覆盖集的构造规则，覆盖数有界于

$$\mathcal{N}(\mathcal{A}(d_{\mathrm{attn}}, H, B_a, S_a); \bar{\epsilon})_\infty$$

$$\lesssim \left(\binom{d_{\text{attn}}^2}{S_a} \tilde{\epsilon}^{-S_a} \right)^{4H+1}$$
$$\lesssim \left(\binom{d_{\text{attn}}^2}{S_a} \cdot \exp\left(O(S_a \log(H) + S_a^3 B_a^2 B_x B_y) \right) \left(\bar{\epsilon}^{-1} + 1 \right)^{S_a} \right)^{4H+1}.$$

□

综上，这些结果完整刻画了测度理论设定下的第一注意力层：它能够准确实现 ϕ_2，以稳定的方式实现这一点，并且其覆盖数随 D 和 ϵ_2 有利地扩展。

B.5 步骤2-3：第二MLP层

在前面的小节中已经确立：第一MLP层能够逼近Mercer基函数 e_j，且注意力机制能够提取与相关分量测度相关联的对应系数$\int e_j \mathrm{d}\mu^{(i^*)}$，现在我们转向该架构的下一阶段。

在这一步中，模型的输入被有效地约化为Mercer系数的有限集合（b_1，.....，b_D）以及查询向量 x_o。于是，该统计问题被转化为对定义在($D + O(1)$)维定义域上的Lipschitz 函数的逼近。

本节的目标有两方面：第一，确定适当的截断维数 D，以在逼近误差与复杂度之间取得平衡；第二，建立使用神经网络对 D 个变量的利普希茨函数的逼近结果。

B.5.1 确定维数 D

如上所述，经过第一个多层感知机和注意力层后，输入测度 μ_0 的有效表示被简化为相对于核特征基 $\{e_i\}_{i\geq 1}$的 Mercer 系数。然而在实际应用中，只能保留有限个系数。因此，一个关键问题是：截断展开中应保留多少项 D，才能在控制模型统计复杂度的同时，使逼近误差仍可忽略? 以下引理量化了将 μ_0 逼近为其在前 D 个特征函数上的投影时产生的截断误差。

引理 11. 设 $\mu_0 \in \mathcal{H}_0$ 具有 *Mercer* 展开

$$\frac{\mathrm{d}\mu_0}{\mathrm{d}\lambda} = \sum_{i=1}^{\infty} b_i e_i,$$

其中 $\{e_i\}$ 是与核特征值 $\{\lambda_i\}$ 对应的 *Mercer* 特征函数， λ是勒贝格测度。定义截断逼近

$$\tilde{\mu}_0 = \sum_{i=1}^{D} b_i e_i.$$

若 $\gamma_{\mathrm{f}} < 0$ 且 $\gamma_{\mathrm{b}} > 0$，则 γ_{f}范数下的截断误差满足

$$\|\mu_0 - \tilde{\mu}_0\|_{\gamma_{\mathrm{f}}} \;\leq\; \lambda_{D+1}^{\frac{-\gamma_{\mathrm{f}}+\gamma_{\mathrm{b}}}{2}}.$$

证明。 左边有上界，界为

$$\begin{aligned}
&\|\mu_0 - \tilde{\mu}_0\|_{\gamma_{\mathrm{f}}} \\
=&\sqrt{\sum_{i\geq D+1} \lambda_i^{-\gamma_{\mathrm{f}}} b_i^2} \\
=&\sqrt{\lambda_{D+1}^{-\gamma_{\mathrm{f}}+\gamma_{\mathrm{b}}}} \sqrt{\sum_{i\geq D+1} \frac{\lambda_i^{-\gamma_{\mathrm{f}}}}{\lambda_{D+1}^{-\gamma_{\mathrm{f}}+\gamma_{\mathrm{b}}}} b_i^2}
\end{aligned}$$

$$\leq \sqrt{\lambda_{D+1}^{-\gamma_{\mathrm{f}}+\gamma_{\mathrm{b}}}} \sqrt{\sum_{i \geq D+1} \lambda_i^{-\gamma_{\mathrm{b}}} b_i^2} \qquad (\text{by} \quad \lambda_{D+1} \geq \lambda_i,\ i \geq D+1 \quad \text{and} \quad -\gamma_{\mathrm{f}} > 0)$$

$$\leq \sqrt{\lambda_{D+1}^{-\gamma_{\mathrm{f}}+\gamma_{\mathrm{b}}}} \qquad \left(\text{by} \quad \sqrt{\sum_i \lambda_i^{-\gamma_{\mathrm{b}}} b_i^2} \leq 1\right)$$

其中，我们在第一个不等式中对 $i \geq D+1$ 使用了 $\frac{\lambda_i^{-\gamma_{\mathrm{f}}}}{\lambda_{D+1}^{-\gamma_{\mathrm{f}}}} \leq 1$，在最后一个不等式中对 $i \geq D+1$ 使用了 $\lambda_i^{-\gamma_{\mathrm{b}}} \geq \lambda_{D+1}^{-\gamma_{\mathrm{b}}}$，并利用了 μ_0 位于该球内。 □

在控制了 Mercer 展开的截断误差之后，我们接下来转向目标回归函数 $F^\star$ 的正则性。具体而言，假定 $F^\star$ 关于由测度上的 γ_{f} 加权 RKHS 距离与查询变量上的标准欧几里得距离构成的乘积度量是 Lipschitz 连续的。形式上，存在 $L > 0$ 使得

$$\left|F^\star(\mu, x) - F^\star(\nu, y)\right| \ \leq \ L\left(\|\mu - \nu\|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}} + \|x - y\|_2\right), \quad \forall \mu, \nu \in B(\mathcal{H}_0, \|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}}),\ x, y \in \mathbb{R}^{d_1}.$$

这一利普希茨性质确保，一旦无限维测度 μ被其截断的 D 维逼近所替代， $F^\star$ 上由此引起的误差便可直接界定。下面的推论使这一约化变得明确。

推论2. 定义截断回归函数

$$\bar{F}_D(\mu_0, v) \ \coloneqq \ F^\star\left(\sum_{i=1}^{D} b_i e_i,\ x\right), \quad \frac{\mathrm{d}\mu_0}{\mathrm{d}\lambda} = \sum_{i=1}^{\infty} b_i e_i, \quad x = \begin{bmatrix} v \\ 0 \end{bmatrix}.$$

若对 $i \geq D+1$ 有 $b_i = 0$，我们简记为 $F_D(b_1, \ldots, b_D, v)$。则，

$$\sup_{\mu_0 \in B(\mathcal{H}_0, \|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}})} \left|\bar{F}_D(\mu_0, v) - F^\star(\mu_0, v)\right| \ \lesssim \ \lambda_{D+1}^{\frac{-\gamma_{\mathrm{f}}+\gamma_{\mathrm{b}}}{2}}.$$

此外， F_D 关于系数$(b_1, \ldots, b_D)$和查询 v 是利普希茨的，满足

$$\begin{aligned} \left|\bar{F}_D(b, v) - \bar{F}_D(c, v')\right| \ &\leq L\sqrt{\max_{i \geq 1} \lambda_i^{-\gamma_{\mathrm{f}}}}\, \|b - c\|_2 + L\|v - v'\|_2 \\ &\lesssim O(1) \cdot \left\| \begin{bmatrix} b - c \\ v - v' \end{bmatrix} \right\|_2 \end{aligned}$$

其中 $b = (b_1, \ldots, b_D)$ 和 $c = (c_1, \ldots, c_D)$。注意到 $\gamma_{\mathrm{f}} < 0$，因此 $\|b - c\|_2$ 前的乘法因子为 $\Theta(1)$。

B.5.2 有限维利普希茨函数的逼近

将 Mercer 展开截断至 $O(d_1 + D)$ 个系数后，无限维回归问题便简化为逼近一个 Lipschitz 函数 $f\colon\ [0,1]^{O(d_1+D)} \to \mathbb{R}$，其 Lipschitz 常数为 K。现在我们回顾关于用深层 ReLU 网络逼近此类函数的定量结果。

引理 12 (Schmidt-Hieber (2020))。 对于任意函数 $f \in \mathrm{Lip}_L([0,1]^{\bar{D}})$ 以及任意整数 $m \geq 1$和 $N \geq \exp\big(\Omega(\bar{D})\big)$ 。存在一个网络

$$\tilde{f} \in \mathcal{F}(\ell, (\bar{D}, 6(D+1)N, \ldots, 6(D+1)N, 1), s, \infty)$$

具有深度

$$\ell \simeq (m+1)(1 + \log(\bar{D} + 1))$$

以及参数数量

$$s \lesssim (\bar{D}+1)^{3+\bar{D}} N(m+6),$$

使得

$$\|\tilde{f} - f\|_{L^\infty} \lesssim (L+1)(1+\bar{D}^2) 6^{\bar{D}} N 2^{-m} + L N^{-1/\bar{D}}.$$

该引理表明，深度 ReLU 网络能够在 $[0,1]^{O(D)}$ 上逼近任意 Lipschitz 函数，并明确体现网络深度、宽度与逼近误差之间的权衡关系。下一条备注将这一一般性结果与我们的 Mercer–再生核希尔伯特空间设定联系起来。

引理 13 (Schmidt-Hieber (2020))。 设 $V = \prod_{i=0}^{\ell}(p_i+1)$。那么，

$$\log \mathcal{N}(\mathcal{F}(\ell,p,s,F);\delta)_\infty \lesssim (s+1)\log\big(\delta^{-1}\ell V\big).$$

该界表明，一旦架构参数（ℓ， p， s）固定，覆盖熵至多随分辨率 δ^{-1} 对数增长。应用我们的参数选择，可得以下推论。

最后，为便于后续使用，我们回顾一个关于 ReLU 网络的利普希茨常数与其各层宽度之间关系的有用估计。

引理 14(引自 Schmidt-Hieber (2020) 中引理 5 的证明)。 *NN* 的利普希茨常数（相对于无穷范数）为 $\prod_{i=0}^{\ell} p_i$。

Fro根据上述引理，我们针对 Mercer-再生核希尔伯特空间设定得到了专门的逼近结果 ：

推论 3 (对我们设定的特化)。 在设定 *3* 下，并假设 $d_1 \simeq (\ln \epsilon^{-1})^{\beta^{-1}}$， 则

$$\bar{D} \sim \big(\tfrac{-\gamma_{\mathrm{f}}+\gamma_{\mathrm{b}}}{2c}\big)^{-1/\alpha}(\ln \epsilon^{-1})^{1/\alpha} + d_1 \;\sim\; (\ln \epsilon^{-1})^{1/\min(\alpha,\beta)}.$$

令

$$m = O\big(\bar{D}\cdot \log \epsilon_3^{-1}\big), \qquad N = \epsilon_3^{-\bar{D}}$$

根据引理 *12*，存在一个深度为*.....*的 ***ReLU*** 网络

$$\ell_2 \lesssim (\mathrm{poly}\log(d+D)\cdot(d+D)\cdot \log \epsilon_3^{-1},$$

宽度

$$\|\boldsymbol{p}_2\|_\infty \lesssim (d+D)\epsilon_3^{-d+D}$$

且参数数量为

$$s_2 \lesssim \tilde{O}\Big(\epsilon_3^{-O\left((\ln \epsilon^{-1})^{1/\min(\alpha,\beta)}\right)}\cdot((\ln\epsilon)^{\min(\alpha,\beta)})^{O((\ln\epsilon)^{1/\min(\alpha,\beta)})}\cdot\big(\mathrm{poly}\log\big(d+D+\epsilon_3^{-1}\big)\Big)$$

，它能以 $\lesssim \epsilon$ 的上确界范数误差逼近 F_D； F_D 在推论*2*中定义。

此外，相应假设类的覆盖熵满足

$$\log \mathcal{N}\big(\mathcal{F}(\ell,p,s,F);\epsilon_3\big)_\infty \;\lesssim\; \epsilon_3^{-O((\ln\epsilon^{-1})^{1/\min(\alpha,\beta)})}\cdot \mathrm{poly}\log\big(\epsilon^{-1}\epsilon_3^{-1}\big),$$

并且该网络的 ***Lipschitz*** 常数（相对于 ℓ_∞ 范数）满足如下界

$$\prod_{i=0}^{\ell} p_i \;\lesssim\; \epsilon_3^{-O\left(c_\epsilon(\ln\epsilon^{-1})^{\frac{2}{\min(\alpha,\beta)}}\cdot(\ln\epsilon_3^{-1})\right)}, \qquad c_\epsilon \lesssim \mathrm{poly}\log\log\epsilon^{-1}.$$

我们不再假设 $d_1 \lesssim (\log\epsilon^{-1})^{\beta^{-1}}$，而是假设 $F^\star$ 与 x_{q}相互独立。那么，通过添加一层 $\mathbb{R}^{d+D}\ni x \mapsto x_{d+1:d+D} = -\mathrm{ReLU}(-Ax) + \mathrm{ReLU}(Ax) \in \mathbb{R}^D$，其中 $A = [O_{D\times d};\ I_D]$，，即可构造出近似 F_D 且上确界误差为 $\lesssim \epsilon_3$ 的 ***ReLU*** 网络，其深度为

$$\ell_2 \lesssim (\mathrm{poly}\log(D)\cdot(D)\cdot \log \epsilon_3^{-1},$$

宽度

$$\boldsymbol{p}_2 = (d+D, \underbrace{O(D\epsilon_3^{-D}),\ldots,O(D\epsilon_3^{-D})}_{\ell-1\ \textit{times}})$$

以及参数数量

$$s_2 \lesssim \tilde{O}\Big(\epsilon_3^{-O\left((\ln\epsilon^{-1})^{1/\alpha}\right)}\cdot((\ln\epsilon)^{1/\alpha})^{O((\ln\epsilon)^{1/\alpha})}\cdot\big(\mathrm{poly}\log\big(D+\epsilon_3^{-1}\big)+D\Big).$$

覆盖熵有如下界

$$\log \mathcal{N}\big(\mathcal{F}(\ell,p,s,F);\epsilon_3\big)_\infty \lesssim \epsilon_3^{-O((\ln \epsilon^{-1})^{1/\alpha})} \cdot \operatorname{poly}\log\big(\epsilon^{-1} d \epsilon_3^{-1}\big),$$

且 *Lipschitz* 常数为

$$\prod_{i=0}^{\ell} p_i \lesssim d_1 \cdot \epsilon_3^{-O\left(c_\epsilon (\ln \epsilon^{-1})^{\frac{2}{\alpha}} \cdot (\ln \epsilon_3^{-1})\right)}, \qquad c_\epsilon \lesssim \operatorname{poly}\log\log \epsilon^{-1},$$

其中 $D \simeq (\ln \epsilon^{-1})^{\alpha^{-1}}$。

注释 4. Schmidt-Hieber (2020) 的原始引理针对定义在 $[0,1]^{\bar{D}}$ 上的函数。在我们的设定中，定义域为 $[-O(1)\ O(1)]^{\bar{D}}$。简单的重缩放映射将 $[-O(1)\ O(1)]$ 映射到 $[0,1],$，且这一变换仅使 Lipschitz 常数乘以一个固定因子。因此，上述逼近和覆盖结果在普适常数范围内仍然成立。

该推论归纳了在我们设定下参数选择的结果：有效输入维度 $\bar{D}$ 按 $(\ln \epsilon^{-1})^{1/\alpha}$ 的速度增长，网络规模相对于 $1/\epsilon$ 呈次指数级增长，覆盖熵由 $\epsilon^{-O((\ln \epsilon^{-1})^{1/\alpha})}$ 控制；当 $\epsilon_3^{-1} \simeq$ 为 poly log $\epsilon^{-1} \cdot \epsilon^{-1}$ 时，Lipschitz 常数相对于 ϵ^{-1} 至多呈准多项式级增长

B.6 步骤 2-4：第二注意力层

回顾一下，注意力假设类参数化为

$$\mathcal{A}(d_{\text{attn}}, H, B_a, B'_a, S_a, S'_a),$$

其中，d_{attn} 是嵌入维度，H 是头数，B_a, B'_a 是权重矩阵算子范数的界，而 S_a, S'_a 是稀疏性约束。

I在当前步骤中，我们仅实现标量的跳跃连接。我们将情形特化为 e

$$d_{\text{attn}} = 1, \quad H = 1, \quad B_a = 0,\ B'_a = 1, \quad S_a = 0,\ S'_a = 1.$$

也就是说，第二个注意力层属于以下类

$$\mathcal{A}(1,1,0,1,0,1).$$

这一特定选择对应于一个与输入测度无关的退化注意力算子，它仅实现一个在向量上充当恒等映射的跳跃连接，从而确保与整体 Transformer 类的形式定义保持一致。

引理 15. $\mathcal{A}(d+D,\ 1,\ 0,\ 1,\ 0,\ d+D)$ 的 ϵ_4-覆盖数满足

$$\log \mathcal{N}(\mathcal{A}(1,1,0,1,0,1);\delta)_\infty = O\big(\operatorname{poly}\log \epsilon_4^{-1}\big).$$

证明. 略。 □

引理16. 每个注意力算子 注意力 $\in \mathcal{A}(1,\ 1,\ 0,\ 1,\ 0,\ 1)$ 相对于欧几里得范数是 $O(1)$-利普希茨连续的。

证明. 略。 □

B.7 推导估计误差上界

现在，我们将前几个小节中得到的逼近界结合起来，为 Transformer 类架构推导出估计误差保证。令 TF 表示由满足第 3 节所述架构和参数约束的 Transformer 模型构成的假设类。

引理 17 (Transformer逼近)。 在设定 *3* 中，对于所有 $F^\star \in \mathrm{Lip}_L(B(\mathcal{H}_0, \|\|_{\mathcal{H}_0^{\gamma_b}}) \|\|_{\mathcal{H}_0^{\gamma_f}})$，存在 $F \in \mathrm{TF}$，使得对于任意如下形式的输入

$$(\nu, x_{\mathrm{q}}) \quad \textit{where} \quad \nu = \tfrac{1}{I}\sum_{i=1}^{I} \mu^{(i)}_{v^{(i)}}, \quad x_{\mathrm{q}} = \mathrm{Emb}_{v^{(i^*)}}(0_{d_2})$$

其中 $\mu^{(i)}_{v^{(i)}}$ 按照定义*1*生成，

$$|F^\star(\nu, x_{\mathrm{q}}) - \hat{F}(\nu, x_{\mathrm{q}})| \lesssim \epsilon,$$

其中假设集*TF*的参数$(d_j, H_j, B_{a,j}, B'_{a,j}, S_{a,j}, S'_{a,j}, \ell_j, \boldsymbol{p}_j, s_j)_{j=1}^2$ 分别按引理*8*对 d_1, H_1, B_a*,1*, B'_a*,1*, S_a*,1*, S'_a*,1*、推论 *1*对 $\ell_1, \boldsymbol{p}_1, s_1$、引理 *15*对$d_2, H_2, B_a$*,2*, B'_a*,2*, S_a*,2*, S'_a*,2*、推论 *3*对 $\ell_2, \boldsymbol{p}_2, s_2$加以定义。其中的有效维数 D 由推论*2*确定。 $\epsilon_i, i = 1, \ldots,$ *4*的确定则留待引理*18*。

上述引理表明，Transformer假设类足够丰富，能够在可容许输入域上以一致精度 ϵ逼近任意Lipschitz目标函数 $F^\star$。请注意，每一层的输出都是一致有界的（我们可以为每一层添加一个截断ReLU层）。

为了分析该假设类中 ERM（经验风险最小化器）的统计性能，接下来需要给出其覆盖熵的上界。

引理18 (Transformer 的覆盖熵)。 *Transformer*假设类的覆盖熵满足

$$\log \mathcal{N}(\mathrm{TF}; \epsilon)_\infty \lesssim \epsilon^{-O((\ln \epsilon^{-1})^{1/\min(\alpha,\beta)})} = \exp\Big(O(\ln \epsilon^{-1})^{(1+\min(\alpha,\beta))/\min(\alpha,\beta)}\Big))$$

假设 $I \le d_1 \simeq (\ln \epsilon^{-1})^{\beta^{-1}}$ 且 $d_2 \simeq 1$。

证明。 该结论可通过应用用于覆盖数的复合引理（引理 5）得出。具体而言，

$$\begin{aligned}
&\log \mathcal{N}(\mathrm{TF}; \epsilon)_\infty \\
\lesssim &\log \mathcal{N}\left(\mathcal{A}(1,1,0,1,0,1); \epsilon\right)_\infty + \log \mathcal{N}\left(\mathcal{F}(\ell_2, p_2, s_2); \tilde{\Omega}(\epsilon)\right)_\infty \\
&+ \log \mathcal{N}\left(\{\mathrm{Attn}_{\theta_1} \diamond \mathrm{MLP}_{\xi_1}\}; \Omega(L^{-1}_{\mathrm{MLP},2}\epsilon)\right) \\
\lesssim &\log \mathcal{N}\left(\mathcal{A}(1,1,0,1,0,1); \underbrace{\epsilon}_{=:\epsilon_4}\right)_\infty \\
&+ \log \mathcal{N}\left(\mathcal{F}(\ell_2, p_2, s_2); \underbrace{\tilde{\Omega}(\epsilon)}_{=:\epsilon_3}\right)_\infty + \log \mathcal{N}\left(\mathcal{A}(d_{\mathrm{attn}}, H, B_a, H_a); \underbrace{\tilde{\Omega}(L^{-1}_{\mathrm{MLP},2}\epsilon)}_{=:\epsilon_2}\right)_\infty \\
&+ \log \mathcal{N}\left(\mathcal{F}(\ell_1, p_1, s_1); \underbrace{\tilde{\Omega}(L^{-1}_{\mathrm{MLP},2}(L_{\mathrm{Attn},1,W^1} + L_{\mathrm{Attn},1,\|\|_2})^{-1}\epsilon)}_{=:\epsilon_1}\right)_\infty \\
\lesssim &\log \mathcal{N}(\mathcal{A}(d+D,\ 1,\ 0,\ 1,\ 0,\ d+D);\ \epsilon)_\infty \\
&+ \log \mathcal{N}\left(\mathcal{F}(\ell_2, p_2, s_2); \tilde{\Omega}(\epsilon)\right)_\infty + \log \mathcal{N}\left(\mathcal{A}(d_{\mathrm{attn}}, H, B_a, H_a); \exp\Big(-O(c_\epsilon \ln \epsilon^{-1})^{\frac{2+2\min(\alpha,\beta)}{\min(\alpha,\beta)}}\Big)\right)_\infty \\
&+ \log \mathcal{N}\left(\mathcal{F}(\ell_1, p_1, s_1); I^{-1} \cdot \exp\Big(-O(c_\epsilon \ln \epsilon^{-1})^{\frac{2+2\min(\alpha,\beta)}{\min(\alpha,\beta)}}\Big)\right)_\infty,
\end{aligned}$$

其中 $c_\epsilon \lesssim \mathrm{poly}\log\log \epsilon^{-1}$。这里我们用到了

(i) 令 $\epsilon_4 = \epsilon$，第二个注意力层是 $O(1)$-利普希茨的（引理16）；

˜(ii) 令 $\epsilon_3 \gtrsim \Omega(\epsilon)$，第二个 MLP 层的 Lipschitz 常数有如下界$L_{\mathrm{MLP},\,2} \lesssim \exp\big(-O\big(c_\epsilon((\ln \epsilon^{-1})^{2/\min(\alpha,\beta)} \cdot (\ln \epsilon_3^{-1})^2)\big) = \exp\Big(c_\epsilon(\ln \epsilon^{-1})^{2+2\min(\alpha,\beta)\,\min(\alpha,\,\beta)}\Big)$ ($c_\epsilon \lesssim$polylog log ϵ^{-1}) ~~(推论 3)~~；

(iii) 令 $\epsilon_2 \gtrsim \exp\Big(-O(c_\epsilon \ln \epsilon^{-1})^{2+2\min(\alpha,\beta)\,\min(\alpha,\beta)}\Big)$，则第一注意力层的 Lipschitz 常数有界为 $L_{\mathrm{Attn},1,W^1} + L_{\mathrm{Attn},1,\|\|_2} \lesssim D\exp\big(O(d^2\log(I\epsilon_2^{-1}))\big) \lesssim I^{O(1)} \cdot \exp\Big(O(c_\epsilon d^2(\ln \epsilon^{-1})^{2+2\min(\alpha,\beta)\,\min(\alpha,\beta)})\Big)$ ~~(引理 9)~~；

(iv) We have $\epsilon_1 \gtrsim I^{-O(1)} \cdot \exp\Big(-O(c_\epsilon d^2(\ln \epsilon^{-1})^{\frac{2+2\min(\alpha,\beta)}{\min(\alpha,\beta)}})\Big)$ for the first MLP layer.

根据引理 15 和推论 3,

$$\log \mathcal{N}\left(\mathcal{A}(1,1,0,1,0,1);\epsilon_4\right)_\infty \lesssim \mathrm{poly}\log \epsilon^{-1}$$

根据推论 3,

$$\log \mathcal{N}\left(\mathcal{F}(\ell_2,p_2,s_2);\epsilon_3\right)_\infty \lesssim \epsilon^{-O((\ln \epsilon^{-1})^{1/\min(\alpha,\beta)})}.$$

根据引理 10,

$$\log \mathcal{N}\left(\mathcal{A}(d_{\mathrm{attn}},H,B_a,H_a);\epsilon_2\right)_\infty \lesssim \mathrm{poly}(\log \epsilon^{-1} + \log I)\cdot d^3.$$

根据推论 1

$$\log \mathcal{N}\left(\mathcal{F}(\ell_1,p_1,s_1);\epsilon_1\right)_\infty \lesssim \mathrm{poly}\log\log I \cdot \mathrm{poly}\log \epsilon^{-1} \cdot (((\log I) + \mathrm{poly}\log \epsilon^{-1}\cdot d)^{4d_2}\cdot D + d).$$

假设 $I \le d_1 \simeq (\log \epsilon)^{\beta^{-1}}$ and $d_2 = O(1)$，我们有

$$\begin{aligned}&\log \mathcal{N}\left(\mathcal{A}(d_{\mathrm{attn}1},H_1,B_{a,1},H_{a,1});\epsilon_2\right)_\infty + \log \mathcal{N}\left(\mathcal{F}(\ell_1,p_1,s_1);\epsilon_1\right)_\infty \\ &+ \log \mathcal{N}\left(\mathcal{F}(\ell_2,p_2,s_2);\epsilon_3\right)_\infty + \log \mathcal{N}\left(\mathcal{A}(1,1,0,1,0,1);\epsilon_4\right)_\infty \\ \lesssim\ & \epsilon^{-O((\ln \epsilon^{-1})^{1/\min(\alpha,\beta)})}。\end{aligned}$$

□

有了这些要素，我们可以调用 ERM 的一般统计学习界。

ˆ**引理 19** (Schmidt-Hieber (2020))。 考虑高斯回归，并令 F 为假设类 $\mathcal{F} \subset L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})$上的经验风险最小化器。假设对所有 $f \in \mathcal{F}$，均有 $\|f\|_{L^\infty} \le A$。那么，对于任意 $\delta > 0$，若 $V(\delta)$ 表示 $\mathcal{F}$ 的覆盖熵，则有

$$R(\bar{F}^\star,\hat{F}) \lesssim \inf_{f\in\mathcal{F}} \|\hat{F}-\bar{F}^\star\|^2_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})} + \frac{(A^2+\sigma^2)V(\delta)}{n} + (A+\sigma)\delta.$$

现在我们可以给出 Transformer 经验风险最小化所达到的统计收敛率。

B.7.1 亚多项式收敛率

ˆ**定理 5** (亚多项式收敛)。 设 F 为经验风险最小化器，其假设集由引理 *17* 构造。在设定 *3* 中，我们有

$$R(F^\star,\hat{F}) \lesssim \exp\Big(-\Omega((\ln n)^{\frac{\min(\alpha,\beta)}{\min(\alpha,\beta)+1}})\Big)$$

假设混合分量的数量有界为 $I \le d_1 \lesssim (\ln n)^{\beta^{-1}\min(\alpha,\beta)\min(\alpha,\beta)+1}$ 且$d_2 \simeq 1$。 ————

证明。 令 $\epsilon \simeq \exp\Big(-c'(\ln n)^{\min(\alpha,\beta)\,\min(\alpha,\beta)+1}\Big)$，其中常数 $c' > 0$ 充分小。结合引理 17 和 18 以及引理 19，我们得到

$$R(F^\star,\hat{F}) \lesssim \epsilon + \frac{\exp\big(-O((\ln \epsilon^{-1})^{1/\min(\alpha,\beta)+1})\big)}{n}$$

$$\begin{aligned}
&\lesssim \epsilon + \frac{\exp\Big(O(c'^{(\min(\alpha,\beta)^{-1}+1)}(\ln n))\Big)}{n} \\
&\leq \epsilon + \frac{n^{c''}}{n} \quad \text{for some } 0 < c'' < 1, \\
&\lesssim \exp\Big(-\Omega((\ln n)^{\frac{\min(\alpha,\beta)}{1+\min(\alpha,\beta)}})\Big),
\end{aligned}$$

假设混合分量的数量有界为 $I \leq d_1 \lesssim (\ln \epsilon^{-1})^{\beta^{-1}}$ □

注释 5 （定理 5 的解释）。 非参数回归中常见的统计学习界具有以下形式

$$R(\hat{F}, \bar{F}^\star) \lesssim n^{-\Theta(1/d)},$$

其中 d 是问题的（有效）维数。然而，在我们的设定中，特征值衰减假设 $\lambda_j \simeq \exp(-cj^\alpha)$意味着有效维数的增长仅为

$$d \sim (\ln n)^{1/(\alpha+1)}.$$

因此，定理 5 中的界可解释为经典$n^{-\Theta(1/d)}$ 速率的直接类比，只不过将 d 替换为 $(\ln n)^{1/(\alpha+1)}$。重要的是，这表明该估计量绕过了联想回忆任务通常面临的组合困难。在我们的框架中，每个需要召回的元素并非有限符号，而是一个概率测度，即一个无限维对象。尽管存在这种内在复杂度，分析结果表明，统计行为完全由底层核的特征值衰减所支配，因此呈现出无限维回归特有的速率。

B.7.2 超越对数容量

在定理 5 中，我们讨论了分量数量（就联想记忆而言的“容量”）受限于

$$I \leq d_1 \lesssim (\ln n)^{\frac{\beta^{-1}\min(\alpha,\beta)}{\min(\alpha,\beta)+1}},$$

的情形。该数量关于样本量 n 和 Transformer 模型的“参数”数量（我们将其视为覆盖数）均呈对数级增长。这是因为，当 d_1 很大时， $\mathbb{S}^{d_1-1}$ 上的利普希茨函数（而非 $B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0{}^{\gamma_b}})$）会变得过于复杂。另一方面，在 B.4 节中，我们观察到注意力矩阵的实际参数数量关于 $d_1(\geq I)$ 呈线性增长。

这里我们考虑以下附加假设：

假设 9. 目标函数 $\tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}})$ 与 x_{q} 无关，且仅依赖于 $\mu_0^{(i^*)}$。（即我们可以写成 $\tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}}) = \tilde{F}^\star(\mu_0^{(i^*)})$ 。）

那么，即使在具有无限维测度值分量的关联回忆任务中，我们也有多项式“容量”：

引理 20. 假设

$$I \leq d_1 \simeq \exp\Big(O((\log \epsilon^{-1})^{\frac{\alpha+1}{\alpha}})\Big), \quad d_2 = O(1).$$

在设定 *3* 中，且满足附加假设 *9*，

$$\log \mathcal{N}(\mathrm{TF}; \epsilon)_\infty \lesssim \exp\Big(O(\ln \epsilon^{-1})^{(1+\alpha)/\alpha}\Big))$$

证明。 本引理的主要策略沿用引理18。我们仅说明与前一个引理的不同之处。

¯令 $D = d + D \simeq d_1$ （考虑 $D \ll d_1$ 的情形）。

(i) 令 $\epsilon_4 = \epsilon$，第二个注意力层是 $O(1)$-利普希茨的（引理16）；

(ii) 令 $\epsilon_3 \gtrsim \Omega(\epsilon)$，第二个 MLP 层的 Lipschitz 常数有如下界：$L_{\mathrm{MLP},\ 2} \lesssim d_1 \exp\left(-O\left(c_\epsilon D^2 \cdot (\ln \epsilon^{-1})^2\right)\right)$，($c_\epsilon \lesssim$ poly log log ϵ^{-1})（修改推论 3，忽略前 d 个索引）；

(iii) 令 $\epsilon_2 \gtrsim d_1 \exp\left(-O\left(c_\epsilon D^2 \cdot (\ln \epsilon^{-1})^2\right)\right)$，第一层注意力层的 Lipschitz 常数有如下界：$L_{\mathrm{Attn},\ 1,\ W^1} + L_{\mathrm{Attn},\ 1,\ \|\|\|_2} \lesssim D \exp\left(O(d^2 \log(I\epsilon_2^{-1}))\right) \lesssim \exp\Big(O(c_\epsilon d_1^2 (\ln d_1)^{O(1)} (\ln \epsilon^{-1})^{2+2\alpha^{-1}})\Big)$ (引理 9 和 $I \le d_1$);

(iv) We have $\epsilon_1 \gtrsim \exp\Big(-O(c_\epsilon d_1^2 (\ln d_1)^{O(1)} (\ln \epsilon^{-1})^{2+2\alpha^{-1}})\Big)$ for the first MLP layer.

根据引理 15 和推论 3,

$$\log \mathcal{N}\left(\mathcal{A}(1,1,0,1,0,1);\epsilon_4\right)_\infty \lesssim \mathrm{poly}\log \epsilon^{-1}$$

通过修改推论 3，忽略前 d 个索引（对应于查询），

$$\log \mathcal{N}\left(\mathcal{F}(\ell_2,p_2,s_2);\epsilon_3\right)_\infty \lesssim \epsilon^{-O((\ln \epsilon^{-1})^{1/\alpha})}.$$

根据引理 10,

$$\log \mathcal{N}\left(\mathcal{A}(d_{\mathrm{attn}1},H_1,B_{a,1},H_{a,1});\epsilon_2\right)_\infty \lesssim \mathrm{poly}(\log\left(\epsilon^{-1}d_1\right)) \cdot d^3.$$

根据推论 1

$\log \mathcal{N}(\mathcal{F}(\ell_1,\ p_1,\ s_1);\ \epsilon_1)_\infty \lesssim$多项式对数对数 I·多项式对数 ϵ^{-1}·(((对数 I)+多项式对数 $\epsilon^{-1}\cdot d)^{4d_2}\cdot D+d$)。
假设 $I \le d_1 \simeq \exp\Big(O((\text{对数 }\epsilon)^{\frac{\alpha+1}{\alpha}})\Big)$ 且 $d_2 = O(1)$，我们有

$$\begin{aligned}&\log \mathcal{N}\left(\mathcal{A}(d_{\mathrm{attn}1},H_1,B_{a,1},H_{a,1});\epsilon_2\right)_\infty + \log \mathcal{N}\left(\mathcal{F}(\ell_1,p_1,s_1);\epsilon_1\right)_\infty \\ &+ \log \mathcal{N}\left(\mathcal{F}(\ell_2,p_2,s_2);\epsilon_3\right)_\infty + \log \mathcal{N}\left(\mathcal{A}(1,1,0,1,0,1);\epsilon_4\right)_\infty \\ \lesssim &\epsilon^{-O((\ln \epsilon^{-1})^{1/\alpha})}.\end{aligned}$$

□

同理，我们得到与定理 5 类似的结果：

定理 6。 令 F 为经验风险最小化器，其假设集由引理 *17* 构造。假设

$$I \le d_1 \simeq \exp(o(\log n)) = n^{o(1)}$$

且目标函数 $F^\star(\mu_0^{(i^*)}, x_{\mathrm{q}})$ 独立于 x_{q}(假设 *9*)。在设定 *3* 中，我们有

$$R(F^\star, \hat{F}) \lesssim \exp\left(-\Omega((\ln n)^{\frac{\alpha}{\alpha+1}})\right).$$

证明。 证明与定理 5 中相同。 □

C 极小极大下界

C.1 证明概要

本节的目标是为设定 4 中的联想回忆问题建立信息论极小极大下界。我们的证明策略包含两个主要步骤：归约为纯无限维回归任务，以及推导相应利普希茨泛函的覆盖/打包熵界。

步骤 1：从无限维回归的归约（C.2 节）。 我们首先将测度上的回归问题归约为联想回忆问题。设

$$\mathcal{F}^\star = \mathrm{Lip}_L(B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0^{\gamma_{\mathrm{b}}}}) \times \mathcal{X}_{\mathrm{q}}, d_{\mathrm{prod}}), \qquad \bar{\mathcal{F}}^\star = \mathrm{Lip}_L(B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0^{\gamma_{\mathrm{b}}}}), \|\cdot\|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}}),$$

其中，$\mathcal{F}^\star$ 是依赖于 (μ, x) 的 Lipschitz 函数全类，而 $\mathcal{F}^\star \subset \bar{\mathcal{F}}^\star$ 是仅依赖于 μ的子类。

这里的关键观察是，每个 ν 都可以写成推前测度的平均值

$$\nu \;=\; \frac{1}{I}\sum_{i=1}^{I}\mu_{v^{(i)}}^{(i)} \;=\; \frac{1}{I}\sum_{i=1}^{I}(\mathrm{Emb}_{v^{(i)}})_\sharp\mu_0^{(i)}.$$

因此，使用“噪声”输入 ν 估计 $F(\mu_0^{(i^*)},\ x)$，至少与使用“纯净”输入 $\mu_0^{(i^*)}$ 进行估计一样困难。

具体而言，这两种观测模型的区别如下：

$$\mathcal{S}_n=\{(\nu_t,(x_{\mathrm{q}})_t,y_t)\}_{t=1}^n,\quad y_t=\tilde{F}^\star(\mu_0^{(i^*)},(x_{\mathrm{q}})_t)+\xi_t,\quad \tilde{F}^\star\in\mathcal{F}^\star,$$

$$\mathcal{U}_n=\{(\mu_0^{(i^*)},y_t)\}_{t=1}^n,\quad y_t=\tilde{F}^\star(\mu_0^{(i^*)})+\xi_t,\quad \tilde{F}^\star\in\bar{\mathcal{F}}^\star\subset\mathcal{F}^\star.$$

˜也就是说，在 $\mathcal{S}_n$ 下，我们观测到一般 Lipschitz 函数 $F^\star\in\mathcal{F}^\star$ 的输出；而在 $\mathcal{U}_n$˜下，输出则被限制在子类 $\mathcal{F}^\star$ 中。

因此，注意到 $F^\star(\nu,\ x):=\tilde{F}^\star(\mu_0^{(i^*)},\ x)$，极小极大风险满足

$$\inf_{\hat{F}}\sup_{\tilde{F}^\star\in\mathcal{F}^\star}\mathbb{E}_{\mathcal{S}_n}\big[\|\hat{F}(\nu,x)-F^\star(\nu,x)\|^2\big]\;\geq\;\inf_{\hat{F}}\sup_{\tilde{F}^\star\in\bar{\mathcal{F}}^\star}\mathbb{E}_{\mathcal{U}_n}\big[\|\hat{F}(\mu_0)-\tilde{F}^\star(\mu_0)\|^2\big].$$

换言之，数据集为 $\mathcal{S}_n$、假设类为 $\mathcal{F}^\star$ 的联想回忆问题，至少与数据集为 $\mathcal{U}_n$、受限类为 $\mathcal{F}^\star$ 的约化回归问题一样困难。这一约化使我们能够聚焦于无限维回归设定。

步骤 2：利普希茨泛函的熵上界。极小极大下界基于 Yang & Barron（1999）给出的通用高斯回归极小极大下界：简而言之，

$$\log\mathcal{M}(\mathcal{F}^\star;\epsilon)_{L^2(\mathbb{P}_{\mu_0})}\;\simeq\;n\epsilon^2\;\;\Rightarrow\;\;\inf_{\hat{F}}\sup_{F^\star}R(F^\star,\hat{F})\;\gtrsim\;\epsilon^2,$$

其中 $\mathcal{M}(\mathcal{F}^\star;\epsilon)_{L^2(\mathbb{P}_{\mu_0})}$ 表示函数类 $\mathcal{F}^\star$ 关于$L^2(\mathbb{P}_{\mu_0})$ 度量的 ϵ-打包数。因此，要得到下界，只需计算集合 $G^\star$ 中利普希茨泛函关于 $L^2(\mathbb{P}_{\mu_0})$ 的打包熵（即度量熵 $\log\mathcal{M}(\simeq\log\mathcal{N})$）。为应用经典的信息论结果，我们推导相关利普希茨泛函类的覆盖/打包熵上界和下界。

步骤 2-1：上界（第 C.3 节）。已知（参见 Boissard（2011）），对于利普希茨类，

$$\log\mathcal{N}(\mathrm{Lip}(A,d);\epsilon)_{L^\infty}\;\lesssim\;\mathcal{N}(A;\epsilon)_d\cdot(\mathrm{poly}\log\epsilon^{-1}),$$

其中 A 是输入集，d 是底层度量。应用这一原理，我们证明 th at

$$\log\mathcal{N}\big(B(\mathcal{H}_0,\|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}});\epsilon\big)_{\|\cdot\|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}}}\;\lesssim\;\exp\big(c_u\log\epsilon^{-1}\big)^{\frac{\alpha+1}{\alpha}},\quad c_u<\infty.$$

该证明依赖于以下等距变换：对于 $f=\sum_{i=1}^\infty b_ie_i$ 以及参数a,b,c,

$$f\;\mapsto\;\phi_{a,b,c}(f):=\sum_{i=1}^\infty\lambda_i^{\frac{c-b}{2}}b_ie_i,$$

这是从$(B(\mathcal{H}_0,\|\cdot\|_{\mathcal{H}_0^a})\,\|\cdot\|_{\mathcal{H}_0^b})$到$(B(\mathcal{H}_0,\|\cdot\|_{\mathcal{H}_0^{a-b+c}})\ \|\cdot\|_{\mathcal{H}_0^c})$的等距双射。我们将此应用于 $a=\gamma_{\mathrm{b}}$、$b=\gamma_{\mathrm{f}}$和 $c=0$。然后，我们采用覆盖赋予 ℓ^2 度量的无限维椭球的标准论证。

步骤2-2：下界（C.4节）。关键难点在于，Lipschitz 常数是各向异性的：沿低索引方向（较小的 j）的差异会受到很强的惩罚，而具有较大 j 的方向则实际上更加平滑。形式化地，对于我们函数类中的 F，有

$$|F(\sum_j b_je_j)-F(\sum_j c_je_j)|\;\leq\;L\sqrt{\sum_j\lambda_j^{-\gamma_{\mathrm{f}}}(b_j-c_j)^2},\quad\gamma_{\mathrm{f}}<0,$$

这清楚地表明，具有较大特征值 λ_j （较小 j）的方向对 Lipschitz 界的贡献远大于具有较小特征值（较大 j）的方向。换句话说，函数类的几何结构在不同坐标之间存在严重扭曲。

为明确展示这一结构，我们构造了一个重缩放映射，将标准立方体 $[0,1]^d$嵌入我们的测度输入空间。根据特征值衰减 $\{\lambda_j\}$对每个坐标进行重新缩放后， $[0,1]^d$ 上的 Lipschitz 函数会变为 $B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}})$ 上的函数，其 Lipschitz 常数正比于

$$\lambda_d^{(-\gamma_{\mathrm{f}}+\gamma_{\mathrm{d}})/2}.$$

这表明，在缩放因子的作用下，我们的函数类包含了一个嵌入的 d维 Lipschitz 球副本。

因此，在分辨率 $R\epsilon/\lambda_d^{(-\gamma_{\mathrm{f}}+\gamma_{\mathrm{d}})/2}$ 下，我们函数类的打包熵至少与 $\mathrm{Lip}_1([0,1]^d, \|\cdot\|_{\ell^\infty})$ 的打包熵一样大。将这一重标度论证与 $[0,1]^d$ 上 Lipschitz 函数已知的装填下界以及 Yang–Barron 信息论不等式相结合，我们得到了联想回忆问题所需的极小极大下界。

备注 6。 在设定 4 中，我们假设密度 $\mathrm{d}\mu_0/\mathrm{d}\lambda$ 是非负的，这总可以通过添加足够大的常数平移来保证。随后，我们放宽了 μ_0必须是概率测度这一额外约束，转而仅要求 $\mathrm{d}\mu_0/\mathrm{d}\lambda$ 属于环境函数空间中的一个有界球。由于问题的本质困难源于定义域的无限维性，而归一化约束仅对应于一个有限维 限制，因此这一修改不会影响该问题的极小极大难度。

C.2 步骤 1：从无限维回归中进行归约

根据引理 21，我们将证明，联想回忆问题至少与输入变量为测度的高斯回归问题同样困难。回顾一下，标准高斯回归问题定义如下：我们观测到独立同分布的随机变量 $(X_t,\ Y_t)_{t=1}^n$，满足

$$Y_t = F^\star(X_t) + \xi_t, \quad t = 1, \dots, n,$$

其中 ξ_t 是与 X_t独立同分布的高斯噪声。另一方面，在我们的回忆与预测问题中，观测到的输入 ν 带有噪声：$(\mu^{(i)}_{v^{(i)}})_{i\neq i^*}$ 混入了 ν，但它们与输出 y无关。我们将证明，在“消除”输入中的噪声后，可以得到更好的估计量。形式化地，我们得到以下推论：

推论 4。设

$$\mathcal{F}^\star := \mathrm{Lip}_L(B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}}) \times \mathcal{X}_{\mathrm{q}}, d_{\mathrm{prod}}), \qquad \bar{\mathcal{F}}^\star := \mathrm{Lip}_L(B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}}), \|\cdot\|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}}).$$

这里， $\mathcal{F}^\star$ 表示在 (μ, x) 中均为 L-利普希茨函数的函数类，而 $\mathcal{F}^\star$ 表示仅依赖于 μ 的 L-利普希茨函数子类。注意，$\mathcal{F}^\star \subset \mathcal{F}^\star$。
考虑数据集

$$\mathcal{S}_n = \{(\nu_t, (x_{\mathrm{q}})_t, y_t)\}_{t=1}^n, \quad y_t = \tilde{F}^\star(\mu_0^{(i^*)}, (x_{\mathrm{q}})_t) + \xi_t, \quad \tilde{F}^\star \in \mathcal{F}^\star,$$

$$\mathcal{U}_n = \{(\mu_0^{(i^*)}, y_t)\}_{t=1}^n, \quad y_t = \tilde{F}^\star(\mu_0^{(i^*)}) + \xi_t, \quad \tilde{F}^\star \in \bar{\mathcal{F}}^\star \subset \mathcal{F}^\star.$$

其采样方式如设定 *4* 所示。记住 $F^\star(\nu, x_{\mathrm{q}}) = \tilde{F}^\star(\mu_0^{(i^*)},\ x_{\mathrm{q}})$，我们有

$$\begin{aligned}
&\inf_{\mathcal{S}_n \mapsto \hat{F} \in L^2(\mathbb{P}_{\nu, x_{\mathrm{q}}})} \sup_{\tilde{F}^\star \in \mathcal{F}^\star} \mathbb{E}_{\mathcal{S}_n}\big[\|\hat{F}(\nu, x) - F^\star(\nu, x)\|^2_{L^2(\mathbb{P}_{\nu, x_{\mathrm{q}}})}\big] \\
&\geq \inf_{\mathcal{U}_n \mapsto \hat{F} \in L^2(\mathbb{P}_{\mu_0})} \sup_{\tilde{F}^\star \in \bar{\mathcal{F}}^\star} \mathbb{E}_{\mathcal{U}_n}\big[\|\hat{F}(\mu_0) - \tilde{F}^\star(\mu_0)\|^2_{L^2(\mathbb{P}_{\mu_0})}\big].
\end{aligned}$$

换言之：具有依赖查询的目标函数的估计问题，至少与目标函数仅依赖于 μ 的受限问题一样困难。因此，只需为后者建立下界即可。

为证明推论 4，我们陈述以下引理。

引理 21. 设 $\mathcal{S}_n = \{(\nu_t, x_t, y_t)\}_{t=1}^n$ 和 $\bar{\mathcal{S}}_n = \{(\mu_0^{(i^*)}, x_t, y_t)\}_{t=1}^n$ 是按照 $\hat{}$ $\hat{}$ $\tilde{}$ $\bar{}$ $\tilde{}$ 定义 *1* 采样得到的数据集。那么，对于任意估计量 $F\colon \mathcal{S}_n \mapsto F$，都存在一个估计量 $F_1\colon \mathcal{S}_n \mapsto F_1$，使得

$$\mathbb{E}_{\mathcal{S}_n}\big[\|\hat{F}(\nu,x) - F^\star(\nu,x)\|^2_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})}\big] \;\geq\; \mathbb{E}_{\bar{\mathcal{S}}_n}\big[\|\tilde{F}_1(\mu_0,x) - \tilde{F}^\star(\mu_0,x)\|^2_{L^2(\mathbb{P}_{\mu_0,x_{\mathrm{q}}})}\big].$$

$\tilde{}$ $\tilde{}$ 此外，如果 $F^\star$ 与查询 x 相互独立，即 $F^\star(\mu,x) = F_2^\star(\mu)$，那么存在一个估计量 $\tilde{F}_2$，它仅依赖于 $\{(\mu_0^{(i^*)}, y_t)\}_{t=1}^n$，使得

$$\mathbb{E}_{\mathcal{S}_n}\big[\|\hat{F}(\nu,x) - \tilde{F}_2^\star(\mu_0)\|^2_{L^2}\big] \;\geq\; \mathbb{E}_{\{(\mu_0^{(i^*)},y_t)\}_t}\big[\|\tilde{F}_2(\mu_0) - \tilde{F}_2^\star(\mu_0)\|^2_{L^2}\big].$$

证明。 请记住，

$$\nu = \frac{1}{I}\sum_i \mu^{(i)}_{v^{(i)}} = \frac{1}{I}\sum_i (\mathrm{Emb}_{v^{(i)}})_\sharp \mu_0^{(i)}, \quad \mu_0^{(i)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\mu_0}, \quad (v^{(i)})_{i=1}^I \sim \mathbb{P}_v.$$

我们希望从原始估计量中消除 $\mu_0^{(i)}$ 和 $v^{(i)}$ 对 $i \in [1\colon I]\backslash\{i^*\}$ 的依赖，并得到更好的估计。我们将构造一个贝叶斯估计映射

$$\bar{\mathcal{S}}_n \longmapsto \{(\mu_0^{(i^*)}, x_{\mathrm{q}}) \mapsto \mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n}(\nu,x_{\mathrm{q}}) \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] =: \tilde{F}_1(\mu_0^{(i^*)}, x_{\mathrm{q}})\} \in L^2(\mathbb{P}_{\mu_0,x_{\mathrm{q}}}),$$

其中 $\mathcal{T}_n :=$ $\{(\mu_0^{(i)})_t,\ (v^{(i)})_t$ $\mid$ 对应于 $\forall i \in [1\colon I]\setminus\{i^*\}$ 和 $\forall t \in [1,n]\}$，而 $\bar{\mathcal{S}}_n :=$ $((\mu_0^{(i^*)})_t,\ (x_{\mathrm{q}})_t, y_t)_{t=1}^n$。这个估计量 $\bar{\mathcal{S}}_n \mapsto \tilde{F}_1$ 定义良好，因为它将 $\bar{\mathcal{S}}_n$ 映射到 $L^2(\mathbb{P}_{\mu_0,x_{\mathrm{q}}})$ 中的一个函数，这是因为：(i) 我们可以根据 $\bar{\mathcal{S}}_n$ 确定性地构造 $\mathcal{S}_n$，并且 $\mathcal{T}_n$ 是因为 $\nu = \frac{1}{I}\sum_i \mu^{(i)}_{v^{(i)}} = \frac{1}{I}\sum_i (\mathrm{Emb}_{v^{(i)}})_\sharp \mu_0^{(i)}$；(ii) 给定 $(\mu_0^{(i^*)}, x_{\mathrm{q}})$ 后，ν 仅具有噪声 $\{(\mu_0^{(i)}, v^{(i)})\}_{i\neq i^*}$ 带来的随机性。注意，上述估计量并未使用用于采样输入/输出对的神谕 $\tilde{}$。简言之，在标准高斯回归的上下文中，F_1 并未作弊：为了对 $\nu|\mu_0^{(i^*)}$ 和 $\mathcal{T}_n$ 取期望，我们额外观察到的只有输入 ν 的噪声，而这些噪声在标准高斯回归中并不存在。从现在起，我们将明确写出 $F = F_{\mathcal{S}_n}$。 损失的下界为

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_n}[\|\hat{F}_{\mathcal{S}_n}(\nu,x_{\mathrm{q}}) - F^\star(\nu,x_{\mathrm{q}})\|^2_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})}] \\
=&\mathbb{E}_{\mathcal{S}_n}[\|(\hat{F}_{\mathcal{S}_n}(\nu,x_{\mathrm{q}}) - \mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n]) \\
&\quad + (\mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] - \tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}}))\|^2_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})}] \\
=&\mathbb{E}_{\mathcal{S}_n}[\|\hat{F}_{\mathcal{S}_n}(\nu,x_{\mathrm{q}}) - \mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n]\|^2_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})}] \quad \dots(\mathrm{i}) \\
&\quad + 2\mathbb{E}_{\mathcal{S}_n}[\langle \hat{F}_{\mathcal{S}_n}(\nu,x_{\mathrm{q}}) - \mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n], \\
&\quad \mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] - \tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}})\rangle_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})}] \quad \dots(\mathrm{ii}) \\
&\quad + \mathbb{E}_{\mathcal{S}_n}[\|(\mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] - \tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}}))\|^2_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})}] \\
\geq&\mathbb{E}_{\mathcal{S}_n}[\|(\tilde{F}_{1,\bar{\mathcal{S}}_n}(\mu_0,x_{\mathrm{q}}) - \tilde{F}^\star(\mu_0,x_{\mathrm{q}}))\|^2_{L^2(\mathbb{P}_{\mu_0,x_{\mathrm{q}}})}].
\end{aligned}$$

对于项 (i)，它大于零。至于 (ii)，我们有

$$\begin{aligned}
(ii) =&\mathbb{E}_{\mathcal{S}_n}[\langle \hat{F}_{\mathcal{S}_n}(\nu,x_{\mathrm{q}}) - \mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n], \\
&\mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] - \tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}})\rangle_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})}] \\
=&\mathbb{E}_{\bar{\mathcal{S}}_n}[\mathbb{E}_{\mathcal{T}_n}[\langle \hat{F}_{\mathcal{S}_n}(\nu,x_{\mathrm{q}}) - \mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n], \\
&\mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] - \tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}})\rangle_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})} \mid \bar{\mathcal{S}}_n]] \\
=&\mathbb{E}_{\bar{\mathcal{S}}_n}[\langle \mathbb{E}_{\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n}(\nu,x_{\mathrm{q}}) \mid \nu, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] - \mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n], \\
&\mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] - \tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}})\rangle_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})}]
\end{aligned}$$

$$
\begin{aligned}
&(\mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] \text{ and } \tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}}) \text{ are independent of } \mathcal{T}_n \mid \bar{\mathcal{S}}_n \text{ if } \bar{\mathcal{S}}_n \text{ is given})\\
=&\mathbb{E}_{\bar{\mathcal{S}}_n}[\int \Big\{ \Big(\int (\mathbb{E}_{\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n}(\nu, x_{\mathrm{q}}) \mid \nu, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] - \mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n]) \mathrm{d}\mathbb{P}_\nu(\nu) \Big)\\
&\times \big(\mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] - \tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}}) \big) \Big\} \mathrm{d}\mathbb{P}_{x_{\mathrm{q}}}(x_{\mathrm{q}})]\\
=&\mathbb{E}_{\bar{\mathcal{S}}_n}[\int \Big\{ \Big(\int (\mathbb{E}_{\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n}(\nu, x_{\mathrm{q}}) \mid \nu, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] - \mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n]) \mathrm{d}\mathbb{P}_{\nu|\mu_0^{(i^*)}}(\nu) \mathrm{d}\mathbb{P}_{\mu_0^{(i^*)}}(\mu_0^{(i^*)}) \Big)\\
&\times \big(\mathbb{E}_{\nu,\mathcal{T}_n}[\hat{F}_{\mathcal{S}_n} \mid \mu_0^{(i^*)}, x_{\mathrm{q}}, \bar{\mathcal{S}}_n] - \tilde{F}^\star(\mu_0^{(i^*)}, x_{\mathrm{q}}) \big) \Big\} \mathrm{d}\mathbb{P}_{x_{\mathrm{q}}}(x_{\mathrm{q}})]\\
=&0.
\end{aligned}
$$

同理，若 $F^\star$ 与 x_{q} 相互独立，则可以省略对 x_{q} 的依赖。我们给出的估计映射为

$$\mu_0 \mapsto \mathbb{E}_{x_{\mathrm{q}},((x_{\mathrm{q}})_t)_{t=1}^n}[\tilde{F}_{1,\mathcal{S}_n} \mid \mu_0, ((\mu_0^{(i^*)})_t, y_t)_t]$$

只能通过观测 $((\mu_0^{(i^*)})_t,\ y_t)_t$ 构造。第二个陈述的细节从略。

□

C.3 步骤 2-1：熵的上界

得益于推论 4，下界分析可归结为一个高斯回归问题，其中输入 $\mu^{(i^*)}{}_0 \sim \mathbb{P}_{\mu_0}$ 根据设定 4 生成。关键步骤是控制底层函数类的覆盖数/填充数，以便应用 Yang & Barron (1999) 的一般极小极大界。

引理 22 (Yang & Barron (1999))**.** 设 $\mathcal{F}^\star := \mathrm{Lip}_L(B(\mathcal{H}_0, \|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}}) \| \cdot \|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}})$。考虑根据设定 4 生成的数据集 $\mathcal{U}_n = \{(\mu_0^{(i^*)}, y_t)\}_{t=1}^n$。假设存在 $\delta, \epsilon > 0$ 使得

$$\log_2 \mathcal{N}(\mathcal{F}^\star; \epsilon)_{L^2} \le \frac{n\epsilon^2}{2\sigma^2}, \qquad \log_2 \mathcal{M}(\mathcal{F}^\star; \delta)_{L^2} \ge \frac{2n\epsilon^2}{\sigma^2} + 2\log 2.$$

Then

$$\inf_{\mathcal{U}_n \mapsto \hat{F}} \sup_{F^\star \in \tilde{\mathcal{F}}^\star} \mathbb{E}_{\mathcal{U}_n}\left[\|\hat{F} - F^\star\|_{L^2(\mathbb{P}_{\mu_0})}^2 \right] \gtrsim \delta^2.$$

注 7. 在下界分析中，我们在定义覆盖数和填充数时采用 L^2 范数。

本小节的目标。为了应用引理 22，我们需要精确控制 Lipschitz 函数类的覆盖熵。具体而言，我们旨在建立一个上界

$$\log \mathcal{N}\Big(\mathrm{Lip}_1(B(\mathcal{H}_0, \|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}}), \| \cdot \|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}}); \epsilon \Big)_{L^2(\mathbb{P}_{\mu_0})}.$$

下界将在 C.4 节中给出。

根据 Boissard（2011）的命题 B.2，(1-)Lipschitz 类的覆盖熵可以由其输入集的覆盖熵控制：

$$\log \mathcal{N}(\mathrm{Lip}(A, d); \epsilon)_{L^\infty} \lesssim \mathcal{N}(A; \epsilon)_d \cdot \mathrm{polylog}(\epsilon^{-1}),$$

其中，A 是输入域（在某些较弱的假设下），d 是底层度量。因此，我们的任务归结为求带有度量 $\| \cdot \|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}}$ 的输入集 $A = B(\mathcal{H}_0, \|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}})$ 的熵的上界。

为此，我们建立加权希尔伯特空间中各球之间的等距对应，从而可以转而使用标准的 L^2 度量。

引理 23. 设 a、b、$c \in \mathbb{R}$。一个映射 $\phi_{a,b,c}$:

$$f = \sum_{i=1}^\infty b_i e_i \mapsto \phi(f) = \sum_{i=1}^\infty \lambda_i^{\frac{c-b}{2}} b_i e_i$$

是从 $(B(\mathcal{H}_0, \|\|_{\mathcal{H}_0^a}) \|\|_{\mathcal{H}_0^b})$ 到 $(B(\mathcal{H}_0, \|\|_{\mathcal{H}_0^{a-b+c}}) \|\|_{\mathcal{H}_0^c})$的等距双射。

证明。 首先，对所有 $f \in B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0})$，

$$\|\phi(f)\|_{\mathcal{H}_0^c} = \sum_i \lambda_i^{-c}\left(\lambda_i^{\frac{c-b}{2}} b_i\right)^2 = \sum_i \lambda_i^{-c+c-b} b_i^2 = \|f\|_{\mathcal{H}_0^b}.$$

这表明 ϕ 是等距映射且为单射。接下来， ϕ也是满射，因为

$$\begin{aligned} & f = \sum_{i=1}^{\infty} b_i e_i \in B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0^a}) \\ \Leftrightarrow & \sum_i \lambda_i^{-a} b_i^2 \le 1 \\ \Leftrightarrow & \sum_i \lambda_i^{-a+b-c} (\lambda_i^{\frac{c-b}{2}} b_i)^2 \le 1 \\ \Leftrightarrow & \phi(f) = \sum_{i=1}^{\infty} \lambda_i^{\frac{c-b}{2}} b_i e_i \in B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0^{a-b+c}}) \end{aligned}$$

□

利用上述等距映射与特征值衰减性质，我们得到输入集的熵的如下上界：

引理 24. $B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0^{\gamma_b}})$ 在赋予距离$\|\|\|_{\gamma_f}$ 时，其覆盖熵的上界为

$$\log \mathcal{N}(B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0{}^{\gamma_b}}); \epsilon)_{\|\|\|_{\mathcal{H}_0^{\gamma_f}}} \lesssim \ln\left(\epsilon^{-1}\right)^{\frac{\alpha+1}{\alpha}}.$$

证明。 根据引理 23，令 $a = \gamma_b$, $b = \gamma_f$, $c = 0$，只需证明

$$\log \mathcal{N}(B_\phi; \epsilon)_{L^2} \lesssim \ln\left(\epsilon^{-1}\right)^{\frac{\alpha+1}{\alpha}}$$

其中$B_\phi = B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0^{\gamma_b - \gamma_f}})$。我们构造一个$J_\epsilon$维集合 $\mathcal{E}_{J_\epsilon} = \{(b_1, \dots, b_{J_\epsilon}) \mid f = \sum_i b_i e_i \in B_\phi\}$，使得对所有$f = \sum_i b_i e_i \in B_\phi$，

$$\sum_{j \ge J_\epsilon + 1} b_i^2 < \frac{1}{4}\epsilon^2.$$

当 $J_\epsilon \simeq (c^{-1}(\gamma_b - \gamma_f) \ln \epsilon^{-1})^{1/\alpha}$ 时即可满足，因为 $\sum_{j \ge J_\epsilon + 1} b_i^2 \le \sum_{j \ge J_\epsilon + 1} \lambda_j \lesssim \lambda_{J_\epsilon}$ 以指数衰减，并使用 $\lambda_j \simeq \exp(-cj^\alpha)$。为了在 $\mathcal{E}_{J_\epsilon}$ 上构造一个 $\frac{1}{2}\epsilon$-覆盖，每个维度最多需要 $O(1 \vee (\lambda_j^2 / \epsilon))$ 个模式，因此覆盖熵有如下界：

$$\begin{aligned} \log \mathcal{N}(\epsilon) &\lesssim \sum_{j=1}^{J_\epsilon} \log \frac{\lambda_j^{\frac{\gamma_b - \gamma_f}{2}}}{\epsilon} \\ &\lesssim \sum_{j=1}^{J_\epsilon} (-\left(\frac{\gamma_b - \gamma_f}{c}\right) j^\alpha + \left(\frac{\gamma_b - \gamma_f}{c}\right)^{-1} J_\epsilon^\alpha) \\ &\lesssim \left(\frac{\gamma_b - \gamma_f}{c}\right)^{-1} J_\epsilon^{\alpha+1} \\ &\simeq \left(\frac{\gamma_b - \gamma_f}{c}\right)^{-\alpha^{-1}} (\ln \epsilon^{-1})^{(\alpha+1)/\alpha} \end{aligned}$$

□

最后，Boissard（2011）的结果和引理 24 得出了利普希茨类所需的熵界。

引理 25 (基于 Boissard（2011）**)。** $\mathrm{Lip}_1(B(\mathcal{H}_0,\|\|\|_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}));\|\|\|_{\mathcal{H}_0^{\gamma_\mathrm{f}}})$ 关于 *Bochner* $L^2(\mathbb{P}_{\mu_0})$-范数的度量熵满足

$$\log\mathcal{N}(\mathrm{Lip}_1(B(\mathcal{H}_0,\|\|\|_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}),\|\|\|_{\mathcal{H}_0^{\gamma_\mathrm{f}}});\epsilon)_{\|\|\|_{L^2(\mathbb{P}_{\mu_0})}}\lesssim\exp\left(c_u(\ln\epsilon^{-1})^{\frac{\alpha+1}{\alpha}}\right)$$

对某个 $c_u>0$。

证明. 由引理 24，$\log\mathcal{N}(B(\mathcal{H}_0,\|\|\|_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}});\epsilon)_{\|\|\|_{\mathcal{H}_0^{\gamma_\mathrm{f}}}}\lesssim\ln(\epsilon^{-1})^{\frac{\alpha+1}{\alpha}}$。然后，由 Boissard（2011）中的命题 B.2 以及 $\epsilon^{-1}=\exp(\ln\epsilon^{-1})=o(\exp\left((\ln\epsilon^{-1})^{\frac{\alpha+1}{\alpha}}\right))$（对任意 $\alpha>0$），我们有

$$\log\mathcal{N}(\mathrm{Lip}_1(B(\mathcal{H}_0,\|\|\|_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}),\|\|\|_{\mathcal{H}_0^{\gamma_\mathrm{f}}});\epsilon)_{\|\|\|_\infty}\lesssim\exp\left(c_u(\ln\epsilon^{-1})^{\frac{\alpha+1}{\alpha}}\right).$$

由不等式 $\sqrt{\mathbb{P}_{\mu_0}(B(\mathcal{H}_0,\|\|\|_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}))}\|\cdot\|_\infty\geq\|\cdot\|_{L^2(\mathbb{P}_{\mu_0})}$ 以及 $\sqrt{\mathbb{P}_{\mu_0}(B(\mathcal{H}_0,\|\|\|_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}))}=1$，我们得到所需的界。□

C.4 步骤2-2：熵的下界

为了应用引理 22 中的一般极小极大界，我们将给出无限维利普希茨类的下界。

主要困难在于 Lipschitz 常数的各向异性：对于我们类中的 F，有

$$|F(\sum_j b_je_j)-F(\sum_j c_je_j)|\ \leq\ L\sqrt{\sum_j\lambda_j^{-\gamma_\mathrm{f}}(b_j-c_j)^2},\qquad -\gamma_\mathrm{f}>0,$$

这表明该泛函在对应于高指标系数 e_j的方向上显著更光滑。为刻画这一效应，我们构造如下嵌入：

$$\iota_d:L^p([0,1]^d)\hookrightarrow L^p(\mu_{\mathcal{H}_0}),\qquad(\iota_dg)(f_\mu):=g\left(\Phi_1\left(\tfrac{f_{\mu,1}}{\lambda_1^{\gamma_\mathrm{d}/2}}\right),\dots,\Phi_d\left(\tfrac{f_{\mu,d}}{\lambda_d^{\gamma_\mathrm{d}/2}}\right)\right),$$

其中

$$f_\mu=\sum_{j=1}^\infty\lambda_j^{\gamma_\mathrm{d}/2}Z_je_j,\quad\mathrm{d}\mu(x)=f_\mu(x)\,\mathrm{d}x,$$

具有独立系数 $Z_j\ \sim\ \rho_j$ 以及累积分布函数 $\Phi_j(z)=\int_{-\infty}^z\rho_j(u)\mathrm{d}u$。

我们证明

$$\iota_d\big(\mathrm{Lip}_1([0,1]^d)\big)\ \subset\ \mathrm{Lip}_{R/\lambda_d^{(-\gamma_\mathrm{f}+\gamma_\mathrm{d})/2}}\big(B(\mathcal{H}_0,\|\|\|_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}),\|\|\|_{\mathcal{H}_0^{\gamma_\mathrm{f}}}\big),$$

从而得到一个装填下界

$$\log\mathcal{M}\big(\mathrm{Lip}_1(B(\mathcal{H}_0,\|\|\|_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}),\|\|\|_{\mathcal{H}_0^{\gamma_\mathrm{f}}});\epsilon\big)_{L^p}\ \geq\ \log\mathcal{M}\left(\mathrm{Lip}_1([0,1]^d,\|\cdot\|_{\ell^\infty}),\tfrac{R\epsilon}{\lambda_d^{(-\gamma_\mathrm{f}+\gamma_\mathrm{d})/2}}\right)_{L^p}.$$

将这一重标度论证与标准的利普希茨装填下界以及 Yang–Barron 的信息论结果相结合，即可得到关联召回问题所需的极小极大下界。

首先，我们构造一个嵌入 ι_d。这一构造表明，经过适当的坐标重标度后，$[0,1]^d$ 上的函数可以等距嵌入我们的测度输入空间。下一个引理将这一嵌入形式化，并量化 Lipschitz 常数如何重新缩放。

引理 26（Lanthaler (2024) 的推广）。设 $\mathbb{P}_{\mu_0}$ 是设定 *4* 中 $B(\mathcal{H}_0,\|\|\|_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}})$ 上的一个概率测度。对于任意 $p\in[1,\infty)$ 和 $d\in\mathbb{N}$，存在一个等距嵌入

$$\iota_d:L^p([0,1]^d)\hookrightarrow L^p(\mathbb{P}_{\mu_0}),$$

使得 $\iota_d(\mathrm{Lip}_1([0,1]^d;\ \|\|\|_\infty))\ \subset\mathrm{Lip}_{R/(\lambda_d^{(-\gamma_\mathrm{f}+\gamma_\mathrm{d})/2})}(B(\mathcal{H}_0,\|\|\|_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}})\ \|\|\|_{\mathcal{H}_0^{\gamma_\mathrm{f}}}))$，其中 $[0,1]^d$ 上的 *Lipschitz* 范数是相对于 $[0,1]^d$ 上的 ∞- 范数定义的。

证明。 根据假设 8， $\mu \sim \mathbb{P}_{\mu_0}$ 的采样方式为

$$f_\mu = \sum_j \lambda_j^{\frac{\gamma_{\mathrm{d}}}{2}} Z_j e_j, \quad \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} = f_\mu,$$

其中 $Z_j, j \geq 1$ 相互独立，且 $Z_j \sim \rho_j(z)\mathrm{d}z$。我们定义累积分布 $\Phi_j(z) = \int_{-\infty}^{z} \rho_j \mathrm{d}z$。我们知道 F_j 是利普希茨的，其Lipschitz 常数以$\sup_j \|\rho_j\|_\infty \leq R$为上界。我们定义$f_{\mu,i} = \langle f_\mu, e_i \rangle$。我们将证明映射

$$\iota_d : L^p([0,1]^d) \hookrightarrow L^p(\mathbb{P}_{\mu_0}), \quad (\iota_d g)(f_\mu) = g(\Phi_1(f_{\mu,1}/\lambda_1^{\frac{\gamma_{\mathrm{d}}}{2}}), \ldots, \Phi_d(f_{\mu,d}/\lambda_d^{\frac{\gamma_{\mathrm{d}}}{2}}))$$

是我们所需要的等距嵌入。对于 $g \in L^p([0,1]^d)$， $\iota_d g$ 的 $L^p(\mathbb{P}_{\mu_0})$-范数等于

$$\begin{aligned}
\mathbb{E}_{f_\mu} |(\iota_d g)(f_\mu)|^p &= \mathbb{E}_{f_\mu} |g(\Phi_1(f_{\mu,1}/\lambda_1^{\frac{\gamma_{\mathrm{d}}}{2}}), \ldots, \Phi_d(f_{\mu,d}/\lambda_d^{\frac{\gamma_{\mathrm{d}}}{2}}))|^p \\
&= \mathbb{E}_{f_\mu} |g(\Phi_1(Z_1), \ldots, \Phi_d(Z_d))|^p \\
&= \int_{[0,1]^d} |g(x_1, \ldots, x_d)|^p \mathrm{d}x \quad (\Phi_j(Z_j) \sim \mathrm{Unif}[0,1]) \\
&= \|g\|^p_{L^p([0,1]^d)},
\end{aligned}$$

这表明 ι_d 是等距嵌入。接下来，我们计算 ι_d的像。映射

$$h_d : (B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}}), \|\|\|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}}) \to ([0,1]^d, \|\|\|_\infty), \quad f_\mu \mapsto (\Phi_1(f_{\mu,1}/\lambda_1^{\frac{\gamma_{\mathrm{d}}}{2}}), \ldots, \Phi_d(f_{\mu,d}/\lambda_d^{\frac{\gamma_{\mathrm{d}}}{2}}))$$

是利普希茨的，因为

$$\begin{aligned}
&\|h_d(f_{\mu_1}) - h_d(f_{\mu_2})\|_\infty \\
\leq &\max_j \left\{ \left| \Phi_j(f_{\mu_1,j}/\lambda_j^{\frac{\gamma_{\mathrm{d}}}{2}}) - \Phi_j(f_{\mu_2,j}/\lambda_j^{\frac{\gamma_{\mathrm{d}}}{2}}) \right| \right\} \\
\leq &\max_j \left\{ \mathrm{Lip}(\Phi_j) \lambda_j^{\frac{-\gamma_{\mathrm{d}}}{2}} |f_{\mu_1,j} - f_{\mu_2,j}| \right\} \\
\leq &\max_j \left\{ \mathrm{Lip}(\Phi_j) \lambda_j^{\frac{\gamma_{\mathrm{f}}-\gamma_{\mathrm{d}}}{2}} \cdot \lambda_j^{\frac{-\gamma_{\mathrm{f}}}{2}} |f_{\mu_1,j} - f_{\mu_2,j}| \right\} \\
\leq &\max_j \left\{ \mathrm{Lip}(\Phi_j) \lambda_j^{\frac{\gamma_{\mathrm{f}}-\gamma_{\mathrm{d}}}{2}} \right\} \sum_j (\lambda_j^{-\frac{\gamma_{\mathrm{f}}}{2}} |f_{\mu_1,j} - f_{\mu_2,j}|)
\end{aligned}$$

因此

$$\mathrm{Lip}(h_d) \leq \frac{R}{\lambda_d^{\frac{-\gamma_{\mathrm{f}}+\gamma_{\mathrm{d}}}{2}}}.$$

因此， $\forall g \in \mathrm{Lip}_1([0,1]^d, \infty)$ 中 $\iota_d g$ 的 Lipschitz 常数有如下界：

$$\mathrm{Lip}(\iota_d g) = \mathrm{Lip}(g \circ h_d) \leq \mathrm{Lip}(g) \cdot \mathrm{Lip}(h_d) \leq \frac{R}{\lambda_d^{\frac{-\gamma_{\mathrm{f}}+\gamma_{\mathrm{d}}}{2}}}.$$

此外，我们还有$\|\iota_d g\|_{C(B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0^{\gamma_{\mathrm{b}}}}))} \leq 1$。 □

引理 26 确保， $[0,1]^d$ 上的 Lipschitz 类可以视为我们的无限维 Lipschitz 类的子类，至多相差一个依赖于 λ_d的缩放因子。这立即根据 $\mathrm{Lip}_1([0,1]^d)$ 中广为研究的填充数，给出了我们类的填充数下界。

推论 5。 在引理 *26* 的假设下，我们有

$$\log \mathcal{M}(\mathrm{Lip}_1(B(\mathcal{H}_0, \|\|\|_{\mathcal{H}_0{}^{\gamma_{\mathrm{b}}}}), \|\|\|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}}); \epsilon)_{L^p(\mathbb{P}_{\mu_0})} \gtrsim \log \mathcal{M} \left(\mathrm{Lip}_1([0,1]^d, \|\|\|_\infty), \frac{R\epsilon}{\lambda_d^{\frac{-\gamma_{\mathrm{f}}+\gamma_{\mathrm{d}}}{2}}} \right)_{L^p([0,1]^d)}$$

证明。 通过对函数进行重新缩放，我们有

$$\begin{aligned}
&\mathcal{M}(\mathrm{Lip}_1(B(\mathcal{H}_0, |||_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}), |||_{\mathcal{H}_0^{\gamma_\mathrm{f}}}); \epsilon)_{L^p} \\
=&\mathcal{M}\left(\mathrm{Lip}_{R/(\lambda_d^{(-\gamma_\mathrm{f}+\gamma_\mathrm{d})/2})}(B(\mathcal{H}_0, |||_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}), |||_{\mathcal{H}_0^{\gamma_\mathrm{f}}}); \frac{R\epsilon}{(\lambda_d^{(-\gamma_\mathrm{f}+\gamma_\mathrm{d})/2})}\right)_{L^p} \\
\gtrsim&\mathcal{M}\left(\iota_d(\mathrm{Lip}_1([0,1]^d; \ell^\infty)); \frac{R\epsilon}{(\lambda_d^{(-\gamma_\mathrm{f}+\gamma_\mathrm{d})/2})}\right)_{L^p} \\
\gtrsim&\mathcal{M}\left(\mathrm{Lip}_1([0,1]^d; \ell^\infty); \frac{R\epsilon}{(\lambda_d^{(-\gamma_\mathrm{f}+\gamma_\mathrm{d})/2})}\right)_{L^p([0,1]^d)}.
\end{aligned}$$

□

因此，估计无限维类的打包熵问题，可归结为估计立方体上有限维 Lipschitz 函数的熵。幸运的是，后者的精确下界已在文献中给出。注意，当 $\epsilon \to 0$ 时，打包与覆盖几乎等价。

引理 27 (Lanthaler (2024))。 对于 $p \in [1, \infty)$ 和 $d \in \mathbb{N}$，存在一个与 d 无关的常数 $c > 0$，使得

$$\log \mathcal{M}(\mathrm{Lip}_1([0,1]^d, |||_\infty); \epsilon)_{|||_{L^p}} \gtrsim \left(\frac{c}{d\epsilon}\right)^d, \quad \forall \epsilon \in (0, c/d].$$

将嵌入论证与上述有限维下界相结合，我们得到以下结果，该结果给出了所需的熵增长指数下界。

引理 28 (与 Lanthaler (2024) 平行)。 设 $\mathbb{P}_{\mu_0}$ 是设定 *4* 中 $B(\mathcal{H}_0, |||_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}})$ 上的概率测度。关于 *Bochner*$L^p(\mathbb{P}_{\mu_0})$-范数，$\mathrm{Lip}_1(B(\mathcal{H}_0, |||_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}) \ |||_{\mathcal{H}_0^{\gamma_\mathrm{f}}})$ 的打包熵满足

$$\log \mathcal{M}(\mathrm{Lip}_1(B(\mathcal{H}_0, |||_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}), |||_{\mathcal{H}_0^{\gamma_\mathrm{f}}}), \epsilon)_{L^2(\mathbb{P}_{\mu_0})} \gtrsim \exp\left(c_l (\ln \epsilon^{-1})^{\frac{\alpha+1}{\alpha}}\right)$$

对于某个常数 $c_l > 0$。

证明。 结合推论 5 和引理 27,

$$\log \mathcal{M}(\mathrm{Lip}_1(B(\mathcal{H}_0, |||_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}), |||_{\mathcal{H}_0^{\gamma_\mathrm{f}}}), \epsilon)_{L^p} \gtrsim \left(\frac{c_1 \lambda_d^{\frac{-\gamma_\mathrm{f}+\gamma_\mathrm{d}}{2}}}{8 d_\epsilon \epsilon}\right)^d$$

其中 $c_1 > 0$ 是一个常数，条件是 $\epsilon \leq \frac{c_1 \lambda_d{}^{\frac{-\gamma_\mathrm{f}+\gamma_\mathrm{d}}{2}}}{d}$ 然后，根据 $\lambda_d \gtrsim \exp(-cd^\alpha)$，右侧有下界

$$(\log \mathcal{M} \gtrsim) \left(\frac{c_1 \exp(-c' d^\alpha))}{d\epsilon}\right)^d \quad \text{if } \epsilon \leq c_2 d^{-1} \exp(-c' d^\alpha),$$

其中 $c' = \frac{c}{(-\gamma_\mathrm{f}+\gamma_\mathrm{d})\ 2}$，且 $c_2 > 0$ 是另一个常数。假设 ϵ 足够小，令 d 为

$$d = \left(\frac{\log(c_2 \epsilon^{-1})}{c'+1}\right)^{1/\alpha} (\simeq (\ln \epsilon^{-1})^{1/\alpha}).$$

通过重新排列上述不等式，我们还可得到

$$\epsilon = c_2 \exp(-(c'+1) d^\alpha),$$

它渐近地满足 $\epsilon \leq c_2 d^{-1} \exp(-c' d^\alpha)$，因为 $\exp(-d^\alpha) \ll d^{-1}$ 其中 $d \simeq (\ln \epsilon^{-1})^{1/\alpha} \to \infty$ 当 $\epsilon \to 0$ 时。于是，我们有

$$\log \mathcal{M}(\mathrm{Lip}_1(B(\mathcal{H}_0, |||_{\mathcal{H}_0{}^{\gamma_\mathrm{b}}}), |||_{\mathcal{H}_0^{\gamma_\mathrm{f}}}), \epsilon)_{L^p}$$

$$
\begin{aligned}
&\gtrsim \left(\frac{c_1 \exp(-c' d^\alpha))}{d\epsilon}\right)^d \\
&\gtrsim \left(\exp\left(-c' d^\alpha + (c'+1)d^\alpha - \ln d + O(1)\right)\right)^d \\
&\gtrsim \exp\big(\Omega(d^{\alpha+1})\big) \\
&\gtrsim \exp\Big(\Omega((\ln \epsilon^{-1})^{\frac{\alpha+1}{\alpha}})\Big),
\end{aligned}
$$

其中第四个不等式用到了 $d^\alpha \simeq \ln \epsilon^{-1}$。 □

C.5 极小极大下界的证明

定理 7 (极小极大下界)。 在设定 *4* 下，我们有

$$
\inf_{\mathcal{S}_n \mapsto \hat{F}} \sup_{\bar{F}^\star \in \mathcal{F}^\star} \mathbb{E}_{\mathcal{S}_n}\Big[\|\hat{F} - F^\star\|^2_{L^2(\mathbb{P}_{\nu,x_{\mathrm{q}}})}\Big] \gtrsim \exp\big(-O\big((\ln n)^{\frac{\alpha}{\alpha+1}}\big)\big).
$$

证明。 根据引理 21，只需计算

$$
\inf_{\mathcal{U}_n \mapsto \hat{F} \in L^2(\mathbb{P}_{\mu_0})} \sup_{F^\star \in \tilde{\mathcal{F}}^\star} \mathbb{E}_{\mathcal{U}_n}\big[\|\hat{F}(\mu_0) - F^\star(\mu_0)\|^2_{L^2(\mathbb{P}_{\mu_0})}\big]
$$

其中 $\mathcal{U}_n = \{(\mu_0^{(i^*)}, y_t)\}_{t=1}^n$ 按照设定 4 进行采样，以替代原始问题。令 $V(\epsilon_n) := \log \mathcal{N}(\mathrm{Lip}_1(B(\mathcal{H}_0, \|\|_{\mathcal{H}_0^{\gamma_{\mathrm{b}}}}) \ \|\|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}}); \epsilon)_{L^2}$ 和 $M(\delta_n) := \log \mathcal{M}(\mathrm{Lip}_1(B(\mathcal{H}_0, \|\|_{\mathcal{H}_0^{\gamma_{\mathrm{b}}}}) \ \|\|_{\mathcal{H}_0^{\gamma_{\mathrm{f}}}}); \epsilon)_{L^2}$。

首先，令 $\epsilon_n = C_{1,\sigma} \exp\big(-c_{\mathrm{E}} (\ln n)^{\frac{\alpha}{\alpha+1}}\big)$，其中 $c_{\mathrm{E}} = \left(\frac{1}{2c_u^2}\right)^{\frac{\alpha}{\alpha+1}}$。然后，根据引理 25，

$$
\begin{aligned}
\frac{V(\epsilon_n)}{\epsilon_n^2} &\lesssim \exp\left(c_u (\log \epsilon_n^{-1})^{\frac{\alpha+1}{\alpha}} + 2c_{\mathrm{E}} (\ln n)^{\frac{\alpha}{\alpha+1}}\right) \\
&\lesssim \exp\left(c_u (c_{\mathrm{E}} (\ln n)^{\frac{\alpha}{\alpha+1}})^{\frac{\alpha+1}{\alpha}} + 2c_{\mathrm{E}} (\ln n)^{\frac{\alpha}{\alpha+1}}\right) \\
&\lesssim \exp\left(\frac{1}{2}\ln n + O((\ln n)^{\frac{\alpha}{\alpha+1}})\right) \\
&\lesssim \frac{n}{\sigma^2}.
\end{aligned}
$$

注意到 $\sigma = \Theta(1)$。接下来，我们给出打包熵的下界。取 $\delta_n = C_\sigma \exp\big(-c_{\mathrm{d}} (\ln n)^{\frac{\alpha}{\alpha+1}}\big)$，其中 $c_{\mathrm{d}} = c_l^{-\frac{\alpha}{\alpha+1}}$ 且 $C_\sigma \gtrsim \frac{2}{\sigma^2} + \frac{2}{\log 2 \, 1}$ 是一个仅依赖于 $\sigma = \Theta(1)$ 的充分大常数，根据引理 28，

$$
\frac{M(\delta_n)}{\epsilon_n^2} \gtrsim \exp\left(c_l (c_{\mathrm{d}}^{\frac{\alpha+1}{\alpha}} \ln n) + O((\ln n)^{\frac{\alpha}{\alpha+1}})\right) \cdot \left(\frac{2}{\sigma^2} + \frac{2\log 2}{1}\right) \gtrsim n\left(\frac{2}{\sigma^2} + \frac{2\log 2}{n}\right).
$$

最后，应用引理 22，我们得到

$$
\inf_{\mathcal{S}_n \mapsto \hat{F}} \sup_{F \in \mathcal{F}^\circ} \mathbb{E}\Big[\|\hat{F} - F\|^2_{L^2}\Big] \gtrsim \delta_n^2 \gtrsim \exp\big(-O\big((\ln n)^{\frac{\alpha}{\alpha+1}}\big)\big).
$$

□

D 测度值注意力的合成实验

为对我们的风险界进行最小化的经验合理性检验，我们设计了一个简单的合成实验：输入是 $[0,1] \times \{-1,+1\}$ 上的测度值上下文，模型则是一个由单个多层感知机→注意力→多层感知机模块构成的网络。目标是从关联分量与非关联分量的混合分布中，恢复底层某个“关联”测度的标量泛函。

数据生成过程。固定一个截断级别 $M = 16 \in \mathbb{N}$，并在 $[0, 1]$,上取一组正交归一三角基

$$\phi_0(x) = 1, \quad \phi_j(x) = \sqrt{2}\,\sin(\pi j x), \quad j = 1, \ldots, M-1.$$

对于光滑度参数 $\alpha > 0$，我们定义特征值

$$\lambda_j = \exp(-j^\alpha), \qquad j = 1, \ldots, M-1.$$

对于每个训练样本，我们独立采样两组系数 $Z_1, Z_2 \sim \mathcal{N}(0, I_{M-1}) Z_{1,0}, Z_{2,0} = 0$，并在 $[0, 1]$ 上构造（未归一化的）密度

$$\tilde{\mu}_k(x) \;=\; \sum_{j=0}^{M-1} \lambda_j\, Z_{k,j}\, \phi_j(x), \qquad k \in \{1, 2\}.$$

我们在均匀网格 $\{x_t\}_{t=1}^T$, $T = 32$ 上离散化 $[0, 1]$，将密度截断为非负值，并归一化以得到概率质量函数 $(p_k(t))_{t=1}^T$:

$$\mu_k^{\mathrm{raw}}(x_t) = \tilde{\mu}_k(x_t), \qquad \mu_k^+(x_t) = \max\{\mu_k^{\mathrm{raw}}(x_t), \varepsilon\}, \qquad p_k(t) = \frac{\mu_k^+(x_t)}{\sum_{s=1}^T \mu_k^+(x_s)},$$

使用一个较小的截断值 $\varepsilon > 0$[3]

独立地，我们均匀采样一个“查询标签” $v^{(1)} \in \{-1, +1\}$，并令 $v^{(2)} := -v^{(1)}$。然后，我们在 $[0, 1] \times \{-1, +1\}$ 上定义一个乘积测度混合分布:

$$\nu \;=\; \tfrac{1}{2}\big(\mu_1 \otimes \delta_{v^{(1)}}\big) + \tfrac{1}{2}\big(\mu_2 \otimes \delta_{v^{(2)}}\big),$$

其中，μ_k 是一个离散测度，它将质量 $p_k(t)$ 分配给 x_t。为构造输入词元序列，我们抽取 $n_{\mathrm{tokens}} = 5000$ 个独立同分布的蒙特卡罗样本 $(X_i,\ Q_i) \sim \nu$:

$$(X_i, V_i) = \begin{cases} (x_{T_1}, v^{(1)}), & \text{with prob. } \frac{1}{2},\ T_1 \sim p_1, \\ (x_{T_2}, v^{(2)}), & \text{with prob. } \frac{1}{2},\ T_2 \sim p_2, \end{cases} \qquad i = 1, \ldots, n_{\mathrm{tokens}}.$$

最后，我们在序列末尾附加一个单独的“查询词元”$(0,\ v^{(1)})$，因此每个输入示例都是一个序列

$$\big\{(X_i, V_i)\big\}_{i=1}^{n_{\mathrm{tokens}}} \cup \{(0, v^{(1)})\} \;\in\; \big([0, 1] \times \{-1, +1\}\big)^{n_{\mathrm{tokens}}+1}.$$

The 目标输出 Y仅取决于结合测度 μ_1（并且独立于 μ_2 和 $v^{(2)}$):

$$Y \;\simeq\; \tilde{F}^\star(\mu_1, \underbrace{X_{n_{\mathrm{tokens}}+1}}_{\text{query token}}) \;:=\; v^{(1)} \cdot \sum_{j=0}^{M-1} \lambda_j\, Z_{1,j}^2.$$

直观而言，模型必须使用最后的查询词元 $(0,\ v^{(1)})$ 关注与 q_1 一致的词元，并从 μ_1 的蒙特卡罗样本中恢复有关隐藏系数 Z_1 的信息。注意，我们在训练期间加入了标准差为 $= 0.01$ 的少量高斯噪声。

Mo删除和训练。 我们采用一种最小化架构，以镜像理论上的测度注意力 n ope算子:

$$\text{context/query MLP} \;\to\; \text{measure attention} \;\to\; \text{MLP HEAD}.$$

对于每个样例，我们构造一个包含 T_{ctx} 个上下文词元 $(x_t,\ v_t) \in \mathbb{R}^2$ 的序列，并附加一个最终查询词元 $(0,\ v_{\mathrm{query}})$。上下文词元和查询词元分别通过两个两层多层感知机嵌入到 $\mathbb{R}^{d_{\mathrm{model}}}$ with $d_{\mathrm{model}} = 8$，其隐藏宽度为 $d_{\mathrm{hidden}} = 8$。得到的查询嵌入提供 Q向量，而上下文嵌入则为单个 4 头 softmax 注意力层提供 K 和 V向量。这个“测度注意力”层输出一个单一的 d_{model} 维表示ˆ，并将其输入最终的两层多层感知机头，以产生标量预测 Y。我们使用平方损失 $\ell(\hat{Y}, Y) = (\hat{Y} - Y)^2$ 进行训练，采用 Adam 优化器和指数衰减学习率，共训练 20 轮。

对于每个 $\alpha \in \{\alpha_1, \ldots, \alpha_L\}$，我们生成大小为 $n \in \{n_{\min}, \ldots, n_{\max}\}$（$n = 2^k$ 对于 $k = 2,$ ……，6）的相互独立的训练集，并在留出的验证集（$n_{\mathrm{val}} = 2000$）上测量经验风险 $L(n)$。

[3]为简化起见，我们的理论并未明确研究这种截断值或归一化。

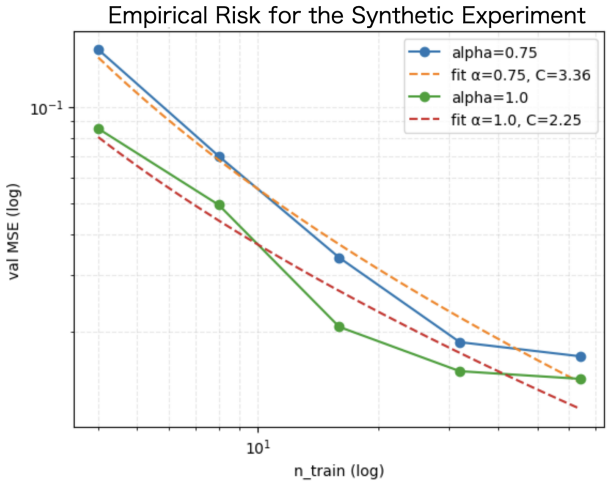

图 4：合成测度值实验的经验风险 $L(n)$，绘制在变换坐标轴（$\log n$）$^{\alpha/(\alpha+1)}$ 上，并与拟合曲线一起展示。每个风险均使用 2000 个未知样本计算。

风险缩放。 理论预测，在此设定下极小极大风险的衰减为

$$L^\star(n) \;\approx\; \exp\!\big(-c\,(\log n)^{\alpha/(\alpha+1)}\big),$$

至多相差乘性常数。为与该预测进行比较，我们对每个 α 拟合如下参数形式

$$\log L(n) \;\approx\; A_\alpha \;-\; C_\alpha\,\big(\log n\big)^{\alpha/(\alpha+1)}$$

即在(A_α, C_α)上，利用测得的数据对 $\{(\log\, n_i, \log\, L(n_i))\}_i$进行最小二乘法拟合。图4展示了$\log L(n)$随$(\log\, n)^{\alpha/(\alpha+1)}$ 变化的情况以及拟合曲线。

作为最基本的合理性检验，该合成实验（图4）表明，改变谱衰减参数 α 会系统性地影响收敛速度：重尾谱（较小的 α）会使经验风险的衰减明显变慢。这在定性上与理论预测一致，尽管我们并未试图匹配精确的渐近速率。

注意力权重分析。 为检验注意力层是否确实使用了查询标签，我们检查训练好的模型在验证集上的softmax 注意力权重。对于给定样本，令

$$\{(X_t, V_t)\}_{t=1}^{T} \in \big([0,1] \times \{-1,+1\}\big)^T$$

表示这 T 个上下文词元，并令

$$(X_{\mathrm{q}}, V_{\mathrm{q}}) = (0, v^{(1)})$$

为追加到序列末尾的查询词元。对于每个注意力头 $h = 1$、……、 H，我们记

$$a^{(h)} \in [0,1]^T$$

表示从查询到 T 上下文位置的 softmax 注意力权重，使得

$$\sum_{t=1}^{T} a_t^{(h)} = 1.$$

我们关注查询词元如何在标签与查询相同的词元和标签相反的词元之间重新分配注意力质量。因此，我们定义索引集

$$S_{\mathrm{same}} \;:=\; \{1 \le t \le T : V_t = V_{\mathrm{q}}\}, \qquad S_{\mathrm{diff}} \;:=\; \{1 \le t \le T : V_t \neq V_{\mathrm{q}}\},$$

也就是说，我们只考虑上下文词元，并从两个集合中都排除查询词元本身。对于每个注意力头 h，我们接着计算查询分配给标签相同词元和标签不同词元的平均单词元注意力权重，

$$\bar{w}^{(h)}_{\text{same}} := \frac{1}{|S_{\text{same}}|} \sum_{t \in S_{\text{same}}} a^{(h)}_t, \qquad \bar{w}^{(h)}_{\text{diff}} := \frac{1}{|S_{\text{diff}}|} \sum_{t \in S_{\text{diff}}} a^{(h)}_t,$$

以及注意力质量总量

$$m^{(h)}_{\text{same}} := \sum_{t \in S_{\text{same}}} a^{(h)}_t, \qquad m^{(h)}_{\text{diff}} := \sum_{t \in S_{\text{diff}}} a^{(h)}_t.$$

在实践中，我们通过向注意力模块添加一个标志来实现这一点；启用该标志后，前向传递结束时会将最近一次 softmax 注意力张量 $A \in \mathbb{R}^{B \times H \times 1 \times T}$ 存储在 CPU 上（其中 B 为批量大小），随后我们提取 $a^{(h)}$ 作为长度为 T的向量 $A_{b,\ h,\ 1,\ :}$，它对应于每个样本 b 和注意力头 h中查询到上下文的权重。

我们在表 1 中报告了每个注意力头 h在 1000 个验证样本上的 $m^{(h)}_{\text{same}}$和 $m^{(h)}$ 差异 的均值和标准差。在这一合成任务中，四个注意力头中有两个几乎将全部注意力质量集中在标签与查询标签匹配的词元上；另一个注意力头则表现出相反的偏好，而还有一个注意力头则近乎对称。对各注意力头取平均后，查询词元分配给标签相同词元的总质量，大于分配给标签相反词元的总质量，这表明整体上存在偏向标签条件检索的趋势。

平均权重的绝对尺度 $\bar{w}_{\text{same}}$和 $\bar{w}_{\text{diff}}$ 很小（约为 10^{-4}），这仅仅是因为注意力分布是在长度为 $T_{\text{ctx}} \approx n_{\text{tokens}} = 5000$的位置序列上归一化的。在近似均匀基线下，我们有

$$\bar{w}_{\text{unif}} \approx \frac{1}{T_{\text{ctx}}} \approx \frac{1}{5000} \approx 2 \times 10^{-4},$$

因此，报告的数值应相对于这一 $1/T_{\text{ctx}}$ 尺度来解读，而非视为绝对概率。在我们的构造中，每个上下文词元都独立地以概率 $1/2$来自 $\mu_1 \otimes \delta_{v^{(1)}}$ 或 $\mu_2 \otimes \delta_{v^{(2)}}$，因此通常有 $|S_{\text{same}}| \approx |S_{\text{diff}}| \approx T_{\text{ctx}}/2$。因此，约为 $\bar{w} \approx 2 \times 10^{-4}$ 的数值对应于在所有上下文词元上近乎均匀的注意力，而约为 $\bar{w}_{\text{same}} \approx 4 \times 10^{-4}$和 $\bar{w}_{\text{diff}} \approx 0$ 的数值则表明，某个注意力头几乎将全部注意力质量都放在同标签子集上（相反标签的情况也类似）。

作为一项合理性检查，为确认模型确实使用了查询输入，我们在评估时进行了“查询打乱”实验：在每个小批量内，随机打乱不同样本之间最后一个（查询）词元，同时保持上下文词元和目标不变，然后重新计算验证损失（见表 1 底部）。在这项合成任务中，打乱查询词元会增大验证均方误差，证实模型确实在一定程度上依赖查询输入。

这项简要实验并非旨在进行全面的实证研究，但它提供了一项合理性检查，表明理论预测的风险定性排序在一个简单的测度值注意力设置中可以复现。

表 1：从查询词元到同标签与不同标签上下文词元的平均注意力权重（在 1000 个验证样例上的均值和标准差，使用 64 条训练数据以及 $\alpha = 1.0$）。在 $T_{\mathrm{ctx}} = 5000$ 且大约一半词元与查询标签相同的情况下，均匀基线为$\bar{w} \approx 2 \times 10^{-4}$，而几乎只关注同标签词元的注意力头达到 $\bar{w}_{\mathrm{same}} \approx 4 \times 10^{-4}$。我们还报告了使用原始查询以及在 1000 条验证数据中打乱查询后的验证均方误差。

Head	$\bar{w}_{\mathrm{same}}$	$\bar{w}_{\mathrm{diff}}$	标准差(w_{same})	标准差(w_{diff})
0	1.96×10^{-4}	2.04×10^{-4}	2.80×10^{-4}	2.85×10^{-4}
1	1.44×10^{-19}	4.00×10^{-4}	2.10×10^{-19}	4.00×10^{-4}
2	$\mathbf{4.00 \times 10^{-4}}$	$\mathbf{1.17 \times 10^{-14}}$	4.00×10^{-4}	1.70×10^{-14}
3	$\mathbf{4.00 \times 10^{-4}}$	$\mathbf{2.86 \times 10^{-27}}$	4.00×10^{-4}	**0**（太小）
mean	2.49×10^{-4}	1.51×10^{-4}	2.70×10^{-4}	1.71×10^{-4}

	原始查询	打乱的查询
验证集均方误差	1.44×10^{-2}	7.75×10^{-1}