arXiv:2302.02774v2 [stat.ML] 1 Jun 2023

自监督学习 相互作用：增强、归纳偏置与泛化

Vivien Cabannes [1] Bobak T. Kiani [2] Randall Balestriero [1] Yann LeCun [1] Alberto Bietti [1]

摘要

自监督学习（SSL）已成为一种强大的框架，能够在无需监督的情况下从原始数据中学习表征。然而，在实践中，工程师面临着诸多问题，例如优化器调优不稳定，以及训练过程中表征坍塌。这些挑战促使我们需要一种理论，以阐明数据增强的选择、网络架构与训练算法之间复杂的相互作用。我们在一个便于理论分析的设定中，对预训练和下游任务上的泛化性能进行精确分析，以研究这种相互作用，并提出若干源自理论、对自监督学习实践者有启发的见解。

1. 引言

自监督学习（SSL）旨在无需预构造标签，便能构建有用的数据表征。鉴于自监督学习近期取得的成功及其广泛适用性，如今成熟的大型神经网络训练方法已将模型的无监督预训练纳入其中：先在大量数据上进行预训练，再在下游数据集上对模型进行微调/探测（Devlin 等人，2019年；Chen 等人，2020 年；Brown 等人，2020 年；Radford 等人，2021年）。自监督预训练通常旨在使模型对输入的某些失真/视图保持不变，以便提取可用于下游任务的有用特征（例如，Chen 等人，2020年；Caron 等人，2020 年；Grill 等人，2020 年；Caron 等人，2021年；Bardes 等人，2022）。尽管自监督学习方法功能强大，但要正确实现并非易事。它们往往会遇到各种实际问题，例如训练过程中的不稳定与崩溃，以及需要仔细调整与架构、优化算法、表征维度和数据增强形式相关的参数。预训练的这些不同方面可能导致行为和表征大相径庭，例如，图1对此进行了说明。这些挑战促使人们寻求新的理论洞见，以更好地理解这些问题为何出现，以及如何更有效地解决它们。

我们的研究聚焦于联合嵌入框架，并刻画在给定输入分布、数据增强和架构选择下学习到的表示。为了获得细致入微的认识，我们研究配备再生核的线性函数类，并分析一种理论上便于研究的损失函数，该函数同时对对比方法和非对比方法进行建模。我们的工作将HaoChen 等人（2021年的离散数据设置以及 Saunshi 等人（2022）的有限维设置加以推广，涵盖了表达能力更强的非参数模型，这类模型可能具有通用逼近性质，并且能够通过其极限核捕捉架构的某些属性（Jacot 等人，2018）。

我们的贡献如下：

1. 我们揭示了两个核心积分算子：一个是“内在”算子，取决于输入分布和增强方式的选择；另一个则刻画了与计算模型相关的归纳偏置。2. 我们针对下游泛化误差给出了新的上界，较以往工作更为紧致，并且能够处理数据在增强前后分布发生偏移的情况。3. 我们借助凸分析工具，为预训练超额风险提出了新的泛化界。该分析带来了新的见解，包括阐明每个样本采用多次增强（例如“多裁剪”）的益处。

4. 我们详述了若干示例，其中最优表征可用闭式形式求得，展示了数据增强、架构和正则化在形成表征中的作用。

5. 我们讨论了理论带给自监督学习实践者的若干实践见解，尤其是预训练中的设计选择可能如何影响下游性能，以及如何避免表征坍塌。

相关工作。过去几年中，理论分析自监督学习的基础逐渐形成。与我们的工作尤其相关的是，Balestriero 与 LeCun (2022年); Kiani 等人

[1]Meta AI，美国纽约州纽约市 [2]麻省理工学院电气工程与计算机科学系，美国马萨诸塞州剑桥市。联系邮箱：VivienCabannes < vivc@meta.com>。

《 40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023年作者。

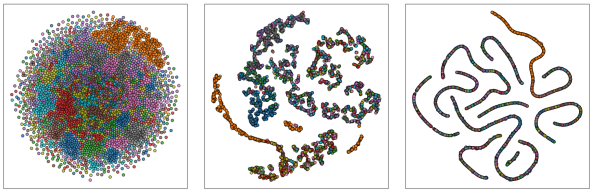

图 1。数据增强和架构的影响。MNIST 上在无数据增强（左）、采用旋转和多层感知机（中）或卷积神经网络（右）的情况下学习到的表征的 t-SNE 可视化。这些表征同时取决于数据增强和架构。

(2022年) 为许多自监督学习设定提供了便于理论分析的刻画，包括核设定下表征的闭式解。对于对比学习，Arora等人最早在理论上对自监督学习进行了分析 Arora等人 (2019年);Tosh等人 (2021a;b); Tian等人 (2021年)。值得注意的是，HaoChen 等人 (2021年) 近期利用谱图理论中的工具，在聚类假设下刻画了自监督学习性能的保证。这些假设被Saunshi 等人 (2022年) 认为不切实际；他们强调，引入归纳偏置对于获得可证明保证的重要性。这项工作后来由Lee 等人 (2021年) 扩展到多模态自监督学习，其中本质上是将中心对称算子 T 替换为非对称算子，并将特征分解替换为奇异分解。Wen与Li和Wen与Li(2021年) 以及Tian(2022年) 也通过对训练动力学中特征学习的分析，进一步审视了归纳偏置的作用。

2. 设置

机器学习简化了创建用于在数据中寻找模式的算法这一任务。算法可被视为从输入 $x \in \mathcal{X}$ 到输出 $y \in \mathcal{Y}$的映射 f。要构建此映射 $f\colon \mathcal{X} \to \mathcal{Y}$，可以选择不一致性度量 $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$，并最小化风险

$$\mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\rho}[\ell(f(X), Y)], \tag{1}$$

对于 $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ 上的输入/输出对分布。我们将根据风险得到的最优输入/输出映射记为 $f^* \in \arg\min \mathcal{R}$。将原始输入（例如像素数组）映射为输出（例如对图像中的动物进行分类）通常是一项具有挑战性的任务。一种有效的技术是，先从输入数据中提取（或设计）有意义的特征 $\psi\colon \mathcal{X} \to \mathbb{R}^k$，再利用这些特征以 $g \circ \psi$ 的形式搜索 f，从而得到 $g\colon \mathbb{R}^k \to \mathcal{Y}$一个简单函数。[1]

尽管特征 ψ 可以通过人工设计获得，表征学习旨在通过无监督学习过程改进这种设计。一方面，基于重建的方法 通过映射 Mx 对输入 x 进行遮蔽或添加噪声，并旨在根据特征 $g\circ\psi$重建原始输入

[1]为方便起见，可测性等若干技术细节将在附录中讨论。

实践	理论	数量
数据增强	谱嵌入	T
架构	函数空间	K
优化	正则化	λ
微妙的相互作用		T_λ

表 1。实践与理论的类比，本文提出这一类比，以帮助厘清自监督学习训练中的各种现象。

使用 g，即一个简单的预测头。大语言模型很大程度上依赖这一范式，通常通过补全句子来进行学习 $\psi\ Mx$，其中词元会被遮蔽（例如 Devlin 等，2019）。另一方面，联合嵌入方法通过利用输入中所含语义信息对微小扰动的不变性来学习 ψ。这是我们将要关注的范式。近年来，联合嵌入方法很大程度上依赖数据增强的概念，例如对像空间进行小幅旋转、平移和颜色抖动。特别是，对比方法通过强制要求：如果两个增强 ξ 和 ξ' 来自同一个数据点，则它们的表示 $\psi(\xi)$ 和 $\psi(\xi')$彼此接近；而如果它们来自不同的数据点，则其表示彼此远离，从而进行学习 ψ（例如 Chen 等，2020）。非对比方法只要求增强数据点之间具有相似性，并通过强制要求表示具有丰富性来避免坍塌（例如参见 Bardes 等人，2022）。接下来，我们将重点介绍 VICReg（Balestriero 与 LeCun，2022）的一种理论上友好的变体，其参数为 $\beta > 0$，定义于 $\psi\colon \mathcal{X} \to \mathbb{R}^k$ 由

$$\mathcal{L}(\psi) = \beta\, \mathbb{E}_X\, \mathbb{E}_{\xi,\xi'}\left[\|\psi(\xi) - \psi(\xi')\|^2 \,\middle|\, X\right] + \left\|\mathbb{E}_\xi[\psi(\xi)\psi(\xi)^\top] - I\right\|_2^2, \tag{2}$$

其中，输入/数据增强对 $(X,\ \xi)$ 服从分布 $\mu \in \Delta_{\mathcal{X}\times\mathcal{X}}$，其条件分布 $(\xi \mid X)$ 由数据增强的选择决定。$\mathcal{L}$ 中的第一项强制表示 ψ 对同一输入 X 的两种数据增强 ξ 和 ξ' 保持不变性；第二项则通过推动 ψ_i 的坐标：$\mathcal{X} \to \mathbb{R}$ of$\psi = (\psi_i)_{i\in[k]}$ 在 L^2中彼此正交，来降低表示坍塌的风险。

注 1 (采用 $\mathcal{L}$ 的对比学习)。 当 $\beta = 1$时，总体损失 $\mathcal{L}$ 等价于 *HaoChen 等人(2021年)* 研究的谱对比损失，它是 *SimCLR*（*Chen 等人，2020 年*）的一种理论上易于分析的替代目标。换言之，$\mathcal{L}$ 同时分析了对比式和非对比式表征学习方法。

给定一个表示 ψ，可以通过线性探测来优化 f：构造 $f = g \circ \psi$，其中 g 是一个线性函数。因此，f 属于函数类

$$\mathcal{F} = \left\{x \mapsto w^\top \psi(x) \,\middle|\, w \in \mathbb{R}^k\right\}. \tag{3}$$

在实践中，我们可能不知道最优的 ψ，但可以根据经验数据将其估计为 ψ，从而得到这一类函数的估计值 $\mathcal{F}$。

3. 表征学习

本节中，我们研究通过使用特定的数据增强和归纳偏置进行预训练所诱导的表征。

3.1. 闭式解

方程（2）注意到不变部分是一个二次型后，可为 ψ 导出闭式解。

引理 2（谱嵌入）。存在一个属于 L^2 的线性正对称算子 L，使得算子 $I - T$ 为正值，且

$$\mathbb{E}_X \, \mathbb{E}_{\xi,\xi'} \left[\|\psi(\xi) - \psi(\xi')\|^2 \,\middle|\, X \right] = \sum_{i \in [k]} \psi_i^\top L \psi_i .$$

为与先前文献保持一致，我们将改用 $T = I - L/2$，它也是一个线性正对称算子，定义如下：对于 ψ_1，$\psi_2 \in L^2$

$$\psi_1^\top T \psi_2 = \mathbb{E}_X \, \mathbb{E}_{\xi,\xi'} \left[\psi_1(\xi)^\top \psi_2(\xi) \,\middle|\, X \right]$$

因此，若 (λ_i) 是 T 的特征值，且 (f_i) 是相应的特征向量，则 $\mathcal{L}$ 的一个最小化器为 $\psi_i = \sqrt{\mu_i} f_i$，其中 $\mu_i = 1 - \overline{\beta} + \beta \lambda_i$。

引理 2 与无监督学习中的指导原则密切相关：数据的良好表征应使数据流形上的变化最小化（Cabannes 等人，2023），以及通过对中心算子进行谱分解来学习此类表征的技术（例如，Coifman 和 Lafon，2006）。

3.2. 在线性函数类中搜索

在这个更具技术性的章节中，我们研究属于线性函数类的 $\mathcal{L}$，其中 ψ的映射坐标 $\psi\colon \mathcal{X} \to \mathbb{R}^k$ 通常在函数空间 $\Psi \subset \mathbb{R}^{\mathcal{X}}$中搜索，从而得到 $\psi \in \Psi^k$。在便于理论分析的设定中，我们假设 Ψ 是一个赋有希尔伯特拓扑的线性函数类，使得线性求值 $\psi \mapsto \psi(x)$ 对几乎所有 $x \in \mathcal{X}$都是连续的。再生核希尔伯特空间理论（Scholkopf & Smola，2001年）指出，Ψ 可以由一个希尔伯特空间 $\mathcal{H}$ 和一个映射 $\varphi : \mathcal{X} \to \mathcal{H}$ 参数化，使得

$$\Psi = \{ x \mapsto f_\theta(x) \,|\, f_\theta(x) = \langle \theta, \varphi(x) \rangle_{\mathcal{H}}, \theta \in \mathcal{H} \} . \quad (4)$$

这推广了 HaoChen 等人(2021年的设定，在该设定中，$\mathcal{X}$被假设为有限，并且 Ψ 由 $\mathcal{H} = \mathbb{R}^{\mathcal{X}}$ 和 $\varphi(x) = \delta_x$参数化；它也推广了Saunshi 等人 (2022) 的设定，在该设定中，$\mathcal{H}$ 被假设为有限维。

为了描述具有这种线性结构的架构（例如神经网络），通常会将这些模型线性化（例如 Jacot 等人(2018)），形式如下

$$\psi_\theta(x) = \psi_{\theta_0}(x) + \langle \nabla_{\theta_0} \psi_{\theta_0}(x), \theta - \theta_0 \rangle + o(\|\theta - \theta_0\|),$$

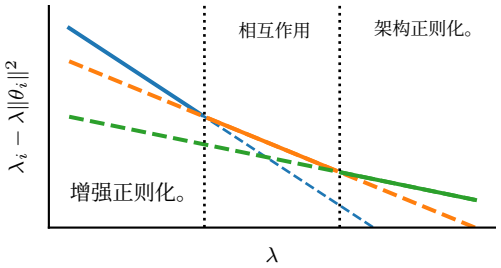

图 2. T与 K 之间的相互作用随 λ变化。当 $(\lambda_i) = (.9, .75, .5)$，且 $(\|\theta_i\|^2) = (.4, .25, .125)$时，对命题 4 的示意说明。图中展示了与三个不同特征函数相关的特征值随 λ的变化；为方便起见，β 设为 1。当 $\lambda = 0$，最小化器ψ_*：$\mathcal{X} \to \mathbb{R}$ 由 (2) 定义为 T，此处为 $\varphi_* = f_1$（$i = 1$，蓝色所示）；当 λ 较大时，$\psi_* = f_3$（绿色）主要取决于 K。在中间区域，这两种状态相互作用，产生 $\psi_* = f_2$（橙色）。这三种状态分别称为“数据增强”状态、“架构”（或 VCReg）状态和“相互作用”状态。图 8 详细说明了如何将这一抽象设定具体化为两层 ReLU 网络和裁剪。

其中 θ 是网络参数，假设其接近初始化值 θ_0，而 ψ_θ 是神经网络。在这种情况下，我们可以取 $\varphi = \nabla_{\theta_0} \psi_{\theta_0}$，这或许描述了宽神经网络的一些情形（Lee 等人，2019年）。

为了在实践中最小化 $\mathcal{L}$ 并改善泛化，通常会引入一个正则化参数。[2] 下面的引理给出了 $\mathcal{L}$的正则化变体的闭式解。

引理 3（正则化总体损失）。对于 $\Theta \in \mathbb{R}^k \otimes \mathcal{H}$以及正则化项 $\lambda > 0$，正则化损失 $\mathcal{L}(S\Theta) + \lambda \|\Theta\|_2^2$ 可以通过下述算子以闭式形式最小化：

$$T_\lambda = (1 - \beta) I + \beta T - \lambda K^{-1} . \quad (5)$$

其中，对于 S，$K = SS^\top$：$\mathcal{H} \to L^2(\mu_\Xi)$；$\theta \mapsto f_\theta$ 是 $\mathcal{H}$ 在 L^2中的嵌入。具体而言，若 (λ_i) 是 T_λ 的（递减）特征值，而 (f_i) 是相应的特征函数，则一个最小化器为 $\psi_i = \max\{\lambda_i, 0\} f_i$。

3.3. 对归纳偏置的需求

引入正则化项 $\lambda \|\Theta\|$ 并得到算子 T_λ的动机来自两种不同的观点。

在统计学习理论的经典观点中，人们会尝试求出 T 的特征函数，以最小化 $\mathcal{L}$（引理 2）。然而，仅能访问有限数量的数据样本时，应在 T 中搜索特征函数

[2]虽然为简便起见，我们在此研究 Tikhonov正则化的偏差，但也可以对早停梯度下降或随机梯度下降进行类似研究，方法是将它们表示为谱滤波器，如Lin 等人(2020 年) 所示；另可参见矩阵分解问题优化相关的文献（Chi 等人，2019年），Simon 等人已将其应用于自监督学习Simon 等人(2023)。

一个容量有限的空间中（即 $\{f \in \Psi \,|\, \|f\|_\Psi^2 \leq \lambda^{-1}\}$）。尽管较小的模型所需样本较少（例如，深度网络中的神经元和层数较少），但这类小模型不太可能具有足够的表达能力来表示理想解。这体现了近似误差与估计误差之间的经典权衡。对于拉普拉斯算子，可以假设 T 的特征函数是光滑的，因此属于一个较小的函数空间，并且可以用有限的计算模型很好地近似。感兴趣的读者可参阅 Cabannes 等人 （2021年a），其中给出了这方面的结果，条件是 $I - T$ 是 L^2 中的实拉普拉斯算子。

另一种观点由 Saunshi 等人(2022) 提出，他们指出，在现实情形下，T的特征值可能具有较大的重数（特别是在第 4.2 节的非局部化增强设置中），这意味着空间 $\mathcal{F}$无法由损失 $\mathcal{L}$唯一确定。因此，仅根据 T 来定义最优解有些不适定；而当 K 被恰当选择时， T_λ 便可以定义一种“更具原则性”的表示 ψ。矛盾的是，从这一观点来看，归纳偏置可以降低近似误差。图 3 阐释了这一思路。它利用了在 T 与 K 可交换这一友好设置下，对归纳偏置的一种解释。

命题 4。 若 T 与 K 可交换，且 (λ_i)是 T 的特征值，而(f_i)是其特征函数，则存在(θ_i)，使得 $f_i = f_{\theta_i}$ (4)。此外，使正则化损失最小的最优表示是那些能最大化 f_i 的 $\beta\lambda_i - \lambda \|\theta_i\|^2$. 。换言之，正则化会偏向于相对于计算模型而言复杂度较低的表示。

引理3 展示了 VCReg 损失的一个有趣特性（$\beta = 0$，即不含不变性项的 VICReg）。在这种设置下，最优的 ψ 会检索 K最大的特征函数，从而恢复核主成分分析。通过对所得 ψ 进行线性探测来学习下游任务，等价于采用特征值截断的线性回归，这是一种强大的谱滤波技术（例如，参见Bun 等人，2017年）。

4. 说明性示例

本文的分析依赖于两个中心算子：T，它由数据分布和增强“内在地”定义；以及 K，它与计算模型（例如网络架构）相关。一旦选定这些算子，第5节将给出精确分析

在核区间中使用自监督学习时的收敛与泛化。实质上，假设 3要求在拥有无限数据的情况下，下游目标函数与上游学习到的表示具有良好的一致性。在此， T 以及 λK^{-1} 所引入的归纳偏置对学习到的表示的影响可能显得较为抽象。为了帮助直观理解这些算子的作用并概述其重要属性，本节列举了几个具体示例，以帮助实践者更好地理解数据增强及其与架构归纳偏置之间的相互作用。

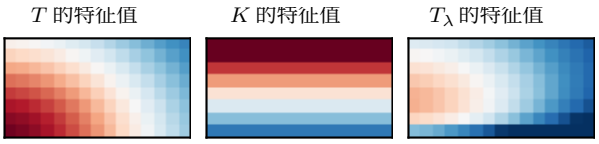

图3。特征值在 T 与 K之间的权衡。 调和设定的示意图，其中 T和 K 在同一基下对角化。该基由“不变性评分”（$x = m$ 见（54））和“复杂度评分”（$y = |S|$ 见（54））进行参数化。对于 $A \in \{T, K, T_\lambda\}$，特征值 $\lambda_{x,\,y}(A)$ 以不同颜色表示，并显示在与 $x \in [15]$ 和 $y \in [8]$相关联的网格中。单独使用算子 T会偏向不变性（较低的 x）和高复杂度（较低的 y），而单独使用 K 则偏向低复杂度。二者的相互作用产生了 T_λ，其最大的特征函数具有高不变性和低复杂度，并对应于理想表示 ψ。

践者更好地理解增强及其与架构归纳偏置之间的相互作用。

理解 SSL 中所学习表示的两种不同视角已经出现。一种直觉源自谱聚类文献，是小节4.1的研究对象。理解 SSL 的另一种直观方式基于调和分析，是小节 4.2的研究对象。总的来说，本节通过去除数据中的强聚类假设，推广了以往的工作，表明真正重要的是 T的特征函数；当采用此类聚类假设时，这些特征函数最终会捕捉聚类结构。本节进一步使用调和分析工具，更好地描述这些特征函数，正如Saunshi 等人 （2022）所建议的，并在表 2中进行了详细说明。

4.1. 局部增强下的低方差

当增强$(\xi \,|\, X)$ 局部于输入 X周围时，优化损失 $\mathcal{L}$ (2) 会促使 ψ 沿增强方向的梯度较小。形式化地，对于 $\psi : \mathcal{X} \to \mathbb{R}$ 利用一阶泰勒展开，

$$\mathbb{E}\left[\|\psi(\xi) - \psi(\xi')\|^2 \,\middle|\, X\right] \simeq \mathbb{E}\left[\langle \nabla\psi(X), \xi - \xi'\rangle^2 \,\middle|\, X\right].$$

在各向同性增强下，目标函数可简化为

$$\mathbb{E}\left[\langle \nabla\psi(X), \xi - \xi'\rangle^2 \,\middle|\, X\right] \propto \|\nabla\psi(X)\|^2,$$

这会使 ψ 在输入空间中密集分布的区域呈现较小的变化——这让人想起过去二十年中用于处理表示和半监督学习的主流方法（vanEngelen &Hoos，2020 年）。更一般地，增强决定了 ψ的重要不变方向，重新得到了切线传播算法的有限差分方法（Simard et al.，1991）。

当数据呈现聚类结构时，低变化方法尤其有用（参见图 10，其中展示了神经网络的一个示例）。如果数据增强保持

聚类结构，$\mathcal{L}$ 则在每个簇上取分段常数函数可使其最小化，从而为涉及不同簇分类的下游任务带来有用特征（HaoChen 等人，2021年；Schiebinger 等人，2015）。归纳偏置进一步使这些顶层特征函数发生变形，使其在 Ψ （4）所定义的意义下具有正则性；例如，当我们使用径向基函数核时，它们是解析的（Sun&Zhou，2008）。

4.2. 增强的作用

当增强不具备局部性时——实践中通常如此——调和分析为深入研究增强的作用提供了有用工具，尤其是在数据均匀分布于球面或超立方体上的情况下。我们关于超立方体的研究结果总结于表 2 中。在这种设置下，我们证明，常见的增强会强制产生平滑性、局部性，或对某些对称性保持不变。例如，裁剪促使 ψ 关注裁剪窗口大小范围内可能出现的细节，从而滤除输入各部分之间可能属于伪特征的远程交互。下面的示例将这一点形式化。

示例 1（裁剪）。考虑超立方体设置，其中 $\mathcal{X} = \{-1, 1\}^n$ 和 X 均匀分布。L^2（$\mathcal{X}$、$\mathbb{R}$）的一组基由奇偶函数 $\chi_S\colon x \mapsto \prod_{i\in S} x_i$ 给出，其中包含所有子集 $S \subseteq [n]$。通过窗口大小为 $v \times w$ 的裁剪进行预训练时，对于所有 S，其支撑集形成了大于 $v \times w$的窗口，集合 $T\chi_S = 0$。对于其他所有 S，$T\chi_S = \lambda_S\chi_S$，其中 λ_S 随 S的直径增大而减小。换言之，采用二维裁剪进行预训练会消除在裁剪窗口之外进行全局作用的函数的影响。实际上，这会赋予诱导表示 ψ 一定的局部性，而这通常有利于泛化。

这个示例表明，理想的裁剪大小应与 ψ所需的细节尺度相匹配；例如，对于 iNaturalist 这类包含细粒度细节的数据集，相较于 ImageNet 这类数据集，应减小裁剪窗口大小。附录 D进一步讨论了其他数据增强示例，例如随机噪声或平移，并说明它们如何使特征函数偏向于平滑或不变。

4.3. 与架构的相互作用

尽管数据增强和架构的设计可以分别进行，但架构和优化方案的改变会对最终的最优 ψ产生重要影响。一般来说，通过增加 λ 来增强归纳偏置，会使 ψ 趋向于更平滑的函数，这一点可由 $\mathcal{H}$ 范数体现，如图4所示。在实践中，应该施加多大程度的归纳偏置（此处由参数 λ表示），通常是根据直觉、常识和实证结果综合确定的。例如，Caron 等人(2021年)指出，提前停止所带来的归纳偏置会产生有益效果，并指出“训练时间更长 [.....] 会导致更差的表现 ce”.

示例 *2*（点积核）。在示例 *1* 的布尔超立方体设定中，许多线性模型 *(4)* 具有 $\varphi(x)^\top\varphi(y) = h(\langle x, y\rangle)$ 的形式（例如，全连接层的经典神经切线核线性化），从而导出一个可由奇偶函数对角化的积分算子 K。更准确地说，存在 $(\nu_i) \in \mathbb{R}^d$，使得 $K\chi_S = \nu_{|S|}\chi_S$，其中 $|S|$ 是 S 的基数，且 $\nu_{|S|}$ 随 $|S|$增大而减小。在裁剪的设定中，T 促使表示偏向于直径较小的奇偶函数（对于直径较小的 S，$\psi = (\chi_S)_S$），而归纳偏置则作用于集合 S的基数，促使模型偏向于使$\nu_{|S|}$最大化的 χ_S。形式化推导见附录 *D*。

附录 D提供了更多示例。例如，在平移的情况下，K 中的低阶偏置与 T中的不变性偏置之间也存在类似的相互作用。我们还考虑了卷积架构，它们除了低阶偏置之外，还可以通过对$\mathrm{diam}(S)$施加约束来实现局部性。图 3 展示了特征值中的此类权衡，图 4 以可视化方式展示了这种相互作用可能如何影响球面设置中的主导特征函数，而图5 说明了由此对不同下游任务产生的影响。

5. 收敛分析

下一节将分析预训练任务和下游任务的保证。[3] 为简便起见，我们考虑均方损失 $\ell(y,\ y') = \|y - y'\|^2$，其中 $\mathcal{Y} = \mathbb{R}^{d_y}$。得益于校准不等式（Bartlett 等人，2006）或自协调性质（Ostrovskii 与 Bach，2018），许多损失的研究都可以归约为最小二乘情形。为精确研究收敛速率，我们考虑第 3.2节的核区间，其中 $\mathcal{F}$ 通过引理 Θ_* 中的3指定为

$$\mathcal{F} = \left\{x \to w^\top\Theta_*\varphi(x) \,\middle|\, w \in \mathbb{R}^k\right\}, \tag{3}$$

$\hat{}\hat{}$，$\mathcal{F}$的定义类似，只是将 Θ 作为 Θ_* 的估计。下文中，(f_i) 表示按特征值降序排列的 T_λ 的特征函数，并且在本节中，λ 被视为固定不变。

5.1. 应对分布偏移

自监督学习算法通常会采用强数据增强，从而可能导致输入与增强之间的边缘分布有所不同。这种差异常常被忽视，许多理论研究都隐含地假设 $\rho_{\mathcal{X}} = \mu_\Xi$。在实践中，下游任务中输入的边缘分布 $\rho_{\mathcal{X}}$可能与增强的边缘分布 μ_Ξ存在显著差异。

[3] 预训练和下游任务分别指最小化 $\mathcal{L}$ 和 $\mathcal{R}$。

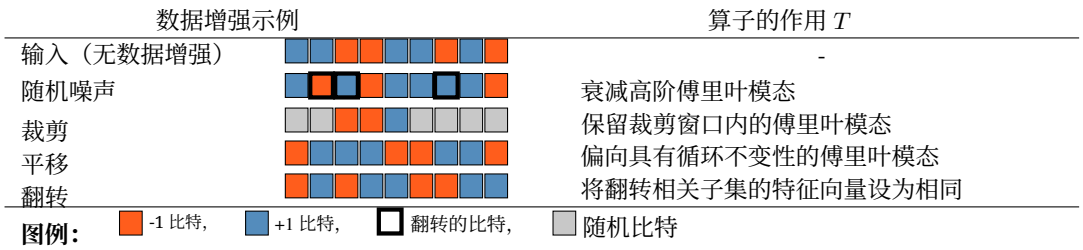

表 2. 常见数据增强 对最优表示 ψ 通过算子 T 的影响。若不进行数据增强，ψ 可以匹配任意傅里叶基函数。数据增强会通过衰减这些基函数在 T 中的特征值，筛除其中一部分，而架构则会通过算子 ψ 促使 ψ 从剩余频率中选取某些特定频率 K。该表展示了常见数据增强对比特流上的奇偶函数的影响。更多细节和推导请参阅附录D。

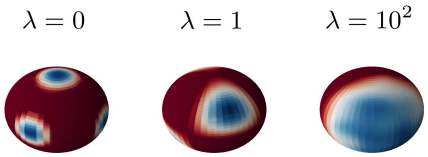

图 4。球面上的相互作用。 T_λ 的第 7 个特征函数在三种不同 λ下的等高线 y， z 坐标以及高斯扰动构成了数据增强。 K 是与径向基函数相关联的积分算子。不进行正则化（左图）时，该特征函数高度局限于与增强作用相对应的簇。递增的正则化会使其倾向于 K 更平滑的调和特征函数（中图和右图）。

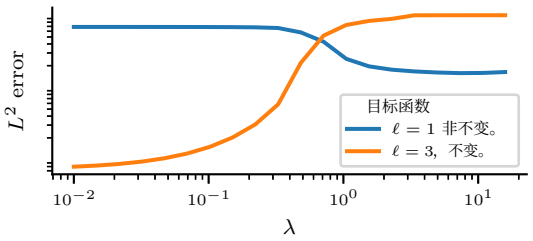

图 5.下游误差的权衡。 预训练正则化 λ 对球面上两个任务的经验下游误差的影响 $\mathbb{S}^7$。目标函数 f_ℓ^* 是次数为 $\ell \in \{1, 3\}$ 的多项式，其中只有 f_3^* 对平移不变。 K 由点积核构成，该核对次数起正则化项的作用，而 T 由局部平移构成。仅根据 ψ 设计 T（$\lambda = 0$）有助于在下游任务中学习全局不变多项式，而增强正则化 λ 有助于学习低次数多项式。实验详情见附录 E.2，图 11 展示了神经网络的类似行为。

在预训练任务中，我们对其施加了表示 ψ 的正交性。然而，最优表示 ψ 很可能对数据增强不变，这意味着理想情况下，当 $X \sim \mu_\Xi$ 或 $X \sim \rho_\mathcal{X}$时， $\psi(X)$应具有相同的分布，我们将其形式化地写作 $\psi_\# \mu_\Xi = \psi_\# \rho_\mathcal{X}$。此外，数据增强很可能会扩展输入数据分布，导致 $\rho_\mathcal{X} \ll \mu_\Xi$占据主导地位。这促使我们提出以下假设和定义。

假设 1 （低扩展性）。 存在 $c_r > 0$，使得对于原函数空间 f 中的任意函数 Ψ，其定义于（4），

$$\|f\|^2_{L^2(\rho_\mathcal{X})} \leq c_r \|f\|^2_{L^2(\mu_\mathcal{X})},$$

假设 2。 对于任意小于 i T_λ的正特征值数量的数，目标 f^* 在 L^2（μ_Ξ） 上关于 f_i的投影，与 L^2（$\rho_\mathcal{X}$） 中关于 f_i 的投影一致。

为了使这两个概念更加具体，下面给出三个示例。

示例 3. 如果 $\rho_\mathcal{X}$ 相对于 μ_Ξ 具有密度，且该密度在其支撑集上有下界 $\delta \in (0, 1]$，即 $\mu_\Xi = \delta\rho_\mathcal{X} + (1 - \delta)$ $\mu_\perp$ 对于 $\mu_\perp \in \Delta_\mathcal{X}$，则满足假设 *1*，且 $c_r = 1/\delta$。

示例 4。 令 $\Sigma_\tau = \mathbb{E}_{X\sim\tau}[\varphi(X)\,\varphi(X)^\top]$ 为 φ 在分布 τ 下的协方差矩阵。当存在c 使得 $\Sigma_{\rho_\mathcal{X}} \preceq c\Sigma_{\mu_\Xi}$（即 $c\Sigma_{\mu_\Xi} - \Sigma_{\rho_\mathcal{X}}$ 是正半定的）时，则假设 *1* 成立，其中 $c_r = c$。

示例 5。 如果 $\psi_\sharp \mu_\Xi = \psi_\sharp \rho_\mathcal{X}$ 对最优表示 $\psi = (f_i)$ 成立，其中 (f_i) 是 T_λ的正特征函数，并且存在一个可测函数 g： $\mathbb{R}^k \to \mathcal{Y}$使得 $f^* = g \circ \psi$，则假设 *2* 得到验证。

本质上，假设1 和2 允许纳入与原始数据并不相似的增强数据，前提是计算模型（假设1）和通过 VICReg 损失进行的训练（假设2）不会过度偏向这些异常的增强数据。示例3 指出，如果增强数据大多与原始样本相似，就不必担心计算模型引入的偏差。示例4 则基于二阶矩给出了更宽松的保证。最后，示例5 指出，如果学习到的表示将增强数据与其对应的原始样本混为一谈，就无需担心增强数据的特殊性。

5.2. 下游任务上的泛化

本节为有监督下游任务提供了全面的泛化保证。以下假设确保目标函数 $f^*(x) =$

$\mathbb{E}_\rho[Y|X=x]$ 所对应的下游任务能够由预训练问题充分表示。

假设 3 (源条件)。 f^* 属于 T_λ 的正特征子空间，即 $f^* \in$ 张成空间 $\{f_i \,|\, \lambda_i > 0\}$。

例6 (簇假设)。 如果密度 μ_Ξ 的支撑集 k 有连通分支，f^* 在这些簇上为常数，且 $\lambda = 0$，则假设 *3* 成立。

现在给出下游保证的简化版本。完整表述见附录 4 B中的定理。

定理 1 (下游误差)。 设 $(X_i,\ Y_i) \sim \rho^{\otimes n}$ 是 n从下游任务的分布中抽取的样本，且 ℓ 为平方损失。定义 $k_\lambda < +\infty$ 为 T_λ的严格正特征值个数。在假设 *1*、*2*和 *3* 成立的条件下，经过过渡阶段后，最优正则化经验风险最小化器的平均超额风险 f_n 为

$$\mathbb{E}[\mathcal{R}(f_n) - \mathcal{R}(f^*)] \leq \frac{2k_e\varepsilon^2}{n} + \frac{\log(n)^{1.1}}{n}\|f^*\|_{L^2(\rho)} + c_{f,T_\lambda}^2\left(\mathcal{L}_k(S\hat{\Theta}) - \mathcal{L}_k(S\Theta_*)\right) + c_{f,k}. \tag{6}$$

其中，ε^2 是 Y 的噪声水平（条件方差的上确界），$k_e \leq k$ 是表示 $\psi = \Theta\varphi$在下游任务上的有效维数，$c_{f,\,k} \leq (k_\lambda - k)_+ \|f^*\|^2_{L^2(\rho_\mathcal{X})}$是一个常数，用于描述目标函数 f^*在下游任务上的能量相对于 T_λ的特征子空间的集中程度；$c_{f,\,T_\lambda} \leq \left\|T_\lambda^{-1} f^*\right\|$是一个类似的常数，同时考虑了 T_λ的特征值衰减；而 $\mathcal{L}_k$ 中的索引 k 表示我们在 $\psi:\ \mathcal{X} \to \mathbb{R}^k$上进行搜索。

定理 1 的结果可视为方差-偏差分解。由模型设定错误的线性回归导致的方差项，其收敛速率为 $k\log(n)/n$。$\log(n)$这一因子实际上是推导过程中的产物，借助 Mourtada 等人（2022）的定理 1 可以将其消去。偏差项与近似误差有关，它通过常数$c_{f,\,T_\lambda}$和$c_{f,\,k}$反映了用 T_λ 学习 f^* 的难度，也反映了通过 $\mathcal{L} - \mathcal{L}^*$产生的预训练任务上的误差。请注意，定理 1 的证明有意避免用 $\left\|T_\lambda^{-1}\right\|_{\mathrm{op}}\|f^*\|$对$c_{f,\,T_\lambda}$作上界，因为这会在上界中引入 T_λ 的谱隙的倒数，而且无法准确描述目标函数 f^* 实际上很容易用 T_λ学习的情形。我们还指出，对于分类任务，近期研究表明，在对 $\rho_\mathcal{X}$作出温和假设且噪声水平较低的条件下，可以将定理 1 的收敛速率转化为零一损失的指数级快速收敛速率（Cabannes 等，2021b）。在所研究的簇设定下，这一点尤其成立。

HaoChen 等人（2021年）。[4] 定理1的理论收敛速率在图6中得到了实验验证。

[4]另请参见 Rigollet (2007，了解此设定下的快速收敛速率。

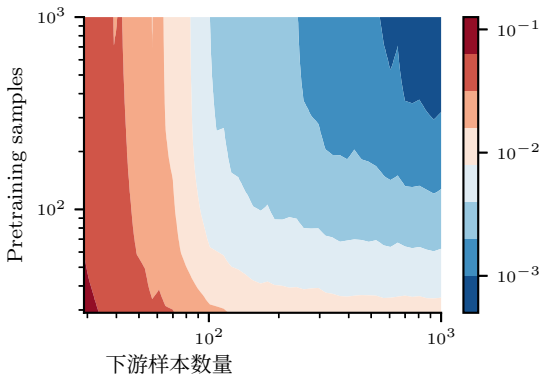

图 *6*。 经验下游性能在一个简单任务上（详见附录D）取决于下游样本数量（x轴）和预训练样本数量（y轴），且均以对数刻度表示。沿图的左侧，可以观察到 $n_{pre}^{-1/2}$相对于预训练样本数量的收敛速率（定理3）；沿图的顶部，可以观察到 n_{down}^{-1} 相对于下游样本数量的收敛速率（定理1）。在底部，可以观察到饱和现象：随着超额风险停滞在 $\mathcal{R}(\Pi_{\mathcal{F}} f^*) - \mathcal{R}(f^*) > 0$，增加下游样本并未带来明显收益。

5.3. 预训练保证

定理1指出，具有较小预训练损失的表征可以解决满足假设3 的下游任务，但并未说明找到此类表征的难度。本节旨在弥补这一不足。以下定理利用拉德马赫复杂度论证，详细说明经验风险最小化器的收敛速率。

定理 2 (经验风险最小化器)。 令 $\Theta_n \in \mathbb{R}^k \otimes \mathcal{H}$ 为基于数据集 $\mathcal{D}_n$、针对 $\mathcal{L}$ 的无偏正则化经验版本的最小化器。假设 $\mathcal{D}_n$ 由 n 个输入样本$(X_i) \sim \mu_\mathcal{X}^{\otimes n}$ 以及每个样本的 m 次数据增强$(\xi_{ij}) \sim \mu|_{X_i}^{\otimes m}$ 构成，则平均超额风险的上界为

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{L}(S\Theta_n)] - \mathcal{L}(S\Theta) \leq \frac{12\kappa^2 k}{\lambda\sqrt{n}}\left(1 + \frac{\kappa^2 k}{\lambda}\right), \tag{7}$$

其中 κ 是 $\|\varphi(X)\|$ 的上界。

请注意，定理的证明2 采用了一个较松的上界来估计经验风险的方差，这主要是因为难以处理样本的非可交换性 (ξ_{ij})。本质上，最小化 $\mathcal{L}$的难易程度取决于两方面：使用经验数据估计时 $\mathcal{L}$ 的方差（或执行随机梯度下降时随机梯度的方差），以及我们旨在寻找表征的空间大小 $\psi:\ \mathcal{X} \to \mathbb{R}^k$。若对 $\varphi(\xi)$的分布作出更强的假设（例如，数据形成簇，且 $(\xi \,|\, X)$ 的分布律在每个簇内不变），则可以证明，超额风险随数据增强数量变化的表现会好得多（例如，将 n替换为一个簇中的最小点数

乘以视图数量）。以下定理给出了随机梯度下降算法的收敛速率，以刻画这种潜在情形。基于凸优化文献的证明及技术细节详见附录 C。

定理 3 (更紧的界)。 存在一种可实现的算法，能够保证平均超额风险

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{L}(S\Theta_n)] - \mathcal{L}(S\Theta) \leq 3\kappa^2 c_\lambda c'_\lambda \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_\xi^2}{nm}} + \frac{4\kappa^6 c_\lambda^2}{n} \tag{8}$$

其中， $c_\lambda = 1+\kappa^2 k_\lambda/\lambda$、 $c'_\lambda = 1+k_\lambda^2/\lambda^2$、 k_λ 是 T_λ的正特征值数量， κ 是 $\|\varphi\|$的上界， σ_X与 $\mathbb{E}[\psi(\xi) \mid X]$ 的方差有关，而σ_ξ 与$(\xi \mid X)$的平均方差有关。此外，当 $K = SS^\top$ 或 $\varphi(\xi)$ 的协方差只有有限个正特征值时（例如， $\mathcal{X}$ 有限或 $\mathcal{H}$ 有限维），其中 c_K是与 K的条件数有关的常数，该上界可以收紧为

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{L}(S\Theta_n)] - \mathcal{L}(S\Theta) \leq \frac{4c_K^2 c_\lambda^2}{n}. \tag{9}$$

在HaoChen 等人(2021年所研究的设定中，我们强调，定理 3 保证的是 $O(n^{-1})$ 的收敛速率，而不是上游损失上的 $O(n^{-1/2})$ 。实际上，我们将 HaoChen 等人HaoChen 等人(2021年，定理 4.3) 中的收敛速率从 $n^{-1/2}$提升至 n^{-1}，适用于预训练和下游任务。

6. 实践见解

在本节中，我们将我们的理论与训练自监督学习算法时常见的现象联系起来，并根据我们的发现提出最佳实践。

6.1. 下游问题

对于下游问题，应区分两种情形。当可用的下游样本较少时，少样本学习需要较小的有效维数 k_e (6)，以降低估计误差并避免拟合噪声。可以通过减小 $\hat{}k_e$ （或等价地，减小 $\mathcal{F}$的容量）来实现，也可以通过降低表征维度 k 或对下游任务应用正则化来实现。有效维数与下游示例数量之间的这种理论权衡，在 He 和 Ozay(2020 年，图 6）中得到了实证说明。相反，当有大量数据可用于训练下游任务时，可以放心地增大表征维度 k 以降低近似误差。Garrido 等人(2022，图 1 ）在大规模数据集上尤其观察到了这一现象：随着 k 递增，有效维数 k_e 趋于一个极限，而下游性能则持续提升直到达到这一极限。值得注意的是，我们的理论解释了这一现象：由于 k_λ 有限，随着 k 递增，有效维数 k_e 将受限于 $k = k_\lambda$ 时的极限情形。[5]

6.2. 预训练问题

每个样本使用多次增强。定理3 展示了多次增强（例如多裁剪）如何使表示更快地收敛到最优状态 ψ。其中，经验风险的方差同时取决于 σ_X 输入变化所导致的方差，以及σ_ξ 数据增强后所得视图变化所导致的方差。对于每个样本使用多次增强，可以降低后一种方差并提升性能；在 Caron 等人（2020 年）引入多裁剪时就观察到了这一点。然而，当预处理数据的总量 $m \times n$ 保持不变时，通常最好处理更多输入，每个输入使用两个视图 $m = 2$，而不是处理较少输入并进行多次增强。这一发现与Bardes 等人（2022）的经验观察相符：如果有可用的新样本，新样本总是优于更多视图。

预训练中的容量问题。 定理 2 和3表明，如果没有正则化来限制计算模型的容量，就不能指望有意义地解决预训练任务。这可以用量 c_λ 来刻画：当 λ 趋近于零时，该量会趋于无穷大。实践中，这类与缺乏正则化相关的问题很常见。给定 $n \times m$ 上游样本 (ξ_{ij})，VICReg 的经验最小化可以通过用 $\sum_{ij} \delta_{(i,\xi_{i,j})}/nm$来近似 μ 。在这种设定下， T 是一个图的邻接矩阵，其连通分支的数量与输入 n的数量相同，详见附录 E。所有连通分支共同构成经验近似 T的一个极大特征向量，从而导致“坍塌”的表示 $\psi = \sum_j \delta_{\xi_{ij}}/m$。正则化会迫使优化器在空间 Ψ 内搜索表示，该空间将这些小簇混合起来，使语义特征函数得以涌现（示意见图7 ）。

6.3. 给实践者的指南

我们的理论研究提供了若干可能对自监督学习实践者有用的见解。下面列出其中几条。

避免坍缩。 常见的坍缩现象是指预训练最终拟合了噪声，而不是学习有用特征；这一问题可以通过多种方式加以解决。我们的理论提出：

- 降低模型容量，通过正则化（例如，提前停止），或使用更简单的架构 （例如，用浅层卷积神经网络替代多层感知机）。因此，

[5]请注意，如果没有正则化，$(1-\beta)I + \beta T$不是迹类的，因此随着 k 增加， k_e 将不会收敛。

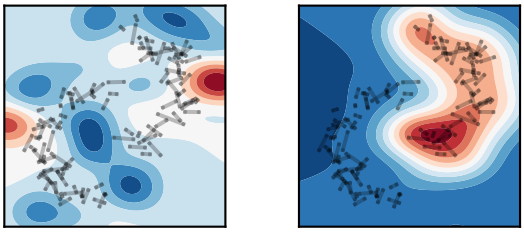

图7。容量问题。当正则化可以忽略（左图）和正则化较小 λ（右图）时， T_λ 经验估计的顶层特征函数的等高线。实验采用尺度约为问题直径十分之一的高斯核，增强以黑点表示；若增强来自同一输入，则用线连接 X。当 λ 可以忽略时，会出现容量问题，妨碍右图中簇结构的恢复。

Ψ 将具有较低的有效维数， K 将鼓励形成“更简单”的表示，这些表示所需的数据更少，即使完全不进行数据增强也能学会。

- 使用更强的数据增强。 T 将变得更紧，从而降低 k_λ T_λ“ 正特征空间 ” 的维数。理想的 ψ 会呈现更多结构，因此可以在更小的空间中进行搜索，从而更难发生坍塌。

纳入先验。表示通常用于解决下游任务，因此在预训练期间纳入恰当的先验至关重要。我们的理论展示了若干因素的重要作用。*(i)* 数据增强决定所施加的不变性的性质（例如，低变化、短程依赖、平移不变性）；影响 T 的顶层特征函数。*(ii)* 架构促进“简单”的表示（例如，平滑性、局部性）；影响 K 的顶层特征函数。*(iii)* 正则化平衡数据增强与架构之间的相互作用；影响 T_λ 的顶层特征函数。*(iv)* 预训练数据会影响 T 和 K 及其特征函数，例如通过聚类结构或自然图像统计特征产生影响。

7. 结论

本文提出了一个用于研究核区间内自监督学习的理论框架。本文考察了三个关键算子及其对收敛和泛化的影响： T 与增强相关， K 与架构选择相关，而 T_λ 则由二者的相互作用产生，并由参数 λ 调节。我们的分析为从业者提供了实用的理论保证和实践指南，以提升自监督学习算法的稳定性和性能。展望核区间之外的情形，未来工作可将我们的分析扩展至核设定之外，特别是理解有限宽度神经网络中的非线性训练动力学和特征学习。

层内的能力。此外，未来研究还可纳入更多有助于提升自监督学习性能的技术，例如在施加损失之前投影表征、对数据进行批处理，或采用不同的损失函数。

参考文献

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O. 和 Saunshi, N. 对对比式无监督表征学习的理论分析。载于*ICML*，2019年。

Bach, F. 利用凸神经网络打破维数灾难。《机器学习研究杂志》，18(1):629–681，2017年。

Bach, F.《第一性原理学习理论》。即将在麻省理工学院出版社出版，2023。

Balestriero, R. 和 LeCun, Y. 对比式与非对比式自监督学习可恢复全局与局部谱嵌入方法。载于神经信息处理系统大会，2022年。Bardes, A., Ponce, J. 和 LeCun, Y. VICReg：用于自监督学习的方差、不变性与协方差正则化。载于国际学习表征会议，2022年。Bartlett, P. 和 Mendelson, S. Rademacher复杂度与高斯复杂度：风险界与结构性结果。《机器学习研究杂志》，2002年。Bartlett, P., Jordan, M. 和 McAuliffe, J. 凸性、分类与风险界。《美国统计学会会刊》，2006年。Bietti, A. 深度卷积模型的近似与学习：核视角。载于国际学习表征会议，2022年。Bietti, A. 和 Mairal, J. 关于神经切线核的归纳偏置。《神经信息处理系统进展》，2019年。Bietti, A., Venturi, L. 和 Bruna, J. 关于不变性与几何稳定性条件下学习的样本复杂度。载于神经信息处理系统进展，2021年。Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. 等。语言模型是少样本学习者。载于神经信息处理系统进展，2020 年。

Bubeck, S.《凸优化：算法与复杂性》。《机器学习基础与趋势》，2015年。

Bun, J., Bouchaud, J.-P., 和 Potters, M.《大型相关矩阵清理：随机矩阵理论的工具》。*PhysicsReports*，2017年。

Cabannes, V., Pillaud-Vivien, L., Bach, F. 和 Rudi, A. 通过拉普拉斯正则化克服半监督学习中的维数灾难。发表于神经信息处理系统大会，2021a。

Cabannes, V., Rudi, A. 和 Bach, F. 结构化预测中的快速收敛率。发表于学习理论会议，2021b。

Cabannes, V., Bietti, A. 和 Balestriero, R. 关于无监督表示学习的最小变差。*ICASSP*，2023。

Caponnetto, A. 和 De Vito, E. 正则化最小二乘算法的最优收敛率。计算数学基础，2007。

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. 和 Joulin, A. 通过对比聚类分配进行视觉特征的无监督学习。发表于神经信息处理系统进展，2020 年。

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., 和 Joulin, A. 自监督视觉变换器中的新兴特性。载于*IEEE/CVF*国际计算机视觉会议论文集，2021年。

Chen, T., Kornblith, S., Norouzi, M., 和 Hinton, G. 用于视觉表征对比学习的简单框架。载于*ICML*，2020 年。

Chi, Y., Lu, Y., 和 Chen, Y. 非凸优化与低秩矩阵分解相遇：概述。*IEEE*信号处理汇刊，2019年。

Coifman, R. 和 Lafon, S. 扩散映射。应用与计算调和分析，2006年。

Davis, C. 和 Kahan, W. 扰动对特征向量的旋转。*SIAM*数值分析杂志，1970年。

Devlin, J., Chang, M.-W., Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向变换器预训练。载于*NAACL-HLT*论文集，2019年。

Efthimiou, C. 和 Frye, C. p 维空间中的球谐函数。世界科技出版社，2014年。

Favero, A., Cagnetta, F. 和 Wyart, M. 局部性在卷积教师-学生场景中战胜维数灾难。发表于神经信息处理系统进展，2021年。

Garrido, Q., Balestriero, R., Najman, L. 和 Lecun, Y. Rankme：通过预训练自监督表征的秩评估其下游性能。*arXiv* 预印本 *arXiv:2210.02885*，2022年。

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M. 等。自举自己的潜在表征——一种自监督学习的新方法。发表于神经信息处理系统进展，2020 年。

HaoChen, J., Wei, C., Gaidon, A. 和 Ma, T. 谱对比损失下自监督深度学习的可证明保证。载于神经信息处理系统大会，2021年。

HaoChen, J. Z. 和 Ma, T. 对比学习中归纳偏置的理论研究。*arXiv* 预印本 *arXiv:2211.14699*，2022年。

He, B. 和 Ozay, M. 探究自监督学习中坍塌特征与白化特征之间的差距。载于*ICML*，2020 年。

Jacot, A., Gabriel, F. 和 Hongler, C. 神经正切核：神经网络中的收敛与泛化。载于神经信息处理系统大会，2018。

Kato，T. 线性算子扰动理论。施普林格，1995。

Kiani, B. T., Balestriero, R., Chen, Y., Lloyd, S. 和 LeCun, Y. 核区间中的联合嵌入自监督学习。*arXiv* 预印本 *arXiv:2209.14884*，2022。

Kolmogorov, A. 和 Tikhomirov, V. 函数空间中集合的 ε-熵与 ε-容量。*Uspekhi Matem-aticheskikh Nauk*，1959。

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J. 和 Pennington, J. 宽神经网络在梯度下降下无论深度如何，都会演化为线性模型。载于神经信息处理系统大会，2019年。Lee, J., Lei, Q., Saunshi, N. 和 Zhuo, J. 预测你已经知道的内容会有所帮助：可证明的自监督学习。载于 神经信息处理系统进展，2021年。Lin, J., Rudi, A., Rosasco, L. 和 Cevher, V. 希尔伯特空间上最小二乘回归的谱算法最优收敛率。应用与计算调和分析，2020 年。Maurer, A. 拉德马赫复杂度的向量收缩不等式。载于算法学习理论，2016年。Mei, S., Misiakiewicz, T. 和 Montanari, A. 随机特征和核模型中的不变性学习。载于学习理论会议，2021年。

Meir, R. 和 Zhang, T. 贝叶斯混合算法的泛化误差上界。机器学习研究期刊，2003年。

Micchelli, C.、Xu, Y. 和 Zhang, H. 通用核。《机器学习》研究期刊，2006 年。

Misiakiewicz, T. 和 Mei, S. 核方法中的卷积与池化操作学习。载于神经信息处理系统进展，2022 年。

Mourtada, J. 和 Rosasco, L. 随机设计下岭回归的初等分析。《科学院院刊·数学》，2022 年。

Mourtada, J., skevi˘cius, T. V., and Zhivotovskiy, N. 分布无关稳健线性回归。数学统计与学习，2022。O'Donnell, R. 布尔函数分析。剑桥大学出版社，2014。Ostrovskii, D. and Bach, F. 使用自协调性对 m-估计量进行有限样本分析。《电子统计学杂志》，2018。Pillaud-Vivien, L. and Bach, F. 核化扩散映射。*arXiv*，2023。Pinelis, I. and Sakhanenko, A. 关于大偏差概率不等式的评述。《概率论及其应用》，1986。Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，2021。Rigollet, P. 簇假设下半监督分类的泛化误差上界。机器学习研究杂志，2007。Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. 理解对比学习需要纳入归纳偏置。载于*ICML*，2022。Schiebinger, G., Wainwright, M., and Yu, B. 核化谱聚类的几何。《统计年鉴》，2015。Scholkopf, B. and Smola, A. 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，2001。Simard, P., Victorri, B., LeCun, Y., and Denker, J. 切向传播——一种用于指定自适应网络中选定不变性的形式体系。载于神经信息处理系统大会，1991。

Simon, J., Knutins, M., Ziyin, L., Geisz, D., Fetterman, A., and Albrecht, J. 关于自监督学习的阶段性特征。*arXiv*，2023。

Smale, S. 和 Zhou, D.-X. 通过积分算子及其近似得到的学习理论估计。《构造性逼近》，2007 年。

Smola, A., Ovári, Z. 和 Williamson, R. C. 点积核正则化。载于神经信息处理系统进展，2000年。Sun, H.-W. 和 Zhou, D.-X. 与解析平移不变 Mercer 核相关联的再生核希尔伯特空间。《傅里叶分析与应用杂志》，2008年。Tian, Y. 理解非线性在对比学习训练动态中的作用。*arXiv* 预印本 *arXiv:2206.01342*，2022年。Tian, Y.、Yu, L.、Chen, X. 和 Ganguli, S. 用双深度网络理解自监督学习，2021年。Tosh, C.、Krishnamurthy, A. 和 Hsu, D. 对比学习、多视图冗余与线性模型。载于算法学习理论，2021a。Tosh, C.、Krishnamurthy, A. 和 Hsu, D. 对比估计揭示线性模型的主题后验信息。*JMLR*，2021b。Tropp, J. 矩阵集中不等式导论。《机器学习基础与趋势》，2015年。van Engelen, J. 和 Hoos, H. 半监督学习综述。《机器学习》，2020 年。Vitushkin, A. 关于希尔伯特第十三问题。《苏联科学院院刊》，1954年。Wen, Z. 和 Li, Y. 探索自监督对比学习中的特征学习过程。载于国际机器学习会议，2021年。Yang, G. 和 Salman, H. 神经网络的细粒度谱视角。*arXiv* 预印本 *arXiv:1907.10599*，2019年。

A. 数学细节与简明证明

A.1. 技术细节

正文中有几处技术细节未作明示，我们现在对此进行讨论。特别是，我们假设存在一个风险 f^* 的最小化器 $\mathcal{R}$；当 $\mathcal{Y}$有限时，或当 ℓ 是最小二乘损失且$(Y \mid X)$ 几乎处处具有二阶矩时，这一假设成立。此外，在最小二乘情形的证明中，为简单起见，我们将假设 $\mathcal{Y} = \mathbb{R}$。当 $\mathcal{Y} = \mathbb{R}^k$ 时，同样的推导仍然成立，不过需要稍加注意，例如在 $\mathcal{Y} \otimes \mathcal{H}$ 而非 $\mathcal{H}$中进行讨论（参见Cabannes 等人，2021b等文献）。

整数。 表征维度 k是一个整数，而 $[k]$ 表示集合 $\{1, 2, \cdots, k\}$。为简便起见，我们滥用符号，将严格正整数集合记作 $\mathbb{N}$。

几何。 乘积 $A \times B$ 表示由元素 (a, b) 构成的集合，其中 $a \in A$ 且 $b \in B$。记号 $a^\top$ 表示 a 的伴随算子，它取决于所考虑的希尔伯特度量空间；我们将 a 视为其中的一部分（例如， $L^2(\mu_\Xi)$中的伴随算子与 $L^2(\rho_\mathcal{X})$中的伴随算子并不相同）。记号 $a^\top b$表示标量积 $\langle a, b \rangle$，其中希尔伯特度量空间 a和 b均视为其中的一部分。矩阵或算子上的希尔伯特范数记作 $\|\cdot\|_F$（弗罗贝尼乌斯）、 $\|\cdot\|_2$ 或 $\|\cdot\|_{HS}$（希尔伯特–施密特）。算子范数记作 $\|\cdot\|_{\mathrm{op}}$。此外，恒等算子始终记作 I。

分布。 为了定义概率，假设 $\mathcal{X}$ 和 $\mathcal{Y}$ 是赋有博雷尔拓扑的波兰空间a。我们使用单纯形记号 Δ_A 来表示 A上的概率测度集合，并使用张量记号 $\rho^{\otimes n}$ 来表示 n 个相互独立且均服从 ρ的随机变量的测度。记号 $\varphi_\# \rho$ 表示当 X 服从测度 ρ时， $\varphi(X)$的测度。记号 $\rho \ll \mu$ 表示，对于任意可测集合 X，若 $\mu(X) = 0$，则 $\rho(X) = 0$。记号 δ_x 表示狄拉克分布，它满足 $\langle f, \delta_x \rangle = f(x)$，其中使用函数与分布之间的对偶配对。对于任意分布 p，空间 $L^2(p)$由平方可积的可测函数构成。

函数。 所有函数，例如 ℓ、 f、 ψ、 φ等，均限定为可测函数。记号 $\circ$ 表示函数复合 $f \circ g(\cdot) = f(g(\cdot))$。函数 $f \colon \mathcal{X} \to \mathcal{Y}$被视为 $\mathcal{Y}^\mathcal{X}$的一个元素，并且我们使用如下同构，例如 $(\mathbb{R}^k)^\mathcal{X} = (\mathbb{R}^\mathcal{X})^k$。我们使用记号 $\mathbb{R}^k \otimes \mathcal{H}$ 表示从 $\mathcal{H}$ 到 $\mathbb{R}^k$的有界线性算子。这种张量积记号推广了带有 $\mathbb{R}^k \otimes \mathbb{R}^{d_h} = \mathbb{R}^{k \times d_h}$的矩阵记号。特别地，

$$\Psi^k = \left\{ x \to \Theta \varphi(x) \,\middle|\, \Theta \in \mathbb{R}^k \otimes \mathcal{H} \right\}.$$

对于 $\Theta \in \mathbb{R}^k \otimes \mathcal{H}$，可以将 Θ 写成行向量形式，作为 $\mathcal{H}^k$ 中的一个元素；其伴随算子 $\Theta^\top \in \mathcal{H} \otimes \mathbb{R}^k$ 则可写成列向量形式。这是因为赋予 $\mathcal{H}^k$ ℓ^2-积拓扑后，它是自伴的。

A.2. 备注1的证明

为了便于使用无偏随机梯度实现（2），我们先对其进行刻画。我们需要将期望移到范数之外。可以通过以下推导做到这一点:

$$\begin{aligned}
\left\| \mathbb{E}[\psi(\xi)\psi(\xi)^\top] - I \right\|^2 &= \mathrm{Tr}\left((\mathbb{E}[\psi(\xi)\psi(\xi)^\top] - I)(\mathbb{E}[\psi(\xi')\psi(\xi')^\top] - I) \right) \\
&= \mathbb{E}_{\xi,\xi'}\left[\mathrm{Tr}\left(\psi(\xi)\psi(\xi)^\top \psi(\xi')\psi(\xi')^\top \right) \right] - 2\,\mathbb{E}_\xi\left[\mathrm{Tr}\left(\psi(\xi)\psi(\xi)^\top \right) \right] + \mathrm{Tr}(I) \\
&= \mathbb{E}_{\xi,\xi'}\left[(\psi(\xi')^\top \psi(\xi))^2 \right] - 2\,\mathbb{E}_\xi\left[\|\psi(\xi)\|^2 \right] + k.
\end{aligned}$$

对于第一部分，我们得到

$$\mathbb{E}_X\, \mathbb{E}_{\xi,\xi'}\left[\|\psi(\xi) - \psi(\xi')\|^2 \,\middle|\, X \right] = 2\,\mathbb{E}_\xi[\|\psi(\xi)\|^2] - 2\,\mathbb{E}_X\, \mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top \psi(\xi') \,\middle|\, X \right]$$

因此，

$$\mathcal{L}(\psi; \beta) = 2(\beta - 1)\,\mathbb{E}_\xi[\|\psi(\xi)\|^2] - 2\beta\,\mathbb{E}_X\, \mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top \psi(\xi') \,\middle|\, X \right] + \mathbb{E}_{\xi,\xi'}\left[(\psi(\xi')^\top \psi(\xi))^2 \right] + k. \tag{10}$$

特别 cular, 当 $\beta = 1$时，我们得到由 HaoChen 等人 （2021年 ），

$$\mathcal{L}(\psi) = -2\,\mathbb{E}_X\, \mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top \psi(\xi') \,\middle|\, X \right] + \mathbb{E}_{\xi,\xi'}[(\psi(\xi)^\top \psi(\xi'))^2] + k.$$

A.3. 引理 2 的证明

首先，注意到如果我们对 ψ 定义： $\mathcal{X} \to \mathbb{R}$，则映射 $\omega(\psi) \to \mathbb{E}_X[\mathbb{E}_{\xi,\xi'}\left[\|\psi(\xi) - \psi(\xi)\|^2 \,\middle|\, X\right]$, ω 是 $L^2(\mu_\Xi)$ 上的二次型。因此，它可以表示为作用于 $L^2(\mu_\Xi)$ 上的线性自伴算子 L ，使得$\omega(\psi) = \langle \psi, L\psi \rangle_{L^2(\mu_\Xi)}$ 。由于 $\omega(\psi) \geq 0$，我们有 $L \succeq 0$（其中 $\succeq$ 洛文纳序指对称算子上的序，即 $A - B$ 为正值时， $A \succeq B$）。以下引理表明 L 有界。

引理 5. 对于任意 $\psi \in L^2(\mu_\Xi)$， $\omega(\psi) \leq 2\|\psi\|^2_{L^2(\mu_\Xi)}$。由此可得， $L \preceq 2I$。

证明。 这是由以下事实得出的：

$$\begin{aligned}
\omega(\psi) &= \mathbb{E}_X[\mathbb{E}_{\xi,\xi'}\left[\|\psi(\xi) - \psi(\xi')\|^2 \,\middle|\, X\right]] \\
&= \mathbb{E}_X[\mathbb{E}_{\xi,\xi'}\left[\|\psi(\xi) - \mathbb{E}[\psi(\xi) \mid X] + \mathbb{E}[\psi(\xi) \mid X] - \psi(\xi')\|^2 \,\middle|\, X\right]] \\
&= \mathbb{E}_X[\mathbb{E}_{\xi,\xi'}\left[\|\psi(\xi) - \mathbb{E}[\psi(\xi) \mid X]\|^2 + \|\mathbb{E}[\psi(\xi) \mid X] - \psi(\xi')\|^2 \,\middle|\, X\right]] \\
&= 2\,\mathbb{E}_X[\mathbb{E}_{\xi}\left[\|\psi(\xi) - \mathbb{E}[\psi(\xi) \mid X]\|^2 \,\middle|\, X\right]] \\
&= 2 \min_{\psi_0 : \mathcal{X} \to \mathbb{R}} \mathbb{E}_X[\mathbb{E}_{\xi}\left[\|\psi(\xi) - \psi_0(X)\|^2 \,\middle|\, X\right]] \\
&\leq 2\,\mathbb{E}_X[\mathbb{E}_{\xi}\left[\|\psi(\xi)\|^2 \,\middle|\, X\right]] = 2\,\mathbb{E}_{\xi}\left[\|\psi(\xi)\|^2\right] = 2\|\psi\|_{L^2(\mu_\Xi)}
\end{aligned}$$

因此，对于任意 ψ，在 $L^2(\mu_\Xi)$ 几何下，我们有 $\psi^\top L\psi \leq \psi^\top\psi$。由于 L 是自伴的，这意味着 $\|L\|_{\mathrm{op}} \leq 2$。 □

由于 $0 \preceq L \preceq 2I$，我们引入 $T = (2I - L)/2$，得到 $0 \preceq T \preceq 1$；并且在 $L^2(\mu_\Xi)$ 几何下，对于ψ： $\mathcal{X} \to \mathbb{R}^k$

$$\mathcal{L}(\psi) = \beta \sum_{i=1}^{k} \psi_i^\top L\psi_i + \left\|\mathbb{E}[\psi(\xi)\psi(\xi')^\top] - I\right\|^2 = 2\beta \sum_{i=1}^{k} \psi_i^\top (I - T)\psi_i + \left\|\mathbb{E}[\psi(\xi)\psi(\xi')^\top] - I\right\|^2.$$

为了在不依赖谱定理的积分形式的情况下对所有算子进行对角化，我们引入以下温和的假设。

假设 4. 假设 T 具有纯点谱。

例 7. 当数据增强分布相对于某个测度具有密度 p ，且$(x, \xi) \to p(\xi \mid x)/p(\xi)$ 属于 $L^2(\mu)$时，或者当 $\mathcal{X}$是有限的， T可以证明是一个紧算子，因此根据谱定理具有纯点谱。

证明。 当 $\mathcal{X}$ 是有限的， L^2 空间是有限维的，因此是局部紧的，这意味着所有算子都是紧的。为了证明密度情形，我们将 T 表示为积分算子。在 $L^2(\mu_\Xi)$几何中，对于$f : \mathcal{X} \to \mathbb{R}$，有：

$$\begin{aligned}
2f^\top(I - T)f &= \mathbb{E}_X\,\mathbb{E}\left[\|f(\xi) - f(\xi')\|^2 \,\middle|\, X\right] = \mathbb{E}_X\,\mathbb{E}\left[\|f(\xi)\|^2 + \|f(\xi')\|^2 - 2\langle f(\xi), f(\xi')\rangle \,\middle|\, X\right] \\
&= 2f^\top f - 2\,\mathbb{E}_X\,\mathbb{E}\left[\langle f(\xi), f(\xi')\rangle \mid X\right].
\end{aligned}$$

这使我们能够 将其与内积对应起来，对于 g，我们有： $\mathcal{X} \to \mathbb{R}$，而 p 是数据增强的密度 s

$$\begin{aligned}
f^\top Tg &= \mathbb{E}_X\,\mathbb{E}\left[\langle f(\xi), g(\xi')\rangle \mid X\right] = \int \langle f(\xi), g(\xi')\rangle\, p(\xi \mid x)\, p(\xi' \mid x)\, \mathrm{d}\xi'\, \mathrm{d}\xi \mu_{\mathcal{X}}(\mathrm{d}x) \\
&= \int \mu_\Xi(\mathrm{d}\xi) \left\langle f(\xi), \int \mu_\Xi(\mathrm{d}\xi') g(\xi') \frac{\int \mu_{\mathcal{X}}(\mathrm{d}x) p(\xi \mid x)\, p(\xi' \mid x)}{p(\xi)p(\xi')} \right\rangle.
\end{aligned}$$

A因此，可以将 T 视为与核相关联的 $L^2(\mu_\Xi)$中的积分算子

$$k(\xi,\xi')=\frac{\int \mu_{\mathcal{X}}(\mathrm{d}x)p\left(\xi\,|\,x\right)p\left(\xi'\,|\,x\right)}{p(\xi)p(\xi')}.$$

当这个核　有界时，或者简单地说，当 $\xi\to k(\xi,\ \xi)$ 属于 $L^2(\mu_\Xi)$ 时，T 是迹类，因此是紧的　t.　□

现在证明，为了最小化 $\mathcal{L}$，应选取算子 $(1-\beta)I+\beta T$ 对应特征值为最大正值的特征函数。利用某个较抽象空间中的简单几何方法即可证明这一点。为此，注意到 $\psi\colon\ \mathcal{X}\to\mathbb{R}^k$ 在 $L^2(\mu_\Xi,\ \mathcal{X},\ \mathbb{R}^k)$ 中可表示为 $\tilde{\psi}\in\mathbb{R}^k\otimes L^2(\mu_\Xi,\ \mathcal{X},\ \mathbb{R})$，其中线性映射将 $(\psi_i^\top\varphi)$ 映射为函数 $\varphi\in L^2(\mu_\Xi,\ \mathcal{X},\ \mathbb{R})$。令 $T_\beta=(1-\beta)I+\beta T$，上游损失可表示为

$$\begin{aligned}
\mathcal{L}(\psi)&=2\beta\sum_{i\in[k]}\langle\psi_i,(I-T)\psi_i\rangle+\left\|\mathbb{E}_\xi[\psi(\xi)\psi(\xi)^\top]-I\right\|^2\\
&=2\beta\sum_{i\in[k]}\left\langle e_i^\top\psi,(I-T)\psi^\top e_i\right\rangle+\left\|\mathbb{E}_\xi[\sum_{i,j\in[k]}e_i^\top\psi(\xi)\psi(\xi)^\top e_je_ie_j^\top]-I\right\|^2\\
&=2\beta\sum_{i\in[k]}e_i\tilde{\psi}(I-T)\tilde{\psi}^\top e_i+\left\|\sum_{i,j\in[k]}e_i^\top\tilde{\psi}\tilde{\psi}^\top e_je_ie_j^\top-I\right\|^2\\
&=2\beta\,\mathrm{Tr}\left(\tilde{\psi}(I-T)\tilde{\psi}^\top\right)+\left\|\tilde{\psi}\tilde{\psi}^\top-I\right\|^2=2\beta\,\mathrm{Tr}\left(\tilde{\psi}(I-T)\tilde{\psi}^\top\right)+\mathrm{Tr}\left(\left(\tilde{\psi}\tilde{\psi}^\top-I\right)^2\right).\\
&=\mathrm{Tr}\left(2\beta\tilde{\psi}(I-T)\tilde{\psi}^\top+\tilde{\psi}\tilde{\psi}^\top\tilde{\psi}\tilde{\psi}^\top-2\tilde{\psi}\tilde{\psi}^\top+I\right).\\
&=\mathrm{Tr}_{L^2(\mu_\Xi)}\left(\tilde{\psi}^\top\tilde{\psi}\tilde{\psi}^\top\tilde{\psi}+(2\beta(I-T)-2I)\tilde{\psi}^\top\tilde{\psi}\right)+k.\\
&=\mathrm{Tr}_{L^2(\mu_\Xi)}\left(\tilde{\psi}^\top\tilde{\psi}\tilde{\psi}^\top\tilde{\psi}-2T_\beta\tilde{\psi}^\top\tilde{\psi}\right)+k.=\mathrm{Tr}_{L^2(\mu_\Xi)}\left((\tilde{\psi}^\top\tilde{\psi}-T_\beta)^2-T_\beta^2\right)+k.
\end{aligned}$$

为了利用这一新表述求得 $\mathcal{L}$ 的最小化器，此处需要稍加注意，因为这两个算子都不是迹类。下面的引理将处理这些注意事项，以完成证明。

引理 6. 设 A 是作用在 $L^2(\mu_\Xi)$上的自伴算子。假设存在 c 使得 $A\preceq cI$，且 A 是纯点谱。那么，若 (λ_i,f_i) 表示 A 的特征分解，且 λ_i 按递减顺序排列，则在约束 B 为秩至多为 k的自伴正算子的条件下，$\mathrm{Tr}\left((B-A)^2-B^2\right)$ 的最小值在$B=\tilde{\psi}^\top\tilde{\psi}$ 处取得，其中 $\psi\colon\ \mathcal{X}\to\mathbb{R}^k$ 满足 $\psi_i=\max(0,\lambda_i)^{1/2}f_i$。

证明。将 A 分解为正部 $A_+\succeq 0$ 和负部 $A_-\succeq 0$，使得 $A=A_+-A_-$。利用 B 是正的自伴算子这一事实，我们得到

$$\begin{aligned}
\mathrm{Tr}\left((B-A)^2-A^2\right)&=\mathrm{Tr}\,B^2-2B^{1/2}AB^{1/2}=\mathrm{Tr}\left(B^2-2B^{1/2}A_+B^{1/2}\right)+2\,\mathrm{Tr}\left(B^{1/2}A_-B^{1/2}\right)\\
&\geq\mathrm{Tr}\left(B^2-2B^{1/2}A_+B^{1/2}\right).
\end{aligned}$$

我们将 B 分解为秩至多为 1 的 k 个对称算子，如 $B=\sum_{i=1}^k B_i$ 所示，使得 $B_iB_j=0$ 对任意

$i \neq j \in [k]$。利用所引入算子的不同性质，我们继续

$$\begin{aligned}
\operatorname{Tr}\left((B-A)^2 - A^2\right) &\geq \sum_{i=1}^{k} \operatorname{Tr}\left(B_i^2\right) - 2\operatorname{Tr}\left(B_i A_+\right) = \sum_{i=1}^{k} \|B_i\|_{\mathrm{op}}^2 - 2\|B_i A_+\|_{\mathrm{op}} \\
&\geq \sum_{i=1}^{k} \|B_i\|_{\mathrm{op}}^2 - 2\|B_i\|_{\mathrm{op}} \|\Pi_{B_i} A_+\|_{\mathrm{op}} \geq \sum_{i=1}^{k} \|B_i\|_{\mathrm{op}}^2 - 2\|B_i\|_{\mathrm{op}} \Big\| \prod_{j<i} (I - \Pi_{B_j}) A_+ \Big\|_{\mathrm{op}} \\
&= \sum_{i=1}^{k} \left(\|B_i\|_{\mathrm{op}} - \Big\| \prod_{j<i} (I - \Pi_{B_j}) A_+ \Big\|_{\mathrm{op}} \right)^2 - \Big\| \prod_{j<i} (I - \Pi_{B_j}) A_+ \Big\|_{\mathrm{op}}^2 \\
&\geq -\sum_{i=1}^{k} \Big\| \prod_{j<i} (I - \Pi_{B_j}) A_+ \Big\|_{\mathrm{op}}^2 \geq -\sum_{i=1}^{k} \sigma_i(A_+)
\end{aligned}$$

其中， Π_B 表示到 B像空间上的正交投影算子， $\sigma_i(A)$表示 A 的第 i个奇异值（按单调顺序排列，其中 $\sigma_1(A)$最大）。最后一个不等式由 Courant-Fisher 极小极大原理得出。令 Π_{B_i} 为投影到 A 的第 i个特征子空间上的投影，且 $\|B_i\|_{\mathrm{op}} = \sigma_i(A)$，即可达到此不等式。换言之， B 应与 A的前 k 个正特征值相匹配。如果 A 的正特征值少于 k 个，那么 B 应与所有正特征值相匹配，并在 A_-的值域上为零。 □

以下是上述证明的直接推论。

命题 7 （极小化者的唯一性）。 $\mathcal{L}$ 的极小化者在正交变换和特征函数选取意义下唯一。更具体地说，若 $U \in \mathbb{R}^{k\times k}$是正交的，即 $U^\top U = I$，则 $\mathcal{L}(\psi) = \mathcal{L}(U\psi)$ ；而若$\lambda_k = \lambda_{k+1}$，则可以在 T_β的特征分解（λ_i, f_i） 中选择不同的特征函数作为 f_k 。

A.4. 引理3的证明

考虑 $\psi\colon \mathcal{X} \to \mathbb{R}^k$，其中 $\psi_i = f_{\theta_i}$ 对于 $\theta_i \in \mathcal{H}$ 和 $S\colon \mathcal{H} \to L^2(\mu_\Xi)$ ； $\theta \to \theta^\top \varphi(\cdot)$ 。我们可以使用前面介绍的张量记号对 $\psi = S\Theta^\top$ 进行参数化，其中 $\Theta = (\theta_i)_{i\in[k]}$ 被视为 $\mathbb{R}^k \otimes \mathcal{H}$的一个元素。引理 3的证明可由以下事实直接得到:

$$\|\Theta\|_2^2 = \left\|\Theta^\top\right\|_2^2 = \left\langle S\Theta^\top, (S^\top S)^{-1} S\Theta^\top \right\rangle_{L^2(\mu_\Xi)} = \sum_{i=1}^{k} S\theta_i^\top K^{-1} S\theta_i = \sum_{i=1}^{k} \psi_i^\top K^{-1} \psi_i.$$

由于 $\Psi = S(\mathcal{H}) = \operatorname{im} S = \operatorname{im} K^{1/2} = K^{1/2}(L^2(\mu_\Xi))$，可以将 K^{-1}视为 K的逆，使得对于$\psi_i \in$ 核 K， $\psi_i^\top K^{-1}\psi_i = +\infty$。这正是我们在正文中隐含采用的假设，因此（$\psi_i$）都属于$\Psi$。需要注意的是，在许多情况下， Ψ 在 $L^2(\mu_\Xi)$ 中稠密（Micchelli 等，2006），因此无需采取这种预防措施，因为核 $K = \{0\}$，且在 $L^2(\mu_\Xi)$ 上定义 K^{-1} 的方式只有一种。

A.5. 引理3的第二个证明：协方差算子

上文给出的引理3 证明对于不熟悉再生核希尔伯特空间的读者来说可能显得相当抽象。本小节将基于协方差算子，给出一个较为易懂的证明。

重新利用 $\mathcal{L}$ 的无偏刻画，我们有

$$\begin{aligned}
\mathcal{L}(\psi;\beta) &= 2(\beta-1)\,\mathbb{E}_\xi[\|\psi(\xi)\|^2] - 2\beta\,\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top\psi(\xi') \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}\left[(\psi(\xi')^\top\psi(\xi))^2\right] + k. \\
&= 2(\beta-1)\operatorname{Tr}\left(\mathbb{E}_\xi[\psi(\xi)\psi(\xi)^\top]\right) - 2\beta\operatorname{Tr}\left(\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\psi(\xi))\psi(\xi')^\top \,\middle|\, X\right]\right) + \operatorname{Tr}\left(\mathbb{E}_\xi\left[\psi(\xi)\psi(\xi)^\top\right]^2\right) + k,
\end{aligned}$$

其中，最后一项源于以下事实:

$$\begin{aligned}
\mathbb{E}_{\xi,\xi'}\left[(\psi(\xi)^\top\psi(\xi'))^2\right] &= \mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top\psi(\xi')\psi(\xi')^\top\psi(\xi)\right] = \mathbb{E}_{\xi,\xi'}\left[\operatorname{Tr}\left(\psi(\xi')\psi(\xi')^\top\psi(\xi)\psi(\xi)^\top\right)\right] \\
&= \operatorname{Tr}\left(\mathbb{E}_\xi[\psi(\xi)\psi(\xi)^\top]\,\mathbb{E}_{\xi'}[\psi(\xi')\psi(\xi')^\top]\right) = \operatorname{Tr}\left(\mathbb{E}_\xi[\psi(\xi)\psi(\xi)^\top]^2\right).
\end{aligned}$$

A.5.1. 算子技术细节

我们将以如下形式 $\Theta\varphi$ 对 $\Theta \in \mathbb{R}^k \otimes \mathcal{H}$ 和 φ 搜索 $\psi\colon\ \mathcal{X} \to \mathcal{H}$。下面讨论与即将出现的无限维算子相关的技术细节。

假设5。 希尔伯特空间 $\mathcal{H}$ 是可分的，且映射 φ 属于 $L^2(\mu_{\mathcal{X}})$ 其中 $\mathcal{X}$ 和 $\mathcal{H}$均赋有博雷尔拓扑。

注8。 算子 $\Sigma = \mathbb{E}_\xi[\varphi(\xi)\varphi(\xi)^\top] \in \mathcal{H} \otimes \mathcal{H}$ 是迹类的。

证明。这由迹的线性性、期望的线性性，以及 $\operatorname{Tr} AB = \operatorname{Tr}\ BA$ 这一事实得出 ，

$$\operatorname{Tr}\Sigma = \mathbb{E}_\xi \operatorname{Tr}\varphi(\xi)\varphi(\xi)^\top = \|\varphi\|^2_{L^2(\mu_{\mathcal{X}})} < +\infty.$$

因此， Σ 是紧的，故具有纯点谱；又由于 $\mathcal{H}$ 是可分的，它可以被对角化，其特征向量构成 $\mathcal{H}$的一组基。

□

我们稍后会看到， $\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}$ 确实与 T等距。因此，在假设4下， $\Sigma^{-1/2}\Sigma_{\mathcal{X}}\Sigma^{-1/2}$ 具有纯点谱。不过，下面的引理表明，无需借助 $T \preceq I$这一事实，该算子也是有界的。

注9。算子$\Sigma_X = \mathbb{E}_X\, \mathbb{E}_{\xi,\ \xi'} \left[\varphi(\xi)\varphi(\xi')^\top \,\middle|\, X\right] \in \mathcal{H} \otimes \mathcal{H}$ 在洛文纳序下满足 $0 \preceq \Sigma_X \preceq \Sigma$，其中 $\preceq$ （若 $B - A$ 是半正定的）。因此， Σ_X 是迹类，且 $\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}$ 是连续的。

证明。 这是由詹森不等式应用于 $A \to AA^\top$ 得出的，而这可以通过协方差算子的正定性来证明。

$$0 \preceq \mathbb{E}[(A - \mathbb{E}[A])(A - \mathbb{E}[A])^\top], \qquad \Rightarrow \qquad \mathbb{E}[A]\,\mathbb{E}[A]^\top \preceq \mathbb{E}[AA^\top].$$

因此，

$$\mathbb{E}_{\xi,\xi'}\left[\varphi(\xi)\varphi(\xi')^\top \,\middle|\, X = x\right] \preceq \mathbb{E}_\xi\left[\varphi(\xi)\varphi(\xi)^\top \,\middle|\, X = x\right],$$

这意味着

$$\Sigma_X \preceq \Sigma.$$

因此，迹 $\Sigma_X \le$迹 $\Sigma < +\infty$ 和 $\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2} \preceq I$，以及 $\left\|\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}\right\|_{\mathrm{op}} \le 1$。其正定性源于 Σ_X 是一个协方差算子 $\Sigma_X = \mathbb{E}_X\left[\mathbb{E}_\xi\left[\varphi(\xi) \mid X\right]\mathbb{E}_\xi\left[\varphi(\xi) \mid X\right]^\top\right]$. 这一事实 □

A.5.2. 算子表述

我们先证明引理的一个变体，其中所有内容都用 $\mathcal{H}$ 表示。稍后我们将详细讨论 $\mathcal{H}$ 与 $L^2(\mu_\Xi)$ 之间的等距同构（这是由 S 与 $\Sigma^{1/2}$ 之间的等距同构得出的），它使我们能够将该变体转化为论文中所写的引理。

引理10. 对于$(\theta_i) \in \mathcal{H}^k$ 和 $f_\theta\colon\ x \to \langle\varphi(x)\ \theta\rangle$，以及一个正则化项 $\lambda \in \mathbb{R}$

$$\mathcal{L}((f_{\theta_i})_{i\in[k]}) + \lambda\sum_{i\in[k]}\|\theta_i\|_2^2 = \operatorname{Tr}\left(\left(\Sigma^{1/2}(\sum_{i\in[k]}\theta_i\theta_i^\top)\Sigma^{1/2} - A)^2 - A^2\right) + k,\right.$$

其中 A 和 Σ是定义在 $\mathcal{H}$ 上的算子，定义如下

$$A = \Sigma^{-1/2}((1-\beta)\Sigma + \beta\Sigma_X - \lambda I)\Sigma^{-1/2}, \quad \Sigma = \mathbb{E}_\xi\left[\varphi(\xi)\varphi(\xi)^\top\right], \quad \Sigma_X = \mathbb{E}_X[\mathbb{E}_{\xi,\xi'}\left[\varphi(\xi)\varphi(\xi')^\top \,\middle|\, X\right]].$$

因此， $\mathcal{L}$ 的一个最小化器 Θ_* 满足： Θ_* 与 A 在正特征值上的特征值分解相匹配，直到第 k 个。形式上，若 $A = \sum_{i\in\mathbb{N}}\lambda_i u_i \otimes u_i$，其中 $u_i \in \mathcal{H}$，且(λ_i) 按降序排列，

$$\Theta_* = (\theta_i)_{i\in[k]},\ \textit{with}\ \theta_i = \sqrt{\max(\lambda_i, 0)}\Sigma^{-1/2}u_i.$$

此外，(f_{θ_i}) 在 $L^2(\mu_\Xi)$中是正交的，其中 μ_Ξ 表示增强上的边缘分布。

证明。 现在我们 根据参数化 $\psi = \Theta\varphi$重写 $\mathcal{L}$ 中出现的不同量。我们有

$$\operatorname{Tr}\left(\mathbb{E}[\psi(\xi)\psi(\xi)^\top]\right) = \operatorname{Tr}\left(\mathbb{E}[\Theta\varphi(\xi)\varphi(\xi)^\top\Theta^\top]\right) = \operatorname{Tr}\left(\Theta\,\mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\Theta^\top\right) = \operatorname{Tr}\left(\Theta\Sigma\Theta^\top\right) = \operatorname{Tr}\left(\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2}\right).$$

伴随算子 $\Theta^\top$ 是相对于 $\mathcal{H}$ 和 $\mathbb{R}^k$上的典范拓扑定义的。同样,

$$\operatorname{Tr}\left(\mathbb{E}_X\,\mathbb{E}\left[\psi(\xi)\psi(\xi')^\top \,\middle|\, X\right]\right) = \operatorname{Tr}\left(\Theta\Sigma_X\Theta^\top\right) = \operatorname{Tr}\left(\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2}\right).$$

对于最后一项，我们得到

$$\operatorname{Tr}\left(\mathbb{E}[\psi(\xi)\psi(\xi)^\top]^2\right) = \operatorname{Tr}\left((\Theta\Sigma\Theta^\top)^2\right) = \operatorname{Tr}\left(\Theta\Sigma\Theta^\top\Theta\Sigma\Theta^\top\right) = \operatorname{Tr}\left(\Sigma^{1/2}\Theta^\top\Theta\Sigma\Theta^\top\Theta\Sigma^{1/2}\right)$$

将各项合并后，我们得到

$$\begin{aligned}
&\mathcal{L}(\Theta\varphi) + 2\lambda\operatorname{Tr}(\Theta^\top\Theta) - k \\
&= \operatorname{Tr}\left(2(\beta-1)\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2} - 2\beta\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2}\right. \\
&\qquad\qquad \left. + \Sigma^{1/2}\Theta^\top\Theta\Sigma\Theta^\top\Theta\Sigma^{1/2} + 2\lambda\Sigma^{-1}\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2}\right) \\
&= \operatorname{Tr}\left(\left(\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2} + (\beta-1)I - \beta\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2} + \lambda\Sigma^{-1}\right)^2 - \left((\beta-1)I - \beta\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2} + \lambda\Sigma^{-1}\right)^2\right) \\
&= \operatorname{Tr}\left(\left(\Sigma^{1/2}\Theta^\top\Theta\Sigma^{1/2} - \Sigma^{-1/2}((1-\beta)\Sigma + \beta\Sigma_X - \lambda)\Sigma^{-1/2}\right)^2 - \left(\Sigma^{-1/2}((1-\beta)\Sigma + \beta\Sigma_X - \lambda)\Sigma^{-1/2}\right)^2\right).
\end{aligned}$$

这就证明了引理的第一部分。值得注意的是，引理中的表达式与推广到连续 $\mathcal{X}$的形式略有不同，后者由 HaoChen 等人 (2021年) 在其附录 F 中提出，并将沿用 Schiebinger 等人(2015) 的工作，考虑带有特征 $\bar{\varphi}(x) = q^{-1/2}(x)\,\mathbb{E}\left[\varphi(\xi)\mid X = x\right]$ 其中 $q\colon\ x \to \mathbb{E}_{X\sim\mu_\Xi}[k(x,\ X)]$ 而不是 $\Sigma^{-1/2}\Sigma_X\Sigma^{-1/2}$。

最后，让我们证明 f_{θ_i} 在 L^2 中正交，我们有

$$\begin{aligned}
\left\langle f_{\theta_i}, f_{\theta_j}\right\rangle_{L^2(\mu_\Xi)} &= \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}\,\mathbb{E}[\left\langle\Sigma^{-1/2}u_i,\varphi(\xi)\right\rangle\left\langle\Sigma^{-1/2}u_i,\varphi(\xi)\right\rangle] \\
&= \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}\,\mathbb{E}[u_i^\top\Sigma^{-1/2}\varphi(\xi)\varphi(\xi)^\top\Sigma^{-1/2}u_j] \\
&= \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}u_i^\top\Sigma^{-1/2}\,\mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\Sigma^{-1/2}u_j \\
&= \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}u_i^\top\Sigma^{-1/2}\Sigma\Sigma^{-1/2}u_j \\
&= \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}u_i^\top u_j = \sqrt{\max(\lambda_i,0)\max(\lambda_j,0)}\delta_{ij}.
\end{aligned}$$

这证明了 f_{θ_i} 在 L^2（μ_Ξ）中的正交性。 □

A.5.3. 等距表述

考虑 $S\colon\ \mathcal{H} \to L^2$（$\mu_\Xi$）； $\theta \to f_\theta$ 再生核希尔伯特空间在 L^2（μ_Ξ）中的嵌入。

引理 11. *S 与 $\Sigma^{1/2}$ 等距，且 $K = S^\top S$ 是一个积分算子，它将 $f \in L^2(\mu_\Xi)$ 映射到 $Kf \in L^2(\mu_\Xi)$，定义为 $\xi \in \mathcal{X}$*

$$Kf(\xi) = \mathbb{E}_{\xi'}\left[\varphi(\xi)^\top\varphi(\xi')f(\xi')\right]. \tag{11}$$

证明. 这由此可得 由以下事实可知， S和 $\Sigma^{1/2}$ 都是 Σ的平方根。事实上， $\Sigma = S^\top S$，因为对于 $\theta \in \mathcal{H}$

$$\begin{aligned}
\left\langle\theta, S^\top S\theta\right\rangle_{\mathcal{H}} &= \left\langle S\theta, S\theta\right\rangle_{L^2(\mu_\mathcal{X})} = \mathbb{E}_\xi[S\theta(\xi)^2] \\
&= \mathbb{E}_\xi[\left\langle\theta,\varphi(\xi)\right\rangle^2] = \mathbb{E}_\xi[\left\langle\theta,\varphi(\xi)\otimes\varphi(\xi)\theta\right\rangle] \\
&= \left\langle\theta,\mathbb{E}[\varphi(\xi)\otimes\varphi(\xi)]\theta\right\rangle = \left\langle\theta,\Sigma\theta\right\rangle.
\end{aligned}$$

因此， S与 $\Sigma^{1/2}$等距同构（如果将 S的奇异值分解写作 $USV^\top$，那么 $\Sigma^{1/2} = USU^\top$）。关于 K中的部分，可以用相同的推导验证， $S^\top f = \mathbb{E}[f(\xi)\varphi(\xi)] \in \mathcal{H}$ 因此（Kf）（ξ）$= (S^\top f)^\top\varphi(\xi) = \mathbb{E}_{\xi'}[f(\xi')\varphi(\xi')^\top\varphi(\xi)]$的值为。 □

利用这个等距同构，可以将 $\|S\theta\| = \left\|\Sigma^{1/2}\theta\right\|$ 替换为 $\mathcal{H}$ 上的希尔伯特范数以及 $L^2(\mu_\Xi)$，这样对于在$\mathcal{H}$中作用的 C，迹 $SCS^\top = $迹 $\Sigma^{1/2}C\Sigma^{1/2}$。回到 $\mathcal{H}$中的证明，可以在适当的位置将所有 $\Sigma^{1/2}$ 替换为 S 或其伴随算子，从而得到以下结论。

引理 12. 对于 $\Theta \in \mathbb{R}^k \otimes \mathcal{H}$，以及一个正则化的 $\lambda \in \mathbb{R}$

$$\mathcal{L}(S\Theta) + \lambda \|\Theta\|_2^2 = \mathrm{Tr}((S\Theta^\top \Theta S^\top - T_\lambda)^2 - T_\lambda^2) + k$$

其中

$$T = S^{-\top}\Sigma_X S^{-1}, \qquad T_\lambda = (1-\beta)I + \beta T - \lambda K^{-1}, \qquad K = SS^\top,$$

其中 $S: \mathcal{H} \to L^2(\mu_\Xi)$; $\theta \to f_\theta$ $\mathcal{H}$ 在 $L^2(\mu_\Xi)$中的嵌入，其中 μ_Ξ 表示增强的边缘分布。因此，$\mathcal{L}$; λ 的一个最小化器 Θ_* 满足：$S\Theta_*^\top$在正特征值上与 T_λ 的特征值分解相匹配，直到第 k 个。

证明。该引理可由前面的讨论得出。$S^{-\top}\Sigma S^{-1}$ 等价于 T在 $L^2(\mu_\Xi)$闭包上的 Ψ，这是由引理 3中的刻画得到的。尽管如此，我们仍可通过改写Saunshi 等人 (2022) 的引理 B.9 来证明本情形，而且证明方式更直接。 □

A.6. 命题 4的证明

命题 4　　依据是：当两个算子可交换时，它们可以在同一个基下对角化　　is.

引理 13. 当 K 和 T 可交换时，K 和 T 可以由相同的特征函数对角化(f_i)。

证明. 当这些算子可交换时，若 f 是 T 对应于 $Tf = \lambda f$ 的一个特征函数，则 $TKf = KTf = \lambda Kf$。这意味着 T 的特征子空间，即 ker$(T - \lambda I)$ 在 K 的作用下保持不变。因此，K 可以按照直和分解为各个 ker$(T - \lambda I)$。将 K 在这些空间上的限制分别对角化后，便存在一个基，能同时对角化 K 和 T。 □

虽然我们没有在正文中讨论这一点，但不应考虑 T 的任意特征值分解，而只应考虑能同时对角化 T 和 K 的特征函数。不过请注意，根据 Courant-Fisher 原理，要找到这些特征函数，可以对 $i \in \mathbb{N}$ 递归地从 ker$(T - \lambda_i I)$ 中选取一个使 $\|\theta_i\|$ 达到最大值或最小值的 $f_i = f_{\theta_i}$ 特征函数。这些特征函数（f_i）将对角化 T_λ，而最优表示会选取使 $f_i^\top T_\lambda f_i$ 最大的那些特征函数，只要该数量为正值。

如果 f_i对角化 K，那么 $f_i \in \mathrm{im}\ K^{1/2} = \Psi = \mathrm{im}\ S$，因此存在一个 $\theta_i \in \mathcal{H}$，使得 $f_i = S\theta_i$。由此，按照 $L^2(\mu_\Xi)$ 几何，$f_i^\top K^{-1} f_i = (S\theta_i)^\top (S^\top S)^{-1} S\theta_i = \|\theta_i\|_2^2$。我们据此推导出

$$f_i T_\lambda f_i = (1-\beta) f_i^\top f_i + \beta f_i^\top T f_i - \lambda f_i^\top K^{-1} f_i = 1 - \beta + \beta\lambda_i - \lambda \|\theta_i\|_2^2 .$$

换言之，T_λ 的最大特征值可通过最大化 $\beta\lambda_i - \lambda\|\theta_i\|^2$ 得到。

注 14。 最近，*HaoChen & Ma (2022)* 从归纳偏置的第二种视角出发，考察了“屏障”情形：在这种情形下，人们只能匹配属于函数空间 Ψ 的特征函数。在核区间中，这种观点具有误导性，因为例如，考虑高斯核 $\varphi(x)^\top \varphi(x') = -\exp(\|x - x'\|^2)$，$\Psi$ 由解析函数构成（*Sun &Zhou，2008*），因此除非处处取值为 *1*，否则它无法参数化任何示性函数。因此，他们的方法无法解释高斯核为何能在簇假设下快速学习。

A.7. 关于 VCReg 的备注

当 $\mathcal{L} = 0$ 时，寻找 ψ 等价于寻找 k 个函数（f_{θ_i}）$_i$，使它们在 L^2（μ_Ξ）中正交，并在乘以（$1 - \lambda\|\theta_i\|^2$）$_+$之前最大化 $1 - \lambda\|\theta\|^2 = 1 - \lambda f_\theta^\top K^{-1} f_\theta$。根据 Courant-Fisher 极小极大原理，函数（f_{θ_i}）$_i$ 由 k 个最大的特征函数给出，其来自 K。

A.8. 示例 3 的证明3

如果 $\mu_\Xi = \delta\rho_\mathcal{X} + (1-\delta)\mu_\perp$，那么对于任意可测函数 f

$$\|f\|^2_{L^2(\mu_\Xi)} = \delta \int_\mathcal{X} f(x)^2 \rho_\mathcal{X}(\mathrm{d}x) + (1-\delta)\int_\mathcal{X} f(x)^2 \mu_\perp(\mathrm{d}x) \geq \delta \|f(X)\|^2_{L^2(\rho_\mathcal{X})} .$$

A.9. 示例的证明4

这可由 $\mathcal{H}$ 分别嵌入到 $L^2(\rho_{\mathcal{X}})$ 和 $L^2(\mu_\Xi)$ 中的 S_ρ 和 S_μ 得出。我们先前已经看到 $S_\mu^\top S_\mu = \Sigma_{\mu_\Xi}$ 和 $S_X^\top S_X = \Sigma_{\rho_{\mathcal{X}}}$。令 $f \in \Psi$，存在 $\theta \in \mathcal{H}$ 使得 $f = f_\theta$，因此，利用 S 和 $\Sigma^{1/2}$ 之间的等距同构，

$$\begin{aligned}\|f_\theta\|^2_{L^2(\rho_{\mathcal{X}})} = \|S_X\theta\|^2_{L^2(\rho_{\mathcal{X}})} &= \left\|\Sigma^{1/2}_{\rho_{\mathcal{X}}}\theta\right\|^2_{\mathcal{H}} = \left\|\Sigma^{1/2}_{\rho_{\mathcal{X}}}\Sigma^{-1/2}_{\mu_\Xi}\Sigma^{1/2}_{\mu_\Xi}\theta\right\|^2_{L^2(\rho_{\mathcal{X}})} \\ &\le \left\|\Sigma^{-1/2}_{\mu_\Xi}\Sigma_{\rho_{\mathcal{X}}}\Sigma^{-1/2}_{\mu_\Xi}\right\|_{\mathrm{op}} \left\|\Sigma^{-1/2}_{\mu_\Xi}\theta\right\|^2_{\mathcal{H}} = \left\|\Sigma^{-1/2}_{\mu_\Xi}\Sigma_{\rho_{\mathcal{X}}}\Sigma^{-1/2}_{\mu_\Xi}\right\|_{\mathrm{op}} \|f_\theta\|^2_{L^2(\mu_\Xi)}.\end{aligned}$$

最后，我们利用等价关系得出 $A \preceq cB$ 蕴含 $B^{-1/2}AB^{-1/2} \preceq c \cdot I$ 这一事实。

A.10. 示例的证明 5

这可由不同对象的定义得出，

$$\Pi^{(\rho_{\mathcal{X}})}_{f_i}(f) = wf_i, \qquad \text{with} \qquad w = \arg\min_{w\in\mathbb{R}} \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[\|f(X) - wf_i(X)\|^2].$$

我们将最后一个目标函数展开为

$$\begin{aligned}\mathbb{E}_{X\sim\rho_{\mathcal{X}}}[\|f(X) - wf_i(X)\|^2] = \mathbb{E}_{X\sim\rho_{\mathcal{X}}}[\|g(\psi(X)) - wf_i(X)\|^2] &= \mathbb{E}_{Z\sim\psi_\#\rho_{\mathcal{X}}}[\left\|g(Z) - w^\top Z\right\|^2] \\ &= \mathbb{E}_{Z\sim\mu_\Xi}[\left\|f(X) - w^\top f_i(X)\right\|^2].\end{aligned}$$

因此， w 与投影相等。

A.11. 示例6的证明

若 μ_Ξ 有 k 个连通分支，则这些分支的指示函数在 $L^2(\mu_\Xi)$中彼此正交，同时使不变项 $\mathbb{E}_x\,\mathbb{E}\left[\|\varphi(\xi) - \varphi(\xi')\|^2 \,\middle|\, X\right]$ 最小化。因此， f^* 属于$(f_i)_{i\le k}$的空间。

B. 下游收敛的控制

本节旨在证明定理1。在下文中， k_λ 表示T_λ 作为作用于 $L^2(\mu_\Xi)$上的算子时正特征值的数量（计入重数）。我们固定 $k \le k_\lambda$，并用 $\mathcal{F}$ 表示$(f_i)_{i\in[k]}$的张成空间。在核区间内，空间 $\mathcal{F}$也可写为 $\mathcal{F} = \left\{w^\top\Theta_*\varphi \,\middle|\, w \in \mathbb{R}^k\right\}$，其中 Θ_* 是引理12中定义的最小化器。 ˆ ˆ我们将由 $\mathcal{F}$ 类似地根据 Θ 的估计 Θ_*所定义的空间记为。

下游任务上的误差可分解为三个量：与ˆ ˆ的容量相关的下游任务误差 $\mathcal{F}$ (12)；与 $\mathcal{F}$ 和 $\mathcal{F}$ (13之间的近似误差相关的上游任务误差；以及由于下游任务可能无法在 $\mathcal{F}$ (14)中得到有效解决而产生的误差。

引理 15 （分解直觉）。 设 $\mathcal{F}$和 ˆ$\mathcal{F}$是 $L^2(\rho_{\mathcal{X}})$中的两个闭凸集，且 $\Pi_{\mathcal{F}}$表示根据 $L^2(\rho_{\mathcal{X}})$ 几何结构在空间 $\mathcal{F}$上的正交投影。对于任意函数 $f\colon \mathcal{X} \to \mathcal{Y}$属于ˆ$\mathcal{F}$，超额风险(1) 可分解为

$$\begin{aligned}\mathcal{R}(f) - \mathcal{R}(f^*) \le \left\|f - \Pi_{\hat{\mathcal{F}}}f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} &\qquad (12)\\ + 2\left\|(I - \Pi_{\hat{\mathcal{F}}})\Pi_{\mathcal{F}}f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} &\qquad (13)\\ + \|(I - \Pi_{\mathcal{F}})f^*\|^2_{L^2(\rho_{\mathcal{X}})}, &\qquad (14)\end{aligned}$$

证明。 该引理的证明可由均方误差的经典刻画和三角不等式得出。引入以下技术假设。

假设 6. 假设 $(X, Y) \to Y$属于 $L^2(\rho)$。

当 $\ell(y,y') = \|y-y'\|^2$ 时，利用 $(X,Y) \to Y - \mathbb{E}[Y \mid X]$ 与任何不依赖于 L^2中的 Y的可测函数正交这一事实，

$$\mathcal{R}(f) = \mathbb{E}[\|f(X)-Y\|^2] = \mathbb{E}[\|f(X)-\mathbb{E}[Y \mid X]\|^2] + \mathbb{E}[\|\mathbb{E}[Y \mid X]-Y\|^2].$$

因此， $f^*(x) = \mathbb{E}[Y \mid X = x]$ 以及

$$\mathcal{R}(f) - \mathcal{R}(f^*) = \|f - f^*\|^2_{L^2(\rho_{\mathcal{X}})}.$$

ˆ 让我们用 $\mathcal{F}$的正交投影来分解超额风险，得到

$$\mathcal{R}(f) - \mathcal{R}(f^*) = \|f - f^*\|^2_{L^2(\rho_{\mathcal{X}}} = \left\|f - \Pi_{\hat{\mathcal{F}}} f^*\right\|^2_{L^2(\rho_{\mathcal{X}}} + \left\|(I - \Pi_{\hat{\mathcal{F}}}) f^*\right\|^2_{L^2(\rho_{\mathcal{X}}}$$

第二项可推导为

$$\begin{aligned}
\left\|(I - \Pi_{\hat{\mathcal{F}}}) f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} &= \left\|(I - \Pi_{\hat{\mathcal{F}}})\Pi_{\mathcal{F}} f^* + (I - \Pi_{\hat{\mathcal{F}}})(I - \Pi_{\mathcal{F}}) f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} \\
&\leq 2\left\|(I - \Pi_{\hat{\mathcal{F}}})\Pi_{\mathcal{F}} f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} + \left\|(I - \Pi_{\hat{\mathcal{F}}})(I - \Pi_{\mathcal{F}}) f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} \\
&\leq 2\left\|(I - \Pi_{\hat{\mathcal{F}}})\Pi_{\mathcal{F}} f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} + \|(I - \Pi_{\mathcal{F}}) f^*\|^2_{L^2(\rho_{\mathcal{X}})}
\end{aligned}$$

其中，最后一个不等式源于投影会收缩距离这一事实。 □

对于线性探测（3），当下游任务使用 n 个数据点进行学习，且噪声水平为 ε 时，（12）预计表现为 $\varepsilon^2 k/n$（Mourtada 与 Rosasco，2022）。在这种线性设定下，（13）应被视为 $\mathcal{F}$ 与 ˆ$\mathcal{F}$ 之间的角度度量，从 f^* （Davis 与 Kahan，1970；Kato，1995）的视角来看。

B.1. 控制 (12)

下游任务误差与模型设定错误的线性模型的泛化误差相关。为了给出其上界，我们将通过 Smale 与 Zhou(2007) 和 Caponnetto 与 De Vito(2007) 关于积分算子集中性的分析来研究收敛速率。这需要对之前的分解进行些许重新处理。

引理 16 (预热)。 令ˆ$\mathcal{F}$ 为$(\psi_i)^{i\in[k]}$的张成空间，其中 S_ψ： $\mathbb{R}^k \to L^2$ 定义为 $S_\psi w = w^\top \psi$， n

$$\Pi_{\hat{\mathcal{F}}} f^* = S_\psi \, \mathbb{E}[\psi(X)\psi(X)^\top]^{-1} \, \mathbb{E}[Y\psi(X)]. \tag{15}$$

基于数据(X_i, Y_i)，可以定义经验风险最小化器 $f_n = S_\psi w_n$，其中 w_n 是以下问题的最小化器

$$w_n \in \operatorname*{arg\,min}_{w\in\mathbb{R}^k} \sum_{i=1}^n \left\|w^\top \psi(X_i) - Y_i\right\|^2 = [\frac{1}{n}\sum_{i=1}^n \varphi(X_i)\varphi(X_i)^\top]^{-1} \frac{1}{n}\sum_{j=1}^n Y_i \varphi(X_i). \tag{16}$$

证明。这两个 f 通过指出，若 $\Pi_{\hat{\mathcal{F}}} f^*$定义为 $S_\psi w$使 w最小化，则可立即证明该公式

$$\mathbb{E}[\left\|w^\top \varphi(X) - Y\right\|^2] = w^\top \, \mathbb{E}[\varphi(X)\varphi(X)^\top] w - 2w^\top \, \mathbb{E}[Y\varphi(X)] + \mathbb{E}[\|Y\|^2].$$

最小化这个二次型可得第一个结果。第二个结果的证明方法相同，只需将 (X,Y) 上的分布替换为经验分布 $n^{-1}\sum_{i\in[n]} \delta_{(X_i,Y_i)}$。 □

作为其推论 预备引理，先引入一些记号。对于 ψ： $\mathcal{X} \to \mathbb{R}^k$ 以及一些数据 (X_i)，定义 e

$$S_\psi : \mathbb{R}^k \to L^2(\rho_{\mathcal{X}}); w \to w^\top \psi, \qquad \hat{S}_\psi : \mathbb{R}^k \to \ell^2(n); w \to (w^\top \psi(X_i))_{i\in[n]}, \tag{17}$$

其中 $\ell^2(n)$ 配备了归一化（即类似概率的）标量积 $\langle a, b\rangle = n^{-1}\sum_{i\in[n]} a_i b_i$。与ˆ引理 11类似，可以证明， S_ψ 和 S_ψ的伴随算子以及协方差算子分别为

$$S_\psi : L^2(\rho_{\mathcal{X}}) \to \mathbb{R}^k; f \to \mathbb{E}_{\rho_{\mathcal{X}}}[f(X)\psi(X)], \qquad \hat{S}_\psi : \ell^2(n) \to \mathbb{R}^k; (Y_i)_{i\in n} \to \frac{1}{n}\sum_{i\in[n]} Y_i \psi(X_i).$$

$$\Sigma_\psi = S_\psi S_\psi^\top = \mathbb{E}_{\rho_{\mathcal{X}}}[\psi(X)\psi(X)^\top], \qquad \hat{\Sigma}_\psi = \hat{S}_\psi \hat{S}_\psi^\top = \frac{1}{n}[\psi(X)\psi(X)^\top], \tag{18}$$

在本小节中，我们只考虑与 ψ 相关联的 S 和 Σ，并为方便起见省略下标。为简化记号，当 $f \in L^2(\rho_{\mathcal{X}})$ 时，我们将 $\hat{} S^\top f$ 记作 $\hat{} S^\top(f(X_i))_{i\in[n]}$。

假设 7 (同方差噪声)。 存在 $\varepsilon > 0$，使得对于 $\rho_{\mathcal{X}}$-几乎所有的 x，$(Y \mid X = x)$ 的方差有界于 ε^2。

引理17 (偏差-方差分解)。 基于数据 (X_i, Y_i)，可以定义带有正则化参数 $\gamma > 0$ 的正则化经验风险最小化器$f_n = S_\psi w_n$，其形式为

$$w_n \in \operatorname*{arg\,min}_{w\in\mathbb{R}^k} \sum_{i=1}^{n} \left\|w^\top \psi(X_i) - Y_i\right\|^2 + \gamma \|w\|^2 . \tag{19}$$

在此过程中，根据假设 7，平均超额风险可以分解为，其中 $M = \sup \|\psi(X)\|$,

$$\begin{aligned}\mathbb{E}_{(X_i,Y_i)}[\left\|f_n - \Pi_{\hat{\mathcal{F}}} f^*\right\|^2_{L^2(\rho_{\mathcal{X}})}] \le \frac{\varepsilon^2}{n}\left(1 + \frac{M^2}{\gamma n}\right) \operatorname{Tr}\left((\Sigma+\gamma)^{-1}\Sigma\right) \\ + 2\gamma\left(1 + \frac{M^2}{\gamma n}\right)^2 \left\langle \Pi_{\hat{\mathcal{F}}} f^*, \Sigma(\Sigma+\gamma)^{-1}\Pi_{\hat{\mathcal{F}}} f^*\right\rangle_{L^2(\rho_{\mathcal{X}})} \\ + 2\,\mathbb{E}_{(X_i)}\left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top(I - \Pi_{\hat{\mathcal{F}}}) f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} .\end{aligned} \tag{20}$$

证明。 重新采用预备引理，可以证明

$$w_n = (\hat{\Sigma}+\gamma)^{-1}\hat{S}(Y_i)_{i\in[n]}.$$

因此，u 利用常见的偏差-方差分解，以及 $f^* = \mathbb{E}_\rho[Y \mid X = \cdot]$ 这一事实，我们推导出

$$\begin{aligned}&\mathbb{E}_{(Y_i \mid X=X_i)}[\left\|f_n - \Pi_{\hat{\mathcal{F}}} f^*\right\|^2] = \mathbb{E}_{(Y_i \mid X=X_i)}[\left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top(Y_i)_{i\in[n]} - \Pi_{\hat{\mathcal{F}}} f^*\right\|^2] \\ &= \mathbb{E}_{(Y_i \mid X=X_i)}\left[\left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top(Y_i - \mathbb{E}[Y \mid X = X_i])_{i\in[n]}\right\|^2\right] + \left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top f^* - \Pi_{\hat{\mathcal{F}}} f^*\right\|^2 .\end{aligned}$$

第一项可利用Mourtada 与 Rosasco (2022) 的方法计算得出

$$\mathbb{E}_{(X_i,Y_i)}\left[\left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top(Y_i - \mathbb{E}[Y \mid X = X_i])_{i\in[n]}\right\|^2\right] \le \frac{\varepsilon^2}{n}\left(1 + \frac{R^2}{\gamma n}\right)\operatorname{Tr}\left((\Sigma+\gamma)^{-1}\Sigma\right)$$

假设 $(Y \mid X)$ 的方差有界于 ε^2。

我们利用以下方法计算第二项

$$\left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top f^* - \Pi_{\hat{\mathcal{F}}} f^*\right\| \le \left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top(f^* - \Pi_{\hat{\mathcal{F}}} f^*)\right\| + \left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top\Pi_{\hat{\mathcal{F}}} f^* - \Pi_{\hat{\mathcal{F}}} f^*\right\| .$$

O同样，最后一部分可利用Mourtada 与 Rosasco (2022) 的方法计算得出 et

$$\mathbb{E}_{(X_i,Y_i)}[\left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top\Pi_{\hat{\mathcal{F}}} f^* - \Pi_{\hat{\mathcal{F}}} f^*\right\|^2] \le \gamma\left(1 + \frac{R^2}{\gamma n}\right)^2 \left\langle \Pi_{\hat{\mathcal{F}}} f^*, \Sigma(\Sigma+\gamma)^{-1}\Pi_{\hat{\mathcal{F}}} f^*\right\rangle_{L^2(\rho_{\mathcal{X}})} .$$

由此得到该引理的分解。 □

我们来推导（20）中的最后一项。

引理 18. 对于 $t = \left\|(\Sigma+\gamma)^{-1/2}(\Sigma - \hat{}\Sigma(\Sigma+\gamma)\right\|_{\mathrm{op}}$ 以及满足 $\|\psi(X)\| \le M$ 几乎处处成立的 M，

$$\left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top(I - \Pi_{\hat{\mathcal{F}}}) f^*\right\|_{L^2(\rho_{\mathcal{X}})} \le \min\left\{\frac{1}{1-t}, 1 + t\cdot\frac{M^2+\gamma}{\gamma}\right\}\left\|\Sigma_\gamma^{-1/2}\hat{S}^\top(I - \Pi_{\hat{\mathcal{F}}}) f^*\right\| . \tag{21}$$

证明。 为简便起见，令 $f = (I - \Pi_{\hat{\mathcal{F}}})f^*$ 和 $A_\gamma = A + \gamma I$。注意到 f 与 S 的像空间正交，因此 $S^\top f = 0$。我们将最后一个数量分解为

$$\begin{aligned}
\hat{\Sigma}_\gamma^{-1}\hat{S}^\top f &= (\hat{\Sigma}_\gamma^{-1} - \Sigma_\gamma^{-1})\hat{S}^\top f + (\Sigma_\gamma)^{-1}\hat{S}^\top f \\
&= \hat{\Sigma}_\gamma^{-1}(\Sigma_\gamma - \hat{\Sigma}_\gamma)\Sigma_\gamma^{-1}S^\top f + \Sigma_\gamma^{-1}\hat{S}^\top f \\
&= \hat{\Sigma}_\gamma^{-1}(\Sigma - \hat{\Sigma})\Sigma_\gamma^{-1}\hat{S}^\top f + \Sigma_\gamma^{-1}\hat{S}^\top f \\
&= \Sigma_\gamma^{-1/2}\left(\Sigma_\gamma^{1/2}\hat{\Sigma}_\gamma^{-1}\Sigma_\gamma^{1/2}\Sigma_\gamma^{-1/2}(\Sigma - \hat{\Sigma})\Sigma_\gamma^{-1/2} + I\right)\Sigma_\gamma^{-1/2}\hat{S}^\top f
\end{aligned}$$

利用 S 与 $\Sigma^{1/2}$ 等距，且后者小于 $\Sigma_\gamma^{1/2}$（按洛文纳序）的事实，我们得到

$$\left\|S(\hat{\Sigma} + \gamma)^{-1}\hat{S}^\top(I - \Pi_{\hat{\mathcal{F}}})f^*\right\|_{L^2(\rho_{\mathcal{X}})} \le \left(1 + \left\|\Sigma_\gamma^{1/2}\hat{\Sigma}_\gamma^{-1}\Sigma_\gamma^{1/2}\right\|_{\mathrm{op}}\left\|\Sigma_\gamma^{-1/2}(\Sigma - \hat{\Sigma})\Sigma_\gamma^{-1/2}\right\|_{\mathrm{op}}\right)\left\|\Sigma_\gamma^{-1/2}\hat{S}^\top f\right\|$$

我们知道

$$\left\|\Sigma_\gamma^{1/2}\hat{\Sigma}_\gamma^{-1}\Sigma_\gamma^{1/2}\right\|_{\mathrm{op}} \le \gamma^{-1}(\|\Sigma\|_{\mathrm{op}} + \gamma) \le \gamma^{-1}(\sup_{x \in \operatorname{supp}\rho_{\mathcal{X}}}\|\psi(x)\|^2 + \gamma)$$

ˆ 我们还知道，对于自伴的 A 和 A 以及任意 $t > 0$，以下蕴含关系链成立

$$\begin{aligned}
\left\|A^{-1/2}(A - \hat{A})A^{-1/2}\right\|_{\mathrm{op}} \le t \quad &\Leftrightarrow \quad -tI \preceq A^{-1/2}(\hat{A} - A)A^{-1/2} \preceq tI \\
&\Leftrightarrow \quad -tA \preceq \hat{A} - A \preceq tA \\
&\Leftrightarrow \quad (1-t)A \preceq \hat{A} \preceq (1+t)A \\
&\Leftrightarrow \quad (1+t)^{-1}A^{-1} \preceq \hat{A}^{-1} \preceq (1-t)^{-1}A^{-1} \\
&\Leftrightarrow \quad (1+t)^{-1} \preceq A^{1/2}\hat{A}^{-1}A^{1/2} \preceq (1-t)^{-1}.
\end{aligned}$$

结合不同的结果即可得到该引理。 □

概率论证将表明， t以及 $\left\| (\Sigma + \gamma^{-1/2}\hat{}S^\top(I - \Pi_{\hat{}\mathcal{F}})f^*\right\|$，在 $n^{-1/2}$中趋于零。我们将使用伯恩斯坦集中不等式。

引理19 （伯恩斯坦集中不等式）。 设 $\mathcal{A}$ 为一个希尔伯特空间，且 $(Z_i)_{i\in[n]}$ 为 $\mathcal{A}$ 上的一列独立随机向量，满足 $\mathbb{E}[Z_i] = 0$；并且存在两个正值常数 M 和 σ，使得对于所有 $m > 2$

$$\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[\|Z_i\|^m] \le m!\sigma^2 M^{m-2}/2$$

对于任意 $t > 0$,

$$\mathbb{P}(\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\| \ge t) \le 2\exp\left(\frac{-nt^2}{2\sigma^2 + 2tM}\right).$$

特别地，当 (Z_i) 被 $3M$有界且 $\sigma^2 = n^{-1}\sum_{i\in n}\mathbb{E}[Z_i^2]$时，该条件成立。另一方面，当 Z_i 为 $\mathbb{R}^{k\times k}$ 中的对称矩阵且 $\|\cdot\|$ 为算子范数时，右侧的上界仍然成立，只需将$2\exp(\cdots)$ 替换为 $k\exp(\cdots)$，其中 $\sigma^2 = \left\|n^{-1}\sum_{i\in[n]}\mathbb{E}[Z_i^2]\right\|$。

证明。 第一部分请参见 Pinelis 与 Sakhanenko （1986），矩阵版本请参见 Tropp （2015）。 □

引理20。 对于任意 $t > 0$，偏差分解最后一项中的向量部分（20）可以控制为

$$\mathbb{P}\left(\left\|\Sigma_\gamma^{-1/2}\hat{S}^\top(I - \Pi_{\hat{\mathcal{F}}})f^*\right\| \ge t\right) \le 2\exp\left(\frac{-nt^2}{a(b + 2M\gamma^{-1/2}t/3)}\right) \tag{22}$$

其中 $b = 2$迹（$\Sigma + \gamma$）$^{-1}\Sigma$， $M =$上确界 $\|\psi$（X）$\|$以及 $a = \|f^*\|_{L^\infty} + M\|f^*\|_{L^2}$。此外，该向量部分有界于 $\gamma^{-1}a^2M^2$。式（20）中最后一项的矩阵部分可以控制为

$$\mathbb{P}\left(\left\|\Sigma_\gamma^{-1/2}(\hat{\Sigma}-\Sigma)\Sigma_\gamma^{-1/2}\right\|_{\text{op}} \geq t\right) \leq k\exp\left(\frac{-nt^2}{2M^2\gamma^{-1}(1+t/3)}\right) \tag{23}$$

此外，这个矩阵部分有界于 $\gamma^{-2}M^4$。

证明。 我们引入

$$Z_i = (I - \Pi_{\hat{\mathcal{F}}})f^*(X_i)(\Sigma+\gamma)^{-1/2}\psi(X_i) \in \mathbb{R}^k. \tag{24}$$

可以验证

$$\frac{1}{n}\sum_{i\in[n]} Z_i = (\Sigma+\gamma)^{-1/2}\frac{1}{n}\sum_{i\in[n]}(I-\Pi_{\hat{\mathcal{F}}})f^*(X_i)\psi(X_i) = (\Sigma+\gamma)^{-1/2}\hat{S}^\top(I-\Pi_{\hat{\mathcal{F}}})f^*,$$

ˆ 同样，由于 $\text{im}\, S = \mathcal{F}$

$$\mathbb{E}[Z_i] = S^\top(I-\Pi_{\hat{\mathcal{F}}})f^* = 0.$$

此外，

$$\|Z_i\| = \left\|(\Sigma+\gamma)^{-1/2}\psi(X_i)\right\|\left|(I-\Pi_{\hat{\mathcal{F}}})f^*(X_i)\right| \leq \gamma^{-1/2}M(\|f^*\|_{L^\infty} + M\|f^*\|_{L^2}).$$

其中 $R = \sup_X \|\psi(X)\|$，并且我们用到了这一事实：

$$\begin{aligned}\left|(I-\Pi_{\hat{\mathcal{F}}})f^*(X_i)\right| &\leq |f^*(X_i)| + \left|\Pi_{\hat{\mathcal{F}}}f^*(X_i)\right| = |f^*(X_i)| + \left|\left\langle SS^{-1}\Pi_{\hat{\mathcal{F}}}f^*, \varphi(X)\right\rangle\right| \\ &\leq |f^*(X_i)| + \left\|SS^{-1}\right\|_{\text{op}}\left\|\Pi_{\hat{\mathcal{F}}}f^*\right\|_{L^2}\|\varphi(X_i)\| \leq \|f^*\|_{L^\infty} + M\|f^*\|_{L^2}.\end{aligned}$$

最后，我们得到

$$\begin{aligned}\mathbb{E}[\|Z_i\|^2] &= \mathbb{E}\left[\left\|(\Sigma+\gamma)^{-1/2}\psi(X_i)\right\|^2\left|(I-\Pi_{\hat{\mathcal{F}}})f^*(X_i)\right|^2\right] \\ &\leq \mathbb{E}\left[\left\|(\Sigma+\gamma)^{-1/2}\psi(X_i)\right\|^2\right](\|f^*\|_{L^\infty} + M\|f^*\|_{L^2}) \\ &= \text{Tr}\left((\Sigma+\gamma)^{-1}\Sigma\right)(\|f^*\|_{L^\infty} + M\|f^*\|_{L^2}).\end{aligned}$$

利用伯恩斯坦不等式可以控制向量项。

对于矩阵项，我们引入

$$Z_i = U_iU_i^\top - \mathbb{E}[U_iU_i^\top], \qquad U_i = (\Sigma+\gamma)^{-1/2}\varphi(X_i).$$

我们有（$\Sigma+\gamma$）$^{-1/2}$（ˆ$\Sigma-\Sigma$）（$\Sigma+\gamma$）$^{-1/2} = \frac{1}{n}\sum_{i\in[n]} Z_i$，并且

$$\sup\|Z\| \leq \sup\|U\|^2 \leq \gamma^{-1}M^2.$$

最后，利用 $\|U_i\|^2 \preceq U_i \sup\|U_i\|$这一事实，并根据均值的变分定义（其中下确界是相对于洛文纳序取的）

$$\mathbb{E}[Z_i^2] = \inf_a \mathbb{E}[(Z_i-a)^2] \preceq \mathbb{E}[(U_iU_i^\top)^2] \preceq \sup\|U\|^2\,\mathbb{E}[U_i^\top U_i] = \sup\|U\|^2(\Sigma+\gamma)^{-1}\Sigma \preceq \sup\|U\|^2 I$$

应用矩阵版本的伯恩斯坦不等式即可得到该引理。 □

现在，我们将上一个引理中的偏差不等式转化为平均值的上界。

引理 21. 沿用上一个引理中的记号。

$$\begin{aligned}\mathbb{E}_{(X_i)}\left\|S(\hat{\Sigma}+\gamma)^{-1}\hat{S}^\top(I-\Pi_{\hat{\mathcal{F}}})f^*\right\|^2_{L^2(\rho_{\mathcal{X}})} &\leq k\exp\left(\frac{-3n\gamma}{(3+\sqrt{2})M^2}\right)(\gamma^{-4}M^6a^2(M^2+2\gamma))^2 \\ &\quad + \frac{16ab}{n} + \frac{512a^2M^2}{9\gamma n^2}.\end{aligned}$$

证明。 本质上，我们有两个随机变量， $X = \left\|\Sigma_\gamma^{-1/2}(\hat{\Sigma}-\Sigma)\Sigma_\gamma^{-1/2}\right\|_{\mathrm{op}}^2$，以及 $Y = \left\|\Sigma_\gamma^{-1/2}\hat{S}^\top(I-\Pi_{\hat{\mathcal{F}}})f^*\right\|^2$这个向量。我们利用以下事实进行计算：对于 X 正值 $\mathbb{E}[X] = \int_{t>0}\mathbb{P}(X>t)\,\mathrm{d}t$，并且 $ab>t$意味着对任意 s，有 $a>1+s$ 或 $b>t/(1+s)$,

$$\begin{aligned}
&\mathbb{E}[\min\left\{\frac{1}{1-X}, 1+\frac{(M^2+\gamma)X}{\gamma}\right\}^2 Y^2] \\
&= \int_{t\in(0,\sup(1+\gamma^{-1}(M^2+\gamma)X)^2Y^2)} \mathbb{P}\left(\min\left\{\frac{1}{1-X}, 1+\frac{(M^2+\gamma)X}{\gamma}\right\}^2 Y^2 > t\right)\mathrm{d}t \\
&\le \int \inf_s \mathbb{P}\left(\min\left\{\frac{1}{1-X}, 1+\frac{(M^2+\gamma)X}{\gamma}\right\}^2 > 1+s\right) + \mathbb{P}(Y^2 > t/(1+s))\,\mathrm{d}t.
\end{aligned}$$

我们不求闭式解，而是采用一个简单得多的上界，即直接取 $s=1$，不进行任何优化。这样可以得到一个简单得多的公式

$$\mathbb{E}[\min\left\{\frac{1}{1-X}, 1+\frac{(M^2+\gamma)X}{\gamma}\right\}^2 Y^2] \le \mathbb{P}(\frac{1}{(1-X)^2}>2)\sup(1+\gamma^{-1}(M^2+\gamma)X)^2Y^2 + 2\,\mathbb{E}[Y^2].$$

对于 Y，我们可以采用与之前相同的技巧。利用 $\exp(-(a+b)^{-1}) \le \exp(-\max(2a,\ 2b)^{-1}) \le \exp(-(2a)^{-1}) + \exp(-(2b)^{-1})$，可得

$$\begin{aligned}
\mathbb{E}[Y^2] &= \int_{t>0}\mathbb{P}(Y^2>t)\,\mathrm{d}t \le \int_{t>0} 2\exp\left(-\frac{-nt}{a(b+2M\gamma^{-1/2}t^{1/2}/3)}\right)\mathrm{d}t \\
&\le 4\int_{t>0}\exp\left(-\frac{-nt}{2ab}\right) + \exp\left(-\frac{-nt^{1/2}}{4aM\gamma^{-1/2}/3)}\right)\mathrm{d}t \\
&= 8abn^{-1} + 256a^2M^2\gamma^{-1}n^{-2}/9
\end{aligned}$$

结合前一个引理，即可证明该引理。 □

现在让我们简化迄今推导出的上界中的常数。

引理 22 （简化常数）。 上一个上界中的常数可以计算如下

$$\mathrm{Tr}\left(\Sigma(\Sigma+\gamma)^{-1}\right) \le k, \qquad M \le \lambda^{-1}k\sup\|\varphi\|, \qquad \left\langle \Pi_{\hat{\mathcal{F}}}f^*, \Sigma(\Sigma+\gamma)^{-1}\Pi_{\hat{\mathcal{F}}}f^*\right\rangle_{L^2(\rho_{\mathcal{X}})} \le \|f^*\|_{L^2(\rho_{\mathcal{X}})}.$$

我们还有

$$\|f^*\|_{L^2(\rho_{\mathcal{X}})} \le \|f^*\|_{L^\infty(\rho_{\mathcal{X}})} \le \sigma, \qquad \varepsilon^2 \le \sigma^2, \quad \sigma^2 = \sup_x \mathbb{E}\left[Y^2 \,\middle|\, X=x\right]$$

因此，前面出现的常数 a 小于 $(1+M)\sigma$。

证明。 第一个上界直接应用了以下事实： $\Sigma \preceq \Sigma+\lambda$，因此 $\mathrm{Tr}((\Sigma+\gamma)^{-1}\gamma) \le \mathrm{Tr}(I) = k$。第二个上界源于以下事实： $\psi = \hat{\Theta}\varphi$，因此 $\|\psi\| \le \left\|\hat{\Theta}\right\|_{\mathrm{op}}\|\varphi\| \le \left\|\hat{\Theta}\right\|_F\|\varphi\|$。与此同时，如果 $\hat{\Theta}$ 经过了正则化

$$\lambda\left\|\hat{\Theta}\right\|_F^2 \le \hat{\mathcal{L}}(\Theta) + \lambda\left\|\hat{\Theta}\right\|^2 \le \hat{\mathcal{L}}(0) = k.$$

对于 f^* 中的部分，我们有

$$\left\|\Sigma^{1/2}(\Sigma+\gamma)^{-1/2}\Pi_{\hat{\mathcal{F}}}f^*\right\| \le \left\|\Pi_{\hat{\mathcal{F}}}f^*\right\| \le \|f^*\|.$$

最后，最后一个等式源于以下事实： $f^*(x)$ 是 $X=x$ 条件下 Y 的均值 ，

$$\|f^*(X)\| = \|\mathbb{E}[Y\,|\,X]\| \le \mathbb{E}\left[Y^2\,\middle|\,X\right]^{1/2} \le \sigma.$$

引理到此结束 □

引理 23. 在假设 7成立时，若 $\gamma = M^2\log(n)^{1+\delta}n^{-1}$，且 $\delta > 0$，则存在一个 $N > 0$，使得对于任意 $n > N$，正则化经验风险 (19)最小化器的超额风险为

$$\mathbb{E}_{(X_i,Y_i)}[\mathcal{R}(f_n) - \mathcal{R}(f^*)] \leq \frac{2k_e\varepsilon^2}{n} + \frac{8M^2\log(n)^{1+\delta}}{n}\|f^*\|_{L^2(\rho_{\mathcal{X}})} + \frac{64ka}{n} + 2\left\|I - \Pi_{\hat{\mathcal{F}}}\Pi_{\mathcal{F}}f^*\right\|^2 + \|I - \Pi_{\mathcal{F}}f^*\|^2 \quad (25)$$

其中 $k_e = \text{迹}\left(\Sigma(\Sigma+\gamma I)^{-1}\right) \leq k$是有效维数，$a = \left\|I - \Pi_{\hat{\mathcal{F}}}f^*\right\|_{L^\infty} \leq \|f^*\|_{L^\infty} + M\|f\|_{L^2}$，且$M =$上确界 $\|\psi\| \leq k\lambda^{-1}$上确界 $\|\varphi\|$。

证明。 当 $\gamma = c\log(n)^{1+\delta}n^{-1}$ 时，超额风险可写为

$$\begin{aligned}\mathbb{E}_{(X_i,Y_i)}[\mathcal{R}(f_n) - \mathcal{R}(f^*)] \leq{}& \frac{k_e\varepsilon^2}{n}\left(1 + \frac{M^2}{c\log(n)}\right) + \frac{2c\log(n)^{1+\delta}}{n}\left(1 + \frac{M^2}{c\log(n)}\right)^2\|f^*\|_{L^2(\rho_{\mathcal{X}})} + \frac{64ka}{n} \\ &+ \frac{114a^2M^2}{9cn\log(n)} + O(\exp(-\log(n)^{1+\delta/2})) + 2\left\|I - \Pi_{\hat{\mathcal{F}}}\Pi_{\mathcal{F}}f^*\right\|^2 + \|I - \Pi_{\mathcal{F}}f^*\|^2\end{aligned}$$

取 $c = M^2$ 即可得到该引理。 □

B.2. 控制 (13)

要理想地控制（13），可以利用总体风险和经验风险的闭式解，并像 Cabannes 等人 (2021b)；Pillaud-Vivien 与 Bach(2023) 一样，对积分算子应用集中不等式。然而，这些证明通过估计 $(\hat{\Sigma}_X + \lambda)^{-1/2}\Sigma(\hat{\Sigma}_X + \lambda)^{-1/2}$ 的最小特征函数，而非 $\Sigma^{-1/2}(\Sigma_X - \lambda)\Sigma^{-1/2}$ 的最大特征函数。在本证明中，我们将采用基于经验过程集中的推导，并结合以下“迁移界”。

引理 24 （迁移界）。 对于 $\hat{\Theta} \in \mathbb{R}^k \otimes \mathcal{H}$，以及 $\hat{\mathcal{F}} = \left\{x \to w^\top\hat{\Theta}\varphi(x) \;\middle|\; w \in \mathbb{R}^k\right\}$，

$$\sum_{i\in[k]}\lambda_i^2\left\|(\Pi_{\mathcal{F}}^{(\mu_\Xi)} - \Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})f_i\right\|_{L^2(\mu_\Xi)}^2 - \sum_{k<i\leq k_\lambda}\lambda_i^2\left\|\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)}f_i\right\|_{L^2(\mu_\Xi)}^2 \leq \mathcal{L}(\hat{\Theta};\lambda) - \mathcal{L}(\Theta_*;\lambda), \quad (26)$$

其中，$\Pi_{\mathcal{F}}^{(\tau)}$ 是 L^2 中 $\mathcal{F}$ 上的正交投影（τ）。

证明。 为简单起见，我们在证明中省略对 μ_Ξ 的依赖。引入 $C = S\hat{\Theta}\hat{\Theta}^\top S^\top$， C 是 L^2 中秩为 k 的正算子，将其写为 $C = \sum_{i\in[k]}\mu_i g_i g_i^\top$，其中 $\mu_i \geq 0$。

$$\mathcal{L}(\hat{\Theta};\lambda) - k = \operatorname{Tr}\left((C - T_\lambda)^2 - T_\lambda^2\right) = \operatorname{Tr}\left(C^2 - 2C^{1/2}T_\lambda C^{1/2}\right).$$

将 $T_\lambda = T$ 分解为 $T_+ - T_-$，其中 T_+ 和 T_- 为正值。由于 $T_\lambda \preceq T_+$、$-C^{1/2}T_\lambda C^{1/2} \succeq C^{1/2}T_+C^{1/2}$，因此

$$\mathcal{L}(\hat{\Theta};\lambda) - k \geq \operatorname{Tr}\left(C^2 - 2C^{1/2}T_+C^{1/2}\right) = \sum_{i\leq k}\mu_i^2 - 2\mu_i\left\|T_+^{1/2}g_i\right\|^2.$$

对 μ_i最小化这个数量，可得

$$\mathcal{L}(\hat{\Theta};\lambda) - \mathcal{L}(\Theta;\lambda) \geq \sum_{i\leq k}\lambda_i^2 - \sum_{i\leq k}\left\|T_+^{1/2}g_i\right\|^4.$$

现在我们引入 (f_i) T_λ的特征函数。由于 $U = (\langle g_i, f_j\rangle^2)_{ij} \in \mathbb{R}^{k\times k_\lambda}$ 和 $\lambda = (\lambda_i) \in \mathbb{R}^{k_\lambda}$，我们有

$$\sum_{i\leq k}\left\|T_+^{1/2}g_i\right\|^4 = \sum_{i\leq k}(g_i^\top T_+ g_i)^2 = \sum_{i\leq k}\left(\sum_{j\leq k_\lambda}\lambda_j\langle g_i, f_j\rangle^2\right)^2 = \sum_{j,m\leq k_\lambda}\lambda_j\lambda_m\sum_{i\leq k}\langle g_i, f_j\rangle^2\langle g_i, f_m\rangle^2 = \lambda^\top U^\top U\lambda.$$

注意，由于（g_i）和（f_i）是正交归一族，因此 U至多是双随机的，即 $\|U\| \le 1$，以及 $U^\top U \preceq I$。如果在 U的定义中将 f_i替换为 $f_i / \left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|$，则其将变为 $\tilde{U} = \mathrm{diag}(\left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2)_{i\le k_\lambda})^{-1} U$，$\tilde{U}$仍然是右随机的。因此

$$U^\top U \preceq \mathrm{diag}(\left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2_{i\le k_\lambda})^2 \,\mathrm{diag}(\left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2_{i\le k_\lambda}).$$

由此可得

$$\sum_{i\le k} \left\|T_+^{1/2} g_i\right\|^4 \le \lambda^\top U^\top U \lambda = \sum_{i\le k_\lambda} \lambda_i^2 \left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2.$$

这使得下界可以简化为

$$\begin{aligned}
\mathcal{L}(\hat{\Theta};\lambda) - \mathcal{L}(\Theta;\lambda) &\ge \sum_{i\le k} \lambda_i^2 \left(f_i^\top f_i - f_i^\top \Pi_{\hat{\mathcal{F}}} f_i\right) - \sum_{k<i\le k_\lambda} \lambda_i^2 f_i^\top \Pi_{\hat{\mathcal{F}}} f_i \\
&= \sum_{i\le k} \lambda_i^2 \left\langle f_i, (I - \Pi_{\hat{\mathcal{F}}}) f_i \right\rangle - \sum_{k<i\le k_\lambda} \lambda_i^2 \left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2 \\
&= \sum_{i\le k} \lambda_i^2 \left\|(\Pi_{\mathcal{F}} - \Pi_{\hat{\mathcal{F}}}) f_i\right\|^2 - \sum_{k<i\le k_\lambda} \lambda_i^2 \left\|\Pi_{\hat{\mathcal{F}}} f_i\right\|^2.
\end{aligned}$$

至此，我们完成了迁移界的证明。 □

引理 24 中的左侧需要与对（13）所需的控制联系起来。为了更细致地处理分布偏移，我们引入以下假设 1 和 2 的一般变体。

假设 n 8 (低扩张)。 假设对于原函数空间中的任意函数 $f \in \Psi$ (4),

$$\|f\|_{L^2(\rho_{\mathcal{X}})} \le \zeta\left(\|f\|_{L^2(\mu_{\mathcal{X}})}\right),$$

满足 $\zeta:\ \mathbb{R} \to \mathbb{R}$ 连续、递增且 $\zeta(0) = 0$。

定义 25 (分布 ε-稳健性)。 若函数凸集 $\mathcal{F}$ 是闭的，则称其在给定函数 f 的条件下对分布偏移具有 ε-稳健性，当

$$\left\|\Pi_{\mathcal{F}}^{(\rho_{\mathcal{X}})} f - \Pi_{\mathcal{F}}^{(\mu_\Xi)} f\right\|_{L^2(\rho_{\mathcal{X}})} \le \varepsilon \|f\|_{L^2(\rho_{\mathcal{X}})},$$

其中 $\Pi_{\mathcal{F}}^{(\tau)}$ 是 $L^2(\tau)$中在 $\mathcal{F}$ 上的正交投影。

假设 9。 存在一个递增且有界的剖面函数 $\sigma:\ \mathbb{R}^2 \to \mathbb{R}$，使得对于任意 $k \in \mathbb{N}$，张成空间 $\{f_i\}_{i\in[k]}$ 对 f^*具有 $\sigma(k)$-稳健性。

引理 26 (分解)。 在假设 8 和 9下，满足 $\mathcal{F}_l$ 其张成空间 $(f_i)^{i\in[l]}$

$$\left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})}) \Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})} f^*\right\|_{L^2(\rho_{\mathcal{X}})} \le \sigma(l) + \zeta\left(\sum_{i\le l} \left|\langle f^*, f_i\rangle_{L^2(\mu_\Xi)}\right| \left\|(\Pi_{\mathcal{F}_l}^{(\mu_\Xi)} - \Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)}) f_i\right\|_{L^2(\mu_\Xi)}\right). \tag{27}$$

证明。利用以下事实：当 是投影时，是投影，并且投影会收缩距离，我们得到

$$\begin{aligned}
\left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})}) \Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})} f^*\right\|_{L^2(\rho_{\mathcal{X}})} &\le \left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})})(\Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})} - \Pi_{\mathcal{F}_l}^{(\mu_\Xi)}) f^*\right\|_{L^2(\rho_{\mathcal{X}})} + \left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})}) \Pi_{\mathcal{F}_l}^{(\mu_\Xi)} f^*\right\|_{L^2(\rho_{\mathcal{X}})} \\
&\le \left\|(\Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})} - \Pi_{\mathcal{F}_l}^{(\mu_\Xi)}) f^*\right\|_{L^2(\rho_{\mathcal{X}})} + \left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})}) \Pi_{\mathcal{F}_l}^{(\mu_\Xi)} f^*\right\|_{L^2(\rho_{\mathcal{X}})}.
\end{aligned}$$

在假设 9下，前一个等式右侧的第一项有界于 $\sigma(l)$。关于第二项，在假设 8下，对于 $f \in \Psi$ 和 $f' \in \hat{\mathcal{F}} \subset \Psi$，我们有

$$\left\|(I - \Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})}) f\right\|_{L^2(\rho_{\mathcal{X}})} \le \|f - f'\|_{L^2(\rho_{\mathcal{X}})} \le \zeta\left(\|(f - f'\|_{L^2(\mu_\Xi)}\right).$$

取右侧的最小值，并利用 ζ 递增这一事实，可得

$$\left\|(I-\Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})})f\right\|_{L^2(\rho_{\mathcal{X}})} \le \zeta\left(\left\|(I-\Pi_{\mathcal{F}_l}^{(\mu_\Xi)})f\right\|_{L^2(\mu_\Xi)}\right).$$

应用于 $\Pi^{(\mu_\Xi)}{}_{\mathcal{F}_l}\ f^*$，可得

$$\left\|(I-\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})\Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})}f^*\right\|_{L^2(\rho_{\mathcal{X}})} \le \zeta\left(\left\|(I-\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})\Pi_{\mathcal{F}_l}^{(\mu_\Xi)}f^*\right\|_{L^2(\mu_\Xi)}\right).$$

We a我们已经处理完所有与分布偏移相关的量。在假设 3下，我们有 e

$$\begin{aligned}\left\|(I-\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})\Pi_{\mathcal{F}_l}^{(\mu_\Xi)}f^*\right\|_{L^2(\mu_\Xi)} &= \left\|\sum_{i\le l}\langle f^*,f_i\rangle_{\mu_\Xi}(I-\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})f_i\right\|_{L^2(\mu_\Xi)}\\ &\le \sum_{i\le l}\left|\langle f^*,f_i\rangle_{\mu_\Xi}\right|\left\|(I-\Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})f_i\right\|_{L^2(\mu_\Xi)}.\end{aligned}$$

汇总前面的方程即可得到该引理。 □

为了更精细地控制 (13)，注意到 (26) 的左侧还满足一些额外约束，这有助于收紧我们的上界。为简单起见，下面将省略所有对 μ_Ξ 的依赖。本质上，我们希望对 λ_i^2给出下界，并对 $\|(\Pi_{\mathcal{F}}-\Pi_{\hat{\mathcal{F}}})f_i\|$给出上界。下一个引理对 $\mathcal{L}(\Theta;\lambda)$施加约束，从而限制在 (13) 上可能产生的最大误差。

引理 27。 当 $\mathcal{F}$ 的维度为 k 且 $\hat{\mathcal{F}}$ 的维度为 k' 时，我们有

$$\sum_{i\le k}\left\|(\Pi_{\mathcal{F}}-\Pi_{\hat{\mathcal{F}}})f_i\right\|^2 = k-k'+\sum_{i>k}\left\|\Pi_{\hat{\mathcal{F}}}f_i\right\|^2 \le k. \tag{28}$$

证明。 考虑两个投影 U 和 V，分别投影到 $(u_i)_{i\in[k]}$ 和 $(v_i)_{i\in[k]}$ 的张成空间上，其中 $(u_i)_{i\in\mathbb{N}}$ 和 $(v_i)_{i\in\mathbb{N}}$ 是环境空间的两个标准正交基。我们在全文中使用 Hilbert-Schmidt 范数，可得

$$\|U(I-V)\|^2 = \|U\|^2-\|UV\|^2 = k-\|UV\|^2 = k-\left\|(UV)^\top\right\|^2 = k-k'+k'-\|VU\|^2 = k-k'+\|V(I-U)\|^2.$$

基于不变性 于 Hilbert-Schmidt 范数在伴随算子下保持不变，以及投影是自伴的这一事实，我们有

$$\|U(I-V)\|^2 = \|(I-V)U\|^2 = k-k'+\|V(I-U)\|^2 = k-k'+\|(I-U)V\|^2.$$

最后，我们还知道，由于投影会收缩距离 $\|(I-V)U\|^2 \le \|U\|^2 = k$。该引理的论断在于将其明确写为

$$\begin{aligned}\left\|(I-\Pi_{\hat{\mathcal{F}}})\Pi_{\mathcal{F}}\right\|^2 &= \left\|(\Pi_{\mathcal{F}}-\Pi_{\hat{\mathcal{F}}})\Pi_{\mathcal{F}}\right\|^2 = \sum_{i\le k}\left\|(\Pi_{\mathcal{F}}-\Pi_{\hat{\mathcal{F}}})f_i\right\|^2\\ &= k-k'+\left\|\Pi_{\hat{\mathcal{F}}}(I-\Pi_{\mathcal{F}})\right\|^2 = k-k'+\sum_{i>k}\left\|\Pi_{\hat{\mathcal{F}}}f_i\right\|^2 \le k.\end{aligned}$$

这引出了引理的陈述。 □

给定对（2）的控制，求（13）的上界可归结为一个纯代数问题。为了找出 $\sum_{i\le k}|\langle f^*,f_i\rangle|\left\|(\Pi^{(\mu_\Xi)}{}_{\mathcal{F}}-\Pi_{\hat{\mathcal{F}}})f_i\right\|_{L^2(\mu_\Xi)}$可能取到的最差值，我们引入

$$x_i = \left\|(\Pi_{\mathcal{F}}-\Pi_{\hat{\mathcal{F}}})f_i\right\|,\quad c_i = |\langle f^*,f_i\rangle|. \tag{29}$$

前一　　我们的结果导出如下最大化问题，以求出（13）的最差值。

$$\begin{aligned} & \max_x \sum_{i\le k} c_i x_i && (30)\\ \text{subject to}\quad & \sum_{i\le k}\lambda_i^2 x_i^2 - \sum_{k<i\le k_\lambda}\lambda_i^2 x_i^2 \le \varepsilon && \text{(Lemma 24)}\\ & \sum_{i\le k} x_i^2 = k - k' + \sum_{k<i\le k_\lambda} x_i^2 \le k && \text{(Lemma 27)} \end{aligned}$$

B.3. 在控制后保持简单并得出结论(14)

要巧妙地解决上述代数问题，以获得关于（13）的最佳上界$(x_i)_{ik}$，需要区分许多情况。尽管区分这些不同情况并展示不同的收敛机制可能具有相关性，但本小节采用一种更简单、尽管不够紧的方式。具体而言，我们可以利用以下事实简化关于 $k' \le k$（它是基于样本的 $\hat{} T_\lambda$ 的正特征值数量与 k 之间的最小值），从而得到 $x_{k+1}^2 = \sum_{i\le k} x_k^2$ 以及 $x_{k+1+j}^2 = 0$，（30）变为

$$\begin{aligned} & \max_x \sum_{i\le k} c_i x_i && (30)\\ \text{subject to}\quad & \sum_{i\le k}(\lambda_i^2 - \lambda_{k+1}^2)x_i^2 \le \varepsilon \end{aligned}$$

一般而言，可以通过引入概率论证来细化这一表述，从而说明可以预期误差在多大程度上集中于与 T_λ^2 的最小特征值相关的特征子空间，即 $\Pi_{\hat{}\mathcal{F}}$ 与 $\Pi_{\mathcal{F}}$ 之间的误差。如果 c_i 比 λ_i 下降得更快，那么我们希望将（x_i）$_{i\le k}$ 的能量归入最小的指标。否则，我们希望将（x_i）$_{i\le k}$ 的能量归入最大的指标。

为简化起见，我们首先在不施加任何秩限制的情况下优化 $\mathcal{L}$，这样便可以考虑 $\lambda_{k_\lambda+1} = 0$，之后再通过阈值处理秩，将其限制到维度为 k 的空间。

引理 28. 在假设 *8* 和 *9*下， $\mathcal{F}_l$ 前 l 个 T_λ的特征函数张成空间，

$$\begin{aligned} &\left\|(I - \Pi_{\hat{\mathcal{F}}}^{(\rho_\mathcal{X})})f^*\right\|_{L^2(\rho_\mathcal{X})}^2 \\ &\quad\le \inf_{l\le k}\left\|(I - \Pi_{\mathcal{F}_l}^{(\rho_\mathcal{X})})f^*\right\|_{L^2(\rho_\mathcal{X})}^2 + 4\sigma(l)^2 + 4\zeta^2\left(\left\|\tilde{T_\lambda}^{-1}\Pi_{\mathcal{F}_l}^{(\mu_\Xi)}f^*\right\|_{L^2(\mu_\Xi)}\left(\mathcal{L}(\hat{\Theta};\lambda) - \mathcal{L}(\Theta;\lambda)\right)^{1/2}\right). \end{aligned} \tag{31}$$

其中 $\tilde{T_\lambda} = \sum_{i\in[k]}(\lambda_i^2 - \lambda_{k+1}^2)^{1/2} f_i f_i^\top$。此外，在搜索 $\hat{}\mathcal{F}$ 时，若不对 Θ施加秩限制，则在$\hat{}$进行阈值处理以将 $\mathcal{F}$ 约减至维度为 k的空间之前，在强假设 *1* 和 *2*以及假设 *3*

$$\left\|(I - \Pi_{\hat{\mathcal{F}}_k})f^*\right\|^2 \le |k - k_\lambda|\,\|f^*\|_{L^2(\rho_\mathcal{X})}^2 + 2c_r\left\|T_\lambda^{-1}f^*\right\|_{L^2(\mu_\Xi)}^2\left\{\mathcal{L}(\hat{\Theta};\lambda) - \mathcal{L}(\Theta_*;\lambda)\right\}. \tag{32}$$

证明。 沿用上述代数记号，这可由柯西-施瓦茨不等式的简单应用得到，对于 $(a_i) \in \mathbb{R}^k$

$$\sum_{i\le l} c_i x_i = \sum_{i\in[l]} \frac{c_i}{a_i} a_i x_i \le \left(\sum_{i\le[l]} \frac{c_i^2}{a_i^2}\right)^{1/2}\left(\sum_{i\in[l]} a_i^2 x_i^2\right)^{1/2}.$$

When 适用于式（29）中的量以及 $a_i = \lambda_i^2 - \lambda_{k+1}^2$ 和 $l \le k$，由前述引理可得

$$\begin{aligned}\left\|(I-\Pi_{\hat{\mathcal{F}}}^{(\rho_{\mathcal{X}})})\Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})} f^*\right\|_{L^2(\rho_{\mathcal{X}})} &\le \sigma(l) + \zeta\left(\sum_{i\le l}\left|\langle f^*, f_i\rangle_{L^2(\mu_\Xi)}\right|\left\|(\Pi_{\mathcal{F}_l}^{(\mu_\Xi)} - \Pi_{\hat{\mathcal{F}}}^{(\mu_\Xi)})f_i\right\|_{L^2(\mu_\Xi)}\right)\\ &\le \sigma(l) + \zeta\left(\sum_{i\le l} c_i x_i\right) \le \sigma(l) + \zeta\left(\left(\sum_{i\le l}\frac{c_i^2}{a_i^2}\right)^{1/2}\left(\sum_{i\in[l]} a_i^2 x_i^2\right)^{1/2}\right)\\ &\le \sigma(l) + \zeta\left(\left(\sum_{i\le l}\frac{c_i^2}{a_i^2}\right)^{1/2}\left(\mathcal{L}(\hat{\Theta};\lambda) - \mathcal{L}(\Theta;\lambda)\right)^{1/2}\right).\end{aligned}$$

最后我们指出 $\sum_{i\le l}\frac{c_i^2}{a_i^2} = \left\|\tilde{T}_\lambda^{-1}\Pi_{\mathcal{F}_l^{(\mu_\Xi)}} f^*\right\|_{L^2}{}^{(\mu_\Xi)}$。

ˆ 对于第二部分，令 $\mathcal{F}_k$ 为通过对 $\mathcal{L}$进行经验最小化而得到的全部 k 个首批特征函数，并令$\mathcal{F}$为与 T_λ的正特征值相关联的所有特征函数的张成空间。我们重新梳理超额风险的分解，可得

$$\begin{aligned}\left\|(I-\Pi_{\hat{\mathcal{F}}_k})f^*\right\|^2 &= \left\|\Pi_{\hat{\mathcal{F}}_{k_\lambda}}(I-\Pi_{\hat{\mathcal{F}}_k})f^*\right\|^2 + \left\|(I-\Pi_{\hat{\mathcal{F}}_{k_\lambda}})(I-\Pi_{\hat{\mathcal{F}}_k})f^*\right\|^2\\ &= \left\|(\Pi_{\hat{\mathcal{F}}_{k_\lambda}} - \Pi_{\hat{\mathcal{F}}_k})f^*\right\|^2 + \left\|(I-\Pi_{\hat{\mathcal{F}}_{k_\lambda}})f^*\right\|^2\\ &\le \left\|(\Pi_{\hat{\mathcal{F}}_{k_\lambda}} - \Pi_{\hat{\mathcal{F}}_k})f^*\right\|^2 + 2\left\|(I-\Pi_{\hat{\mathcal{F}}_{k_\lambda}})\Pi_{\mathcal{F}} f^*\right\|^2 + \|(I-\Pi_{\mathcal{F}})f^*\|^2\\ &\le |k-k_\lambda|\,\|f^*\|^2 + 2\left\|(I-\Pi_{\hat{\mathcal{F}}_{k_\lambda}})\Pi_{\mathcal{F}} f^*\right\|^2.\end{aligned}$$

最后一个上界源于假设 3，以及对两个投影的算子范数所作的宽松界定。虽然可以去掉 $k-k_\lambda$，但我们保留它，因为我们预期该数量会以这种方式变化，其常数与 $\|f^*\|^2/k_\lambda$相近，而不是 $\|f^*\|^2$。

□

现在我们可以陈述主定理。

定理4。 在假设 *3*、*7*、*8* 和*9*下，存在一个正则化项 γ，使得正则化经验风险最小化器满足：对于任意 $\delta > 0$，存在一个 $N_\delta > 0$，使得对于任意 $n > N_\delta$，正则化经验风险最小化器的超额风险 (19)可表示为

$$\begin{aligned}\mathcal{R}(f) - \mathcal{R}(f^*) \le{}& \frac{2k_e\varepsilon^2}{n} + \frac{8M^2\log(n)^{1+\delta}}{n}\|f^*\|_{L^2(\rho_{\mathcal{X}})} + \frac{64ka}{n}\\ &+ \inf_{l\le k}\left\|(\Pi_{\mathcal{F}_{k_\lambda}} - \Pi_{\mathcal{F}_l}^{(\rho_{\mathcal{X}})})f^*\right\|_{L^2(\rho_{\mathcal{X}})}^2 + 4\sigma(l)^2 + 4\zeta^2\left(\left\|\tilde{T_\lambda}^{-1}\Pi_{\mathcal{F}_l}^{(\mu_\Xi)} f^*\right\|_{L^2(\mu_\Xi)}\left(\mathcal{L}_k(\hat{\Theta};\lambda) - \mathcal{L}_k(\Theta;\lambda)\right)^{1/2}\right). \quad (33)\end{aligned}$$

其中 $\mathcal{F}_l$ 是 l个首要特征函数张成的空间，来自 T_λ， k_λ 是 T_λ中严格为正的特征函数数量， $k_e \le k$ 是 ψ在 $L^2(\rho_{\mathcal{X}})$中的有效维数，$a = \left\|I - \Pi_{\hat{}\mathcal{F}} f^*\right\|_{L^\infty} \le \|f^*\|_{L^\infty} + M\|f\|_{L^2}$， $M =$上确界 $\|\psi\| \le k\lambda^{-1}$上确界 $\|\varphi\|$，以及 $\tilde{T}_\lambda = \sum_{i\in[k]}(\lambda_i^2 - \lambda_{k+1}^2)^{1/2} f_i f_i^\top$。此外，仅在假设 *1* 和 *2*下，我们有更简单的上界

$$\begin{aligned}\mathcal{R}(f) - \mathcal{R}(f^*) \le{}& \frac{2k_e\varepsilon^2}{n} + \frac{8M^2\log(n)^{1+\delta}}{n}\|f^*\|_{L^2(\rho_{\mathcal{X}})} + \frac{64ka}{n} + \max(k-k_\lambda, 0)\,\|f^*\|_{L^2(\rho_{\mathcal{X}})}^2\\ &+ 2c_r\left\|T_\lambda^{-1}\Pi_{\mathcal{F}_\lambda} f^*\right\|_{L^2(\mu_\Xi)}^2\left\{\mathcal{L}_{k_\lambda}(\hat{\Theta};\lambda) - \mathcal{L}_{k_\lambda}(\Theta_*;\lambda)\right\} + \|(I-\Pi_{\mathcal{F}_\lambda})f^*\|_{L^2(\mu_\Xi)}\end{aligned}$$

其中，ˆΘ 在最后一个表达式中被视为属于 $\mathbb{R}^{k_\lambda}\otimes\mathcal{H}$，而 $\mathcal{F}_\lambda$ 是与 T_λ的正特征值相关联的特征子空间。

B.4. 讨论

B.4.1. 有限个正特征值

以下结果将 T_λ 的特征值与 K的特征值联系起来。该结果特别证明，当 K 是迹类时， k_λ 是有限的；这是定理1中的一项结论。

引理 29 （关联 K 与 T_λ之间的容量）。若（μ_i）是 K 的特征值，则 T_λ 中大于 $t \in \mathbb{R}$ 的特征值数量小于 $\{i \,|\, \mu_i > \lambda/$ 的基数（$1-t$）$\}$. 此外，若存在 $q > 0$ 使得迹 $\left(K^{1/q}\right) < +\infty$，则存在一个 c_q，使得若（μ_i）是 K 的特征值，则有 $\mu_i \leq c_q i^{-q}$. 因此，在这种情况下，对于任意 $t \in \mathbb{R}$， T_λ 中大于 t 的特征值数量小于（$c_q\left(1-t\right)/\lambda$）$^{1/q}$。

证明。令我们考虑特征值大于 t 的特征向量集合 (f_i)。考虑这个空间的张成空间，我们希望量化其维数。我们知道，这个张成空间中的所有酉向量都满足

$$t \leq x^\top T_\lambda x \leq x^\top T x - \lambda x^\top K^{-1} x \leq 1 - \lambda x^\top K^{-1} x$$

因此

$$x^\top K^{-1} x \leq \frac{1-t}{\lambda}$$

这意味着，该张成空间与由特征值大于 $\lambda/(1-t)$ 的特征向量 K^{-1} 张成的空间不相交。换言之，这个线性空间与余维数为 d 的一个线性空间不相交，其中 d 是引理陈述中提到的基数。记我们感兴趣的空间为 U，记除原点外与其不相交的空间为 V，并记环境空间为 E。由于 $U \cap V = \{0\}$，商空间 $(U+V)/V$ 与 U 同构，因此

$$\dim(U) = \dim\left(\frac{U+V}{V}\right) \leq \dim\left(\frac{E}{V}\right) = \operatorname{codim}(V) = d.$$

这就完成了引理第一部分的证明。
第二个结论由 $\mu_i^{1/q}$ 可求和且递减这一事实推出，因此序列 $S_n = \sum_{i \leq n} \mu_i^{1/q}$ 是一个柯西序列。由此可知，存在 $N \in \mathbb{N}$，使得对任意 $s > N/2$，都有

$$s\mu_{2s}^{1/q} \leq S_{2s} - S_s \leq 1/2.$$

因此，对于所有 $s \geq N$，我们都有 $\mu_s \leq s^{-1/q}$，从而 $\mu_s / s^{-1/q}$ 有界。记其最大值为 c_q，即可得到第一个结论。最后一个陈述是以下事实的推论： $c_q i^{-q} > \lambda/(1-t)$ 蕴含 $i < (c_q(1-t)/\lambda)^{1/q}$。 □

例 8。 考虑径向基函数核 $\varphi(x)^\top \varphi(x') = \exp(-\|x - x'\|^2)$， Ψ 是解析函数空间（*Sun & Zhou, 2008*），与 L^2 空间相比（*Kolmogorov & Tikhomirov, 1959*）已知其规模较小。因此，可以将 $q = +\infty$ 视为前一个引理中的情形。更一般地，当 φ 有界时， K 是迹类的，并且可以取 $q = 1$。

证明。 的容量与 K的容量有关，而后者又与 $K(\left\{ f \,\middle|\, \|f\|_{L^2(\mu_\Xi)} \leq 1 \right\})$的容量有关，后者本身又与$\Psi = \operatorname{im} K^{1/2}$ 的容量有关。这解释了为什么本质上可以将 q 取为任意大（Bach, 2023）。

当 φ 有界时，以下

$$\begin{aligned}\operatorname{Tr}(K) = \operatorname{Tr}\left(SS^\top\right) = \operatorname{Tr}\left(S^\top S\right) = \operatorname{Tr}\left(\mathbb{E}[\varphi(X)\varphi(X)^\top]\right) = \mathbb{E}[\operatorname{Tr}\left(\varphi(X)\varphi(X)^\top\right)] \\ = \mathbb{E}[\varphi(X)^\top \varphi(X)] = \mathbb{E}[\|\varphi(X)\|^2] < +\infty,\end{aligned}$$

证明 K 是迹类。 □

B.4.2. 消失偏差的推导

在正文中，我们假设 T_λ 是定义表征学习解的正确算子（这也解释了假设 3）。这或许会冒犯纯粹主义者；更理想的做法是定义一种有原则的解，不依赖于架构的选择（不过，与某些架构相比，这种解可能更容易用其他架构近似）。因此，值得研究定理 4 中最后一个表达式在 λ 趋于零时的行为。

我们将对这一消失情形下的归纳偏置进行更精确的研究留待未来工作：从本质上说，架构的选择 Ψ 通过 T 扰动 λK^{-1} 使其成为 T_λ；理想情况下，我们希望量化在降低正则化参数时，T_λ 向 T 收敛的速度，前提是从 f^* 的角度来看。在核区间中，可以通过扰动理论（Kato，1995）以及考虑假设Davis 与 Kahan，1970）的 Davis-Kahan定理的改进来刻画这一点，同时纳入假设3。此外，当 K 与 T 可交换时，借助命题4，可以更直接地研究这种相互作用。

C. 上游超额风险的控制

为了控制超额风险，可以使用源自优化的技术，也可以使用源自经典统计学习的技术。

C.1. 拉德马赫复杂度

首先，我们指出，当用 $\Lambda = \Theta^\top\Theta \in \mathcal{H}\otimes\mathcal{H}$ 参数化时，$\mathcal{L}$ 是一个二次函数。

引理 30. 令 $\Theta \in \mathbb{R}^k \otimes \mathcal{H}$，记 $\Lambda = \Theta^\top\Theta \in \mathcal{H}\otimes\mathcal{H}$

$$\mathcal{L}(S\Theta) = 2(\beta-1)\,\mathbb{E}_\xi[\langle \Lambda, \varphi(\xi)\varphi(\xi)^\top\rangle] - 2\beta\,\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\langle \Lambda, \varphi(\xi')\varphi(\xi)^\top\rangle \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}\left[\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top\rangle^2\right] + k. \quad (34)$$

此外，正则化项为 $\lambda\|\Theta\|^2 = \lambda\,\text{迹}\,\Lambda = \lambda\langle\Lambda, I\rangle$。

证明。 考虑 $\psi = \Theta\varphi$，我们有

$$\begin{aligned}
\mathcal{L}(\psi;\beta) &= 2(\beta-1)\,\mathbb{E}_\xi[\psi(\xi)^\top\psi(\xi)] - 2\beta\,\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\psi(\xi)^\top\psi(\xi') \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}\left[(\psi(\xi')^\top\psi(\xi))^2\right] + k.\\
&= 2(\beta-1)\,\mathbb{E}_\xi[\varphi(\xi)^\top\Lambda\varphi(\xi)] - 2\beta\,\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\varphi(\xi)^\top\Lambda\varphi(\xi') \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}\left[(\varphi(\xi')^\top\Lambda\varphi(\xi))^2\right] + k.\\
&= 2(\beta-1)\,\mathbb{E}_\xi[\mathrm{Tr}\left(\Lambda\varphi(\xi)\varphi(\xi)^\top\right)] - 2\beta\,\mathbb{E}_X\,\mathbb{E}_{\xi,\xi'}\left[\mathrm{Tr}\left(\Lambda\varphi(\xi')\varphi(\xi)^\top\right) \,\middle|\, X\right] + \mathbb{E}_{\xi,\xi'}\left[\mathrm{Tr}\left(\Lambda\varphi(\xi)\varphi(\xi')^\top\right)^2\right] + k.
\end{aligned}$$

该引理由希尔伯特-施密特几何的迹刻画、Λ 是自伴的这一事实，以及正则化项写作 $\|\Theta\|^2 = \mathrm{Tr}\,\Theta^\top\Theta$ 得出。 □

我们回顾统计学习文献中的三个有用事实。

引理31。 令 $\mathcal{R}(\zeta) = \mathbb{E}_Z[\ell(\zeta, Z)]$，$\zeta^*$ 为定义域内针对 ζ 的 $\mathcal{L}$ 最小化器，并令 ζ_n 为基于可交换数据 Z_i 的 $\mathcal{R}_{(Z_i)}(\zeta) = \frac{1}{n}\sum_{i\in[n]}\ell(\zeta, Z_i)$ 最小化器，其中 $\mathbb{E}_{(Z_i)}[\mathcal{R}_{(Z_i)}] = \mathcal{R}$。$\zeta_n$ 的平均超额风险由拉德马赫复杂度上界为

$$\mathcal{R}(\zeta_n) - \mathcal{R}(\zeta_*) \le 4\,\mathbb{E}_{(Z_i),(\sigma_i)}\left[\sup_\zeta \frac{1}{n}\sum_{i=1}^n \sigma_i\ell(\zeta, Z_i)\right] \quad (35)$$

其中，σ_i 是独立同分布的随机变量，以二分之一的概率取值为 *1* 或 *-1*。

证明。 该证明是学习理论中的经典结果（Bartlett & Mendelson，2002年），证明过程包括引入 ζ_n 和 ζ的经验风险，并通过 ζ 在整个定义域上的偏差上确界，对经验风险与总体风险之间的差异进行上界估计 ζ。随后，将总体风险替换为平均经验风险，并运用对称化技巧，引入基于(σ_i) 的可交换性而定义的变量(Z_i)。 □

引理32. 对于线性模型，拉德马赫复杂度可以界定为

$$\mathbb{E}_{(Z_i),(\sigma_i)}\left[\sup_{\|\zeta\|\le M}\frac{1}{n}\sum_{i=1}^n \sigma_i\langle Z_i, \zeta\rangle\right] \le \frac{M}{\sqrt{n}}\sqrt{\mathbb{E}[\|Z\|^2]}. \quad (36)$$

证明。 这是一个经典的 r 球约束预测器的拉德马赫复杂度结果（Bartlett &Mendelson，2002年）。 □

L引理 33。 此外，当 $h:\ \mathbb{R}\to\mathbb{R}$ 为利普希茨函数时，以下压缩原理成立

$$\mathbb{E}[\sup_f \frac{1}{n}\sum_{i=1}^n \sigma_i h(f(Z_i))] \le \sup\|\mathrm{d}h(x)\|\,\mathbb{E}[\sup_f \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)]$$

证明。 这源于利普希茨函数对空间容量的压缩（Vitushkin，1954年），关于机器学习背景下的证明，参见 Meir 和 Zhang（2003年）。 □

现在，我们可以基于拉德马赫复杂度陈述收敛性质。

引理 34。 设 $\Theta_n\in\mathbb{R}^k\otimes\mathcal{H}$ 是基于数据集 $\mathcal{D}_n$对 $\mathcal{L}$ 的无偏正则化经验版本的最小化器。假设 $\mathcal{D}_n$ 由 n 个输入样本（X_i）以及每个样本的 m 个数据增强（ξ_{ij}）构成，则平均超额风险的上界为

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{L}(S\Theta_n)] - \mathcal{L}(S\Theta) \le \frac{8\kappa^2\sup\|\Lambda\|_{HS}}{\sqrt{n}}\left(\frac{m+1+\beta}{m} + \frac{\sqrt{2}\kappa^2\sup\|\Lambda\|_{HS}\,(m^2+1)}{m^2}\right), \tag{37}$$

其中 κ 是 $\|\varphi\left(X\right)\|$ 的上界。

证明。 根据前面关于拉德马赫复杂度的引理，我们有

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}[\mathcal{L}(S\Theta_n);\lambda] - \mathcal{L}(S\Theta;\lambda) &\le 8\,\mathbb{E}_{\mathcal{D}_n,\sigma}\left[\sup_\Lambda \frac{1-\beta}{n}\sum_{i\in[n]}\sigma_i\left\langle\Lambda, \frac{1}{m}\sum_{j\in[m]}\varphi(\xi_{ij})\varphi(\xi_{ij})^\top\right\rangle\right] \\
&\quad + 8\,\mathbb{E}_{\mathcal{D}_n,\sigma}\left[\sup_\Lambda \frac{\beta}{n}\sum_{i\in[n]}\sigma_i\left\langle\Lambda, \frac{2}{m}\sum_{j\in[m/2];j+k-1=m}\varphi(\xi_{ij})\varphi(\xi_{ik})^\top\right\rangle\right] \\
&\quad + 4\,\mathbb{E}_{\mathcal{D}_n,\sigma}\left[\frac{2}{n}\sum_{i\in[n/2];i+j-1=n}\sigma_i\frac{1}{m^2}\sum_{k,l\in[m]}\left\langle\Lambda,\varphi(\xi_{ik})\varphi(\xi_{jk})^\top\right\rangle^2\right] \\
&\le \frac{8\sup\|\Lambda\|_{HS}}{\sqrt{n}}\left((1-\beta)\,\mathbb{E}_X\left[\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m\varphi(\xi_i)\varphi(\xi_i)^\top\right\|_{HS}^2\middle|X\right]\right]^{1/2}\right) \\
&\quad + \frac{8\sup\|\Lambda\|_{HS}}{\sqrt{n}}\left(\beta\,\mathbb{E}_X\left[\mathbb{E}\left[\left\|\frac{2}{m}\sum_{i,j=1}^{m/2}\varphi(\xi_i)\varphi(\xi_j)^\top\right\|_{HS}^2\middle|X\right]\right]^{1/2}\right) \\
&\quad + \frac{8\sup\|\Lambda\|_{HS}}{\sqrt{n}}\left(\sqrt{2}\sup\left|\left\langle\Lambda,\varphi(\xi)\varphi(\xi')^\top\right\rangle\right|\mathbb{E}\left[\left\|\frac{1}{m^2}\sum_{i,j=1}^m\varphi(\xi_i)\varphi(\xi_j)^\top\right\|_{HS}^2\right]^{1/2}\right).
\end{aligned}$$

为计算这些项，注意如果 (Z_i) 是独立同分布变量，

$$\mathbb{E}[\left\|\frac{1}{p}\sum_{i\in[p]}Z_i\right\|^2] = \mathbb{E}[\left\|\frac{1}{p}\sum_{i\in[p]}Z_i - \mathbb{E}[Z]\right\|^2] + \|\mathbb{E}[Z]\|^2 = \frac{1}{p}\,\mathbb{E}[\|Z-\mathbb{E}[Z]\|^2] + \|\mathbb{E}[Z]\|^2.$$

虽然可以逐项计算，但该引理的关键在于只需用 κ 对 φ 进行上界估计，因此可以通过将 φ 表示为 κ，得到其均值和标准差。 □

正文中的表达式由以下引理得出。

引理 35。 当 *mini* 通过最小化正则化风险，可以在约束 $\|\Lambda\|_{HS}\le\lambda^-$ 下缩小对 Θ 的搜索范围 ^1k.

证明。 进行正则化时，我们有

$$\|\Lambda\|_{HS} = \left\|\Theta^\top\Theta\right\|_{HS} \le \|\Theta\|_{\mathrm{op}} \|\Theta\|_{HS} \le \|\Theta\|_{HS}^2 ,$$

以及经验风险或总体风险的最小化器

$$\lambda \|\Theta\|^2 \le \mathcal{L}(S\Theta) + \lambda \|\Theta\|^2 \le \mathcal{L}(0) = k,$$

这解释了引理的陈述。 □

细心的读者可能会注意到，与 HaoChen 等人 (2021年) 的上界相比，我们获得了一个 $k^{-1/2}$ 倍的改进。事实上，在 HaoChen 等人 (2021年) 中，若采用 Maurer (2016) 的技巧，而非对以下对象的平凡上界：ˆ向量值函数空间的拉德马赫复杂度 $k\max_{i\in[k]} \mathcal{R}(\mathcal{F}_i)$，并使用 HaoChen 等人 （2021年）的记号，也可以得到这一因子。

C.2. 凸优化

当最小二乘问题具有平滑性或强凸性等额外结构时，凸优化的结果可能会带来改进，使其收敛速率优于 $n^{-1/2}$ 中通常的收敛速率。回顾凸优化的基本结果。

引理 36. 设 $\mathcal{L}(\Theta) = \mathbb{E}_Z[\ell(\Theta, Z)]$ 是定义在凸定义域上的凸函数。给定 n 个样本，采用无偏随机梯度下降并进 (Z_i), 行最终平均，可以达到以下超额风险

$$\mathbb{E}_{(Z_i)} \mathcal{L}(\hat{\Theta}) - \mathcal{L}(\Theta_*) \le \sqrt{2} M V n^{-1/2} \tag{38}$$

其中 $M^2 = \|\Theta_* - \Theta_0\|^2$ 且 $V^2 = \mathbb{E}[\|\nabla_\Theta \ell(\Theta, Z_i)\|^2]$。此外，若 $\mathcal{L}$ 是 α-光滑的，则可以达到

$$\mathbb{E}_{(Z_i)} \mathcal{L}(\hat{\Theta}) - \mathcal{L}(\Theta_*) \le \sqrt{2} M \sigma n^{-1/2} + \alpha M^2 n^{-1} \tag{39}$$

其中 $\sigma^2 = \mathbb{E}[\|\nabla\mathcal{L} - \nabla\ell\|^2]$。最后，当 $\mathcal{L}$ 是 α-强凸时，可以达到

$$\mathbb{E}_{(Z_i)} \mathcal{L}(\hat{\Theta}) - \mathcal{L}(\Theta_*) \le \frac{2V^2}{\alpha(n+1)}. \tag{40}$$

ˆ因此，给定 n 个数据样本，存在一个关于 Θ 的经验估计，可保证这些泛化界。

证明。 此引理可直接由以下定理推出： Bubeck (2015)中的定理 6.1、6.2 和 6.3。 □

需要注意的是，以 $\Lambda = \Theta^\top\Theta$为参数时， $\mathcal{L}$ 根据引理 30是一个二次型，但它是在非凸定义域上进行最小化的，即秩为 k的对称算子所构成的定义域。我们将放宽这一约束，并考虑在自伴正算子集合 Λ 上进行优化。这是合理的，因为即使放宽 Λ上的秩约束，定理 4仍能为下游任务提供保证。

为了利用引理36，应考虑 $\mathcal{L}$ 的无偏表达式。考虑如下小批量方案：对每个 X_i，分别采样两个输入 X_1、 X_2，以及 m 数据增强 ξ_{ij} ，形式化表示为

$$X_i \sim \mu_{\mathcal{X}}^{\otimes 2}, \qquad \xi_{ij} \sim \mu|_{X_i}^{\otimes m}. \tag{41}$$

这里， $\mu_{\mathcal{X}}$ 表示相对于 X的 μ 的边缘分布，它很可能与 $\rho_{\mathcal{X}}$相匹配； $\mu|_X$ 表示 Ξ关于 X的条件分布。

引理37。 $\mathcal{L}$ 的无偏公式基于定义为 ℓ 的量

$$\begin{aligned}\nabla_\Lambda \ell(S\Theta; \lambda) = \frac{2(\beta-1)}{m} \sum_{j\in[m]} \varphi(\xi_{1j})\varphi(\xi_{1j})^\top - \frac{2\beta}{m(m-1)} \sum_{1\le j\ne k\le m} \varphi(\xi_{1j})\varphi(\xi_{1k})^\top \\ + \frac{1}{m^2} \sum_{i,i'=1}^{2} \sum_{j,k=1}^{m} \left\langle \Lambda, \varphi(\xi_{ij})\varphi(\xi_{i'k})^\top \right\rangle \varphi(\xi_{ij})\varphi(\xi_{i'k})^\top .\end{aligned} \tag{42}$$

此外，当 $\mathcal{L}$ 经过正则化时，必须添加 $+\lambda I$ ，才能得到正则化 *ris* 上的梯度 k.

证明。 该公式由引理30得出。 □

为了对梯度范数的平方进行上界估计，可以使用以下引理 a.

引理 38. 对于 ℓ 在 (42)中给出的，梯度范数及其方差的上界为

$$\|\nabla_\Lambda \ell\| \le 2\kappa^2 + \kappa^4 \sup \|\Lambda\|, \qquad \textit{and} \qquad \mathbb{E}[\|\nabla_\Lambda \ell - \nabla \mathcal{L}\|^2] \le (\sigma_X^2 + m^{-1}\sigma_\xi^2)(1 + \sup \|\Lambda\|^2), \tag{43}$$

其中， σ_X 与 $\mathbb{E}[\psi(\xi) \mid X]$ 的方差有关，而σ_ξ 与$(\xi \mid X)$的平均方差有关。

证明。 将 $\nabla \ell$ 分解为 $\nabla \ell = a + b + c$ 中出现的三个项（42），我们有

$$\begin{aligned}
\|a\| &\le 2(1-\beta) \left\|\varphi(\xi)\varphi(\xi)^\top\right\| \le 2(1-\beta)\kappa^2 \\
\|b\| &\le 2\beta \left\|\varphi(\xi)\varphi(\xi')^\top\right\|^2 \le \beta\kappa^2 \\
\|c\| &\le \left\|\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top \right\rangle \varphi(\xi)\varphi(\xi')^\top\right\| \le \sup\|\Lambda\|\,\kappa^4.
\end{aligned}$$

为了对方差进行上界估计，可以如下进行

$$\mathbb{E}\,\|\nabla \ell - \nabla \mathcal{L}\|^2 \le 3\,\mathbb{E}\,\|a - \mathbb{E}[a]\|^2 + 3\,\mathbb{E}\,\|b - \mathbb{E}[b]\|^2 + 3\,\mathbb{E}\,\|c - \mathbb{E}[c]\|^2 .$$

我们先从 a中的部分开始,

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i\in[m]} \varphi(\xi_{1i})\varphi(\xi_{1i})^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i\in[m]} \varphi(\xi_{1i})\varphi(\xi_{1i})^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \,\middle|\, X = X_1]\right\|^2\right] \\
&\quad + \mathbb{E}\left[\left\|\mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \mid X = X_1] - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\right\|^2\right] \\
&= \frac{1}{m}\,\mathbb{E}_X\,\mathbb{E}_\xi\left[\left\|\varphi(\xi)\varphi(\xi)^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \mid X]\right\|^2 \,\middle|\, X\right] + \mathbb{E}\left[\left\|\mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \mid X] - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\right\|^2\right].
\end{aligned}$$

同样， b中的部分可以表示为

$$\begin{aligned}
\mathbb{E}[\|b - \mathbb{E}[b]\|^2] &= 2\beta\left(\frac{2}{m}\,\mathbb{E}_X\,\mathbb{E}_\xi\left[\left\|\varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \mid X]\right\|^2 \,\middle|\, X\right]\right) \\
&\quad + 2\beta\,\mathbb{E}\left[\left\|\mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \mid X] - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top]\right\|^2\right].
\end{aligned}$$

最后,

$$\begin{aligned}
\mathbb{E}[\|c - \mathbb{E}[c]\|^2] &= \frac{1}{m^2}\,\mathbb{E}_X\,\mathbb{E}_\xi\left[\left\|\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top \right\rangle \varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top \right\rangle \varphi(\xi)\varphi(\xi')^\top \mid X, X']\right\|^2 \,\middle|\, X, X'\right] \\
&\quad + \mathbb{E}\left[\left\|\mathbb{E}\left[\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top \right\rangle \varphi(\xi)\varphi(\xi')^\top \mid X, X'\right] - \mathbb{E}[\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top \right\rangle \varphi(\xi)\varphi(\xi')^\top]\right\|^2\right] \\
&= \frac{1}{m^2}\,\mathbb{E}_X\,\mathbb{E}_\xi\left[\left\|\left\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X']\right\rangle\right\|^2 \,\middle|\, X, X'\right] \\
&\quad + \mathbb{E}\left[\left\|\left\langle \Lambda, \mathbb{E}\left[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X'\right] - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top]\right\rangle\right\|^2\right]. \\
&\le \frac{1}{m^2}\,\|\Lambda\|^2\,\mathbb{E}_X\,\mathbb{E}_\xi\left[\left\|\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X']\right\|^2 \,\middle|\, X, X'\right] \\
&\quad + \|\Lambda\|^2\,\mathbb{E}\left[\left\|\mathbb{E}\left[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X'\right] - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top]\right\|^2\right].
\end{aligned}$$

因此，我们得到

$$\mathbb{E}\,\|\nabla \ell - \nabla \mathcal{L}\|^2 \le 3\left(2(1-\beta)\left(\frac{\sigma_{\xi,1}^2}{m} + \sigma_{X,1}^2\right) + 2\beta\left(\frac{2\sigma_{\xi,2}^2}{m} + \sigma_{X,2}^2\right) + \sup\|\Lambda\|^2\left(\frac{\sigma_{\xi,3}^2}{m^2} + \sigma_{X,3}^2\right)\right)$$

其中

$$\sigma_{\xi,1}^2 = \mathbb{E}_X \mathbb{E}_\xi \left[\left\| \varphi(\xi)\varphi(\xi)^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \mid X] \right\|^2 \middle| X \right]$$
$$\sigma_{X,1}^2 = \mathbb{E} \left[\left\| \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top \mid X] - \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top] \right\|^2 \right]$$
$$\sigma_{\xi,2}^2 = \mathbb{E}_X \mathbb{E}_\xi \left[\left\| \varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \mid X] \right\|^2 \middle| X \right]$$
$$\sigma_{X,2}^2 = \mathbb{E} \left[\left\| \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \mid X] - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top] \right\|^2 \right]$$
$$\sigma_{\xi,3}^2 = \mathbb{E}_X \mathbb{E}_\xi \left[\left\| \varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X'] \right\|^2 \middle| X, X' \right]$$
$$\sigma_{X,3}^2 = \mathbb{E} \left[\left\| \mathbb{E}\left[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top \mid X, X' \right] - \mathbb{E}[\varphi(\xi)\varphi(\xi')^\top \otimes \varphi(\xi)\varphi(\xi')^\top] \right\|^2 \right].$$

利用 $m^2 \geq m$ 这一事实，并选择合适的 σ_X 和 σ_ξ，即可得到该引理。 □

下面的引理陈述了 $\mathcal{L}$的凸性

引理 39. 作为 Λ的函数，目标函数 $\mathcal{L}$ 是 α-光滑的，且 $\alpha = \kappa^4$，其中 κ 是 $\|\varphi\|$的一个上界。此外，当 $\mathcal{X}$有限时，它是 α'-强凸的，其中 α' 是 $K = SS^\top$的特征值间隙的平方。

Proof. 这是引理 30的推论， $\mathcal{L}$ 是一个二次函数，其二次项为

$$\mathbb{E}[\langle \Lambda, \varphi(\xi)\varphi(\xi')^\top \rangle^2] = \langle \Lambda, \mathbb{E}[\varphi(\xi')\varphi(\xi')^\top] \otimes \mathbb{E}[\varphi(\xi)\varphi(\xi)^\top]\Lambda \rangle = \langle \Lambda, \Sigma \otimes \Sigma\Lambda \rangle .$$

换言之， $\mathcal{L}$ 的海森矩阵是 $\Sigma \otimes \Sigma \in \mathcal{H}^{\otimes 2} \otimes \mathcal{H}^{\otimes 2}$。因此，$\Sigma \otimes \Sigma \preceq \|\Sigma \otimes \Sigma\|_{\text{op}} I = \|\Sigma\|^2_{\text{op}} I \preceq \kappa^4 I$.类似地，$\Sigma \otimes \Sigma \succeq \left\|\Sigma^{-1}\right\|^{-2}_{\text{op}} I = \gamma_\xi^2 I$，其中 γ_ξ 是 Σ的特征值间隙，因此也是 K的特征值间隙。 □

在结束之前，还有一些问题需要解决。首先，尽管恒等算子不是希尔伯特-施密特算子，但应注意， λ 中的项只会在随机梯度下降中收缩距离。因此，优化正则化风险只会收缩下降轨迹（要对此进行形式化证明，可以参阅 Bubeck (2015)的证明）。最后，我们描述了自伴正算子空间中的一种下降过程，但没有纳入对秩的任何约束 Λ。请注意，根据引理35，可以将 Λ 的搜索范围限制在定义域$\|\Lambda\| \leq k_\lambda/\lambda$之内。最后，如果 Λ 使损失最小化 $\mathcal{L}$，那么可以证明，将其特征值阈值化，使其维数至多为 k，只会使损失 $\mathcal{L}$ 增加一个有界的乘法因子。我们指出，如果没有显式正则化，前文所述的带提前停止的随机梯度下降算法具有正则化效果，可在Lin 等人(2020 年)的谱滤波框架下进行研究。

D. 示例

本节旨在说明　　在简单分布下，借助调和分析技术， T和 K 的含义。

D.1. 比特序列（又称布尔超立方体）上的谐波分析

通过允许进行精确推导的设定，可以细致分析经典增强的作用。我们将聚焦于不变数据分布（例如均匀分布），以及由置换或坐标扰动构成、且能保持该分布不变的增强。虽然这类分布可能缺少真实数据中存在的结构，但它们有助于精确研究特定架构和增强的效果，而这些研究结果也可能部分适用于更贴近现实的数据。此类研究需要构造合适的 L^2 基，以便于研究核算子 K 和平滑算子 T（由数据增强定义）的作用。这与不变核的研究密切相关（例如，Bietti 等人，2021年；Bietti，2022；Mei 等人，2021年；Misiakiewicz 与 Mei，2022）。

这里我们将关注定义在布尔立方体 $\mathcal{X} = \{-1, +1\}^d$ 上、服从均匀分布的 n 位输入。为了充分运用调和分析工具，我们假设输入是从 $\mathcal{X}$ 上的均匀分布中采样得到的。在这种情况下，函数空间 $L^2(\mathcal{X}) = L^2(\mathcal{X}, \mathbb{R}, \mu_\mathcal{X})$ 通过通常的标量积定义，其中 $f, g\colon \mathcal{X} \to \mathbb{R}$,

$$\langle f, g \rangle = \mathbb{E}_{x\sim\tau}[f(x)g(x)] = \frac{1}{2^d} \sum_{x\in\mathcal{X}} f(x)g(x).$$

D.2. 数据增强的作用

现在让我们分析数据增强在布尔立方体上 T 的定义中的作用。为简便起见并便于记号，我们假设比特的索引按模 d 计算，例如 $x_{-1} = x_d$。

D.2.1. 通过奇偶函数进行研究

奇偶函数。 该空间中的一个有用基是奇偶函数，它们可以看作是这个L^2-空间中的傅里叶函数（O'Donnell，2014）。对于每个子集 $S \subseteq [d]$，其定义为统计 x 在该集合内的奇偶性

$$\chi_S(x) = \prod_{i \in S} x_i. \tag{44}$$

引理 40。 奇偶函数 χ_S 构成了 $L^2(\mathcal{X})$ 的标准正交基。

证明。 不难验证 $\langle \chi_S,\ \chi_S \rangle = 1$。若 $S \neq S'$，则不失一般性，存在一个 $i \in S \setminus S'$，于是有

$$\langle \chi_S, \chi_{S'} \rangle = \mathbb{E}_x[x_i \chi_{S \setminus \{i\}}(x) \chi_{S'}(x)] = \mathbb{E}_{x_i}[x_i]\, \mathbb{E}_{x_{-i}}[\chi_{S \setminus \{i\}}(x) \chi_{S'}(x)] = 0.$$

这证明了这个基的正交性。 □

让我们从那些易于用奇偶基研究的增强开始。

命题 41 (随机噪声**)**。 考虑以概率 p 翻转 x 的每一位；形式上，可通过以下操作表示

$$B_y^p(x) = x \odot y, \quad y \sim \mathrm{Ber}(\{-1, +1\}, p)^{\otimes d}, \tag{45}$$

其中，操作 $x \odot y$ 应用逐点乘法，而分布$\mathrm{Ber}(\{-1, +1\}, p)$ 以概率 p 返回值 -1，以概率 $1 - p$返回值 $+1$。在增强 $\xi = X \otimes y$下，T在奇偶基中被对角化，其中

$$T\chi_S = |1 - 2p|^{|S|}\, \chi_S. \tag{46}$$

换言之，T施加一个因子 $|1 - 2p|^{|S|}$，以减弱高阶傅里叶函数的影响。

证明。 回忆公式$g^\top T f = \mathbb{E}_X\, \mathbb{E}_{\xi,\xi'}\, [\langle f(\xi)\ g(\xi') \rangle \,|\, X]$。因此，令 y, y' 表示噪声串（每一位以概率 p等于 -1），且 $S \,\triangle\, S' = (S \cup S')/(S \cap S')$,

$$\begin{aligned}
\chi_S^\top T \chi_{S'} &= \mathbb{E}_X[\mathbb{E}_{y,y'}[\chi_S(X \odot y)\chi_{S'}(X \odot y')]] = \mathbb{E}_X\left[\mathbb{E}_{y,y'}\left[\prod_{i \in S} X_i y_i \prod_{j \in S'} X_j y'_j\right]\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_y\left[\prod_{i \in S \triangle S'} X_i y_i\right]\right] \mathbb{E}_{y,y'}\left[\prod_{i \in S \cap S'} y_i y'_i\right] \\
&= \mathbb{E}[\chi_{S \triangle S'}(X)] \cdot |1 - 2p|^{|S \triangle S'|} |1 - 2p|^{2|S \cap S'|} = |1 - 2p|^{|S|} \delta_{S,S'}.
\end{aligned}$$

因此，对于比特翻转增强，T在奇偶基下对角化。 □

命题42 (裁剪　　与掩蔽**)**。 考虑大小为 w的窗口内的裁剪操作，形式化定义如下：

$$[M_a^w(x)]_i = \begin{cases} x_i & \textit{if } i \in [a, a + w) \\ \mathrm{Ber}(\{-1, +1\}, 0.5) & \textit{otherwise} \end{cases}, \qquad a \sim \mathcal{U}([d]), \tag{47}$$

其中 $[a, a + w) = \{a, a + 1, \ldots, a + w - 1\}$，$a$ 从 $[d]$上的均匀分布中抽取，而分布伯努利$(\{-1, +1\}$,0.5) 以相同概率返回 $+1$ 和 -1 的随机比特，从而有效地掩蔽 $[a, a + w)$中窗口外的值。在增强 $\xi = M_a^w(X)$下，T在奇偶基下对角化，且

$$T\chi_S = \frac{\max\{1 + w - \mathrm{diam}(S), 0\}^2}{d^2} \cdot \chi_S \qquad \textit{with} \qquad \mathrm{diam}(S) = \min\{v \,|\, v, a \in [d]; S \subseteq [a, a + v)\}. \tag{48}$$

换言之，裁剪操作实际上会消除对高阶奇偶函数对应的核的任何依赖，只要其支撑集落在大小为 w的窗口之外。

证明。 在此设定下，

$$\begin{aligned}\chi_S^\top T\chi_{S'} &= \mathbb{E}_X[\mathbb{E}_{a,b}[\chi_S(M_a^w(X))\chi_{S'}(M_b^w(X))]] \\ &= \frac{1}{d^2}\sum_{a,b=1}^{d}\mathbb{E}_{X,\nu,\nu'}\left[\prod_{i\in S\cap[a,a+w)}x_i\prod_{i'\in S\setminus[a,a+w)}\nu_{i'}\prod_{j\in S'\cap[b,b+w)}x_j\prod_{j'\in S'\setminus[b,b+w)}\nu'_{j'}\right] \\ &= \frac{1}{d^2}\sum_{a,b=1}^{d}\mathbf{1}_{S\subseteq[a,a+w)}\,\mathbf{1}_{S'\subseteq[b,b+w)}\,\mathbb{E}_X[\chi_S(X)\chi_{S'}(X)] = \frac{1}{d^2}\sum_{a,b=1}^{d}\mathbf{1}_{S\subseteq[a,a+w)}\,\mathbf{1}_{S'\subseteq[b,b+w)}\,\delta_{S,S'} \\ &= \left(\frac{1}{d}\sum_{a=1}^{d}\mathbf{1}_{S\subseteq[a,a+w)}\right)^2\delta_{S,S'}.\end{aligned}$$

总和的项数与 S的直径有关。 □

命题 43 (二维裁剪**)**。 考虑二维设置 $\mathcal{X}=\{-1,+1\}^{m\times d}$，其中输入按 $m\times d$网格排列。考虑对大小为 $v\times w$的窗口进行裁剪操作，形式化表示为

$$[M_{a,b}^{v\times w}(x)]_{i+jm}=\begin{cases}x_{i+jm} & \text{if } i\in[a,a+v), j\in[b,b+w)\\ \mathrm{Ber}(\{-1,+1\},0.5) & \text{otherwise}\end{cases},\qquad (a,b)\sim\mathcal{U}([m]\times[d]). \tag{49}$$

在数据增强 $\xi=M_{a\,b}^{v\times w}$，(X)下， T 可在奇偶基中对角化，并且

$$T\chi_S=\frac{1}{m^2d^2}\left(1+v-\mathrm{diam}_{e_1}S\right)_+^2\cdot\left(1+w-\mathrm{diam}_{e_2}S\right)_+^2\chi_S, \tag{50}$$

其中，直径$_{e_1}S$ 是 S 投影到第一维后的直径。

证明。 这可由一维情形的证明得出。 □

命题 44 (翻转**)**。考虑这样一个算子：它以概率 p将索引逆序排列，形式化地说，

$$[R(x)]_i=x_{-i}. \tag{51}$$

在数据增强 $\xi=R(X)$下，

$$T=(1-2p+2p^2)I+2p(1-p)J, \tag{52}$$

其中， J是一个对合，它将任意集合 S映射到其镜像 $S=\{-i\,|\,i\in S\}$。在这种情况下， T由$\sqrt{2}(\chi_S+\chi_{\tilde{S}})$和$\sqrt{2})(\chi_S-\chi_{\tilde{S}})$对角化，其中 $S\subseteq[d]$。

证明。在这种情况下，

$$\begin{aligned}\chi_S^\top T\chi_{S'} &= \left((1-p)^2+p^2\right)\mathbb{E}_X\left[\chi_S(X)\chi_{S'}(X)\right]+2p(1-p)\,\mathbb{E}_X\left[\chi_{\tilde{S}}(X)\chi_{S'}(X)\right]\\ &= (1-2p+2p^2)\delta_{S,S'}+2p(1-p)\delta_{\tilde{S},S'},\end{aligned}$$

这解释了该引理。 □

注记 45。到目前为止，我们研究了空间 $L^2(\mathcal{X},\mathbb{R},\mu_X)$ 中的所有算子，而正文考察的是 $L^2(\mathcal{X},\mathbb{R},\mu_\Xi)$中的算子；这是合理的，因为此前研究的所有变换都使均匀分布保持不变，因此

$$L^2(\mu_\mathcal{X})=L^2(\mu_\Xi). \tag{53}$$

D.2.2. 循环奇偶函数中的平移研究

为了研究由置换，更具体地说是平移，构成的增强，奇偶基并不适合对角化 T。因此，我们在下文中定义了另一种融入循环对称性的基（Misiakiewicz 与 Mei，2022年）。我们注意到，也可以对其他分布开展类似研究，例如球面、球面的乘积或环面上的均匀分布（Bietti 等人，2021年；Bietti，2022；Favero 等人，2021年；Mei 等人，2021年）。

循环奇偶函数。 函数 χ_S 是多项式，可按其次数 $\ell = |S|$ 归类到空间 $V_{d,\ell}$ 中；这些空间的直和构成完整的 $L^2(\mathcal{X})$ 空间，其中

$$\dim V_{d,\ell} = |\{S \subseteq [d], |S| = \ell\}| = \binom{d}{\ell}.$$

这些不同的空间还可以进一步分解为群作用下的轨道。具体而言，对于置换群 $G = \mathfrak{S}_d$，我们定义作用 $\mathcal{A}$：$G \times \mathcal{X} \to \mathcal{X}$，记作 $\mathcal{A}(a,\ x) = a \cdot x$，如下：

$$(a \cdot x)_i = x_{a^{-1}(i)}.$$

为了通过调和分析具体研究数据增强，我们更具体地考察平移作用，它构成置换的一个子群。为简便起见，我们将这个群记作 $[d]$，它可理解为 $\mathbb{Z}/d\mathbb{Z}$，作用于 $\mathcal{X}$ 上，如下：

$$(a \cdot x)_i = x_{i-a}$$

其中 $i - a$ 按模 d 理解。将此作用的轨道定义为 $\{S + a \,|\, a \in [d]\}$，其中 $S \subseteq [d]$。在这些不同的轨道上，可以定义以下"循环奇偶函数"$\psi_{m,\ S}$：$\mathcal{X} \to \mathbb{C}$:

$$\psi_{m,S} = \frac{1}{\sqrt{k_S}} \sum_{k \in [k_S]} e^{2i\pi km/k_S} \chi_{S+k} = \frac{\sqrt{k_S}}{d} \sum_{k \in [d]} e^{2i\pi k \frac{md}{k_S}/d} \chi_{S+k} \qquad \text{where} \qquad k_S = |\mathrm{orb}(S)|\,, \tag{54}$$

其中 $m \in [k_S]$，而 S 被视为一个轨道的代表元。

引理 46。 对于 $m \in [k_S]$，且 S 是平移作用下每个轨道的一组代表元，循环奇偶函数$(\psi_{m,\ S})$构成 $L^2(\mathcal{X}, \mathbb{C}, \mu)$ 的一组正交基，其中 μ 是 $\mathcal{X}$ 上的均匀测度。此外，它们还将算子A：$L^2 \to L^2$对角化，该算子定义为 $Af(x) = f(a \cdot x)$，对任意 $a \in [d]$ 均成立。

证明。 第一部分由以下事实得出：$L^2(\mathcal{X})$ 可以分解为与 $V_{d,\ell}$ for$\ell \in [0, d]$ 相关联的直和；每个子空间都可以分解为平移作用 $\mathrm{orb}(S) = \{S + a \,|\, a \in [d]\}$ 的轨道（注意，平移不会改变集合 S 的基数）。这些后续空间可以通过离散傅里叶变换进行参数化，从而得到 $\psi_{m,S}$。

找到这些基的一种自然方式，是尝试对算子 T 进行对角化，使得 $(\chi_S^\top T \chi_{S'})_{S,S' \subseteq [d]}$ 为分块对角形式，其中每个块对应于某个轨道上的循环矩阵，并且可通过离散傅里叶变换对角化。引理中的算子尤其属于这种情况

$$A\chi_S = \chi_{S+a}$$

只有当 $[d] \cdot S$ 与 $[d] \cdot S'$ 相交时，上述项才非零，这意味着 $\mathrm{orb}(S) = \mathrm{orb}(S')$，由此构造出分块对角结构。将 i 块中的元素按 $S_{i,k} = S_i + k$ for $k \in [d]$ 编号，我们有

$$\chi_{S_{i,k}}^\top A \chi_{S_{i,k'}} = \mathbf{1}_{S_{i,k} = S_{i,k'} + a}, = \mathbf{1}_{S_i + k = S_i + k' + a}, = \mathbf{1}_{k - k' = a},$$

它只取决于 $(k - k')$ 的值。因此，上述每个块都是一个由离散傅里叶变换对角化的循环矩阵。该矩阵的特征向量为

$$v_m = \frac{1}{\sqrt{k_S}} \sum_{k \in [k_S]} e^{2i\pi km/k_S} e_k, \qquad \text{where} \qquad k_S = |\mathrm{orb}(S)|\,,$$

对于 $m \in [k_S]$，相应的特征值为

$$\mu_m = \sum_{k \in [k_S]} c_{k_S - k} \exp\left(\frac{2i\pi km}{k_S}\right),$$

其中

$$c_i = \mathbf{1}_{i=a},$$

利用我们为 $e_i \simeq \chi_{S+i}$ 写出那些矩阵这一事实，可以得到该引理。 □

对 ope 的研究　　算子 T可借助其平方根 A 进行简化： $L^2(\mu_\Xi) \mapsto L^2(\mu_\mathcal{X})$形式定义为

$$Af(x) = \mathbb{E}_\xi \left[f(\xi) \,|\, X = x \right] \tag{55}$$

并进行验证

$$\langle f, Tg \rangle_{L^2(\mu_\Xi)} = \mathbb{E}_X \, \mathbb{E}_{\xi,\xi'} [f(\xi) g(\xi') | X] = \langle Af, Ag \rangle_{L^2(\mu_\mathcal{X})}. \tag{56}$$

当 $\mu_\mathcal{X}$ 在置换作用下不变时，这种分解将特别有用，这意味着$\mu_\mathcal{X} = \mu_\Xi =: \mu_\circ$

引理 47. 在均匀布尔设置中，当增强定义为 $\xi = a \cdot X$ *where* a 是从概率分布 $p \in \Delta_{\mathfrak{S}_d}$中抽取的一个置换时，

$$Tf(x) = \sum_{a,b \in \mathfrak{S}_d} p(a) p(b) f((a^{-1}b) \cdot x).$$

证明。 T的平方根定义为 $Af(x) = \sum_{a \in \mathfrak{S}_d} p(a) f(a \cdot x)$。我们重点考虑 $p(b) = \delta_{a=b}$ 的情形，并利用 $\mu_\mathcal{X}$ 是均匀测度，因而在平移下保持左不变这一事实，计算 A 的伴随算子，其

$$\begin{aligned}
\langle Af, g \rangle_{L^2(\mu_\mathcal{X})} &= \frac{1}{2^d} \sum_{x \in \mathcal{X}} Af(x) g(x) = \frac{1}{2^d} \sum_{x \in \mathcal{X}} p(a) f(a \cdot x) g(x) \\
&= \frac{1}{2^d} \sum_{x \in \mathcal{X}} p(a) f(a \cdot x) g(a^{-1} \cdot a \cdot x) = \frac{1}{2^d} \sum_{x \in \mathcal{X}} p(a) f(x) g(a^{-1} \cdot x) \\
&= \left\langle f, x \mapsto p(a) g(a^{-1} \cdot x) \right\rangle_{L^2(\mu_\mathcal{X})}. \\
&= \left\langle f, A^\top g \right\rangle_{L^2(\mu_\Xi)}.
\end{aligned}$$

在一般情况下，由线性性可得，

$$A^\top f(x) = \sum_{a \in \mathfrak{S}_d} p(a) f(a^{-1} \cdot x).$$

计算 $T = A^\top A$ 即可得出结果。注意，如果我们进一步假设 p 是对称的（即 $p(a) = p(a^{-1})$），那么我们有 $A^\top = A$，因此 $T = A^2$。 □

这使我们能够更精细地刻画平移对算子 T的影响。

命题 48 (平移**).** 考虑形式上定义如下的平移算子

$$[T_a(x)]_i = x_{i-a}, \qquad a \sim p \in \Delta_{[d]} \tag{57}$$

U在数据增强 $\xi = T_a(X)$下， T 在 $\mathbb{C}$ 中由循环奇偶函数(54)对角化。

$$T\psi_{m,S} = \frac{d^2}{k_S^2} \left| \hat{p}\left(\frac{md}{k_S}\right) \right|^2 \psi_{m,S}, \tag{58}$$

其中$\hat{}p$ 是 p的傅里叶变换，对 $\omega \in [d]$ 定义为

$$\hat{p}(\omega) = \sum_{a \in [d]} p(a) \exp\left(\frac{-2i\pi a\omega}{d}\right) \tag{59}$$

证明。 在平移的情况下，我们有

$$Af(x) = \sum_{a \in [d]} p(a) f(a \cdot x) = \sum_{a \in [d]} p(a) A_a f(x),$$

若 A_k 是将 f 与 $x \to f(k \cdot x)$ 关联起来的算子，它就是一个平移算子。重新采用引理 46 的证明，并取 $A_k \psi_{m,S} = e^{-2i\pi km/k_S} \psi_{m,S}$，可得

$$A\psi_{m,S} = \sum_{a \in [k_S]} p(a) \exp\left(\frac{-2i\pi am}{k_S}\right) \psi_{m,S} = \sum_{a \in [d]} \frac{d}{k_S} \cdot p(a) \exp\left(\frac{-2i\pi am}{k_S}\right) \psi_{m,S} = \frac{d}{k_S} \cdot \hat{p}\left(\frac{md}{k_S}\right) \psi_{m,S},$$

and

$$T = A^*A = \sum_{m,S} \frac{d^2}{k_S^2} \left|\hat{p}\left(\frac{md}{k_S}\right)\right|^2 \psi_{m,S}\psi^*_{m,S}.$$

这就证明了该引理。 □

现在我们展示，不同的平移采样分布如何在算子 T 中产生不同的平滑效应。

示例 9（平移的平滑效应）。 为了了解增强强度的影响，考虑一种平移分布 p，其形式为 $p(a) = \omega p_0(\omega a)$，其中 p_0 是一种总和为 1 的局部化窗口形状（例如均匀分布或高斯分布）。这里，$\omega \approx 1/\Delta$ 与窗口大小 Δ 成反比，窗口大小控制着数据增强的“强度”或范围。于是我们有

$$|\hat{p}(m)|^2 = |\hat{p}_0(m/\omega)|^2 \approx |\hat{p}_0(\Delta m)|^2.$$

这里，平方傅里叶系数 $|\hat{} p_0(m)|^2$ 通常会随频率 m 增大而衰减，这表明 T 具有平滑效应，会惩罚特征函数 $\psi_{m,S}$ 中 m 较大的那些，即振荡更快的特征函数。上述公式还表明，增强强度 Δ 的增加会使其随 m 更快衰减，同时不影响平移不变特征函数（$m = 0$）。

D.3. 架构的作用

一种特别有用的特征空间 φ，用于定义线性函数类 Ψ，是以下集合：

$$\varphi : \mathcal{X} \mapsto \mathbb{R}^{2^d}; x \mapsto (e_S \chi_S(x))_{S\subseteq[d]}. \tag{60}$$

对于任意序列 $(e_S) \in \mathbb{R}^{2^d}$。这种形式的线性模型可以在奇偶基下对角化，从而能够有效研究数据增强的作用与架构的作用之间的相互作用。

引理 49. 对于通过特征定义的任意线性模型 φ 在 (60)中，积分算子 $K\colon L^2(\mathcal{X}) \mapsto L^2(\mathcal{X})$ 在奇偶基下对角化，

$$K\chi_S = e_S^2 \chi_S. \tag{61}$$

证明。 这是由以下事实得出的：$Kf(x) = d^{-1}\sum_{x'\in[d]} k(x, x')f(x')$，其中 $k(x, x') = \varphi(x)^\top \varphi(x')$. □

这些函数类别中包括满足 $k(x, y) := \varphi(x)^\top\varphi(y) = \tilde{h}(\|x - y\|^2) = h(x^\top y)$ 的点积核。这些核再次表明，它们特别适合布尔超立方体的傅里叶几何。

引理 50（点积核的谱分解）。 任意点积核都可在奇偶基下对角化。具体而言，存在 $(\nu_i)_{i\in[0,d]} \in \mathbb{R}^{d+1}$，使得当 $\mu_{\mathcal{X}}$ 是超立方体上的均匀分布时，

$$K\chi_S = \nu_{|S|}\chi_S. \tag{62}$$

证明。 可以验证，$x^\top y = d - 2k$ 表示 k 中 x 与 y不同的比特数。定义 Q_ℓ 次数为ℓ、由次数为 ℓ的多项式取平均得到的多项式如下：

$$\sum_{S\subseteq[d],|S|=\ell} \chi_S(x)\chi_S(y) = \binom{d}{\ell} Q_{\ell,d}(\langle x, y\rangle), \tag{63}$$

对于任意布尔字符串 x 和 y。$Q_{\ell,d}$ 是定义良好的，因为左侧具有平移不变性。此外，利用 χ_S 的正交性，可以证明 $(Q_{\ell,d})_{\ell\in[0,d]}$ 构成 $\{d - 2k \mid k \in [0, d]\}$上的函数基。更确切地说，$m \to \binom{d}{\ell}^{-1/2} Q_{\ell,d}(m)$ 是赋予 L^2 空间以均匀分布经映射 $x \to \langle x, y\rangle$（对任意固定的 y）所得推前测度后的标准正交基，并且维数相匹配。因此，存在 ν_ℓ 使得

$$h(\langle x, y\rangle) = \sum_{\ell\in[0,d]} \nu_\ell \binom{d}{\ell} Q_{\ell,d}(\langle x, y\rangle)$$

其中， ν_ℓ 可以通过计算 h 与 Q_ℓ 在 $L^2(\tau)$ 中的标量积得到。

$$\nu_\ell = \langle h, Q_\ell \rangle_{L^2(\tau)} . \tag{64}$$

最后，利用在均匀设定下， $Kf(x) = d^{-1}\sum_{x'\in[d]} k(x, x')f(x')$，其中 $k(x, x') = \varphi(x)^\top\varphi(x')$，我们得到

$$K\chi_S(x) = \mathbb{E}_Y[h(\langle x, Y\rangle)\chi_S(Y)] = \sum_\ell \nu_\ell \sum_{S'\subset[d],|S'|=\ell} \chi_{S'}(x)\,\mathbb{E}[\chi_{S'}(Y)\chi_S(Y)] = \nu_\ell\chi_S(x).$$

这就完成了该引理的证明。 □

引理 49 也可以在球面上证明。其证明展示了 Q_ℓ，它们是归一化的勒让德（或 Gegenbauer）多项式。详情参见 例如 Smola 等人 (2000年)； Bietti 等人 (2021年)； Mei 等人 (2021年)。注意，对于球面上常见的核函数，例如 NTK 中出现的核函数， ν_k 会随 k (Bach，2017年； Bietti 与 Mairal，2019年) 多项式衰减。

特征（60）足够丰富，可以描述具有全连接层或卷积层的简单架构的神经切向核。首先，我们描述这类 NTK 的一般形式如下。

命题 51 (简单网络的线性化)。 定义一种简单的神经网络架构如下：

$$f(x) = \sqrt{\frac{\Delta}{N\omega d}} \sum_{i\in[N]} \sum_{k\in[d/\Delta]} a_{ik} \sum_{s\in[\omega]} \sigma\left(\left\langle w_i, x^{(q)}_{(k\Delta+s)}\right\rangle\right), \tag{65}$$

其中 $x^{(q)}{}_{(k)} = (x_k, x_{k+1}, \cdots, x_{k+q-1})$ 是大小为 q 的局部补丁（索引按模 d 定义）， w_i 是从旋转不变分布初始化的权重 $\mathcal{W}$, σ： $\mathbb{R}\mapsto\mathbb{R}$ 是激活函数， $\omega\in\mathbb{N}$ 是平均池化窗口的大小， $\Delta\in\mathbb{N}$ 是池化窗口， N 是通道数。该网络在初始化附近的线性化得到核

$$k(x, x') = \varphi(x)^\top\varphi(x') = \frac{\Delta}{d\omega} \sum_{k\in[d/\Delta]} \sum_{s,s'\in[\omega]} h\left(\left\langle x^{(q)}_{(k\Delta+s)}, y^{(q)}_{(k\Delta+s')}\right\rangle/q\right) \tag{66}$$

其中

$$h(\langle u, v\rangle/q) = \mathbb{E}_{w\sim\mathcal{W}}\left[\sigma(\langle u, w\rangle/\sqrt{q})\sigma(\langle v, w\rangle/\sqrt{q}) + \sigma'(\langle u, v\rangle/\sqrt{q})\sigma'(\langle u, v\rangle/\sqrt{q})\cdot\langle u, v\rangle/q\right. \tag{67}$$

证明。 例如，Misiakiewicz 与 Mei (2022）的命题 3 给出了这样的线性化). □

命题 化（全连接网络的线性化）。单隐藏层全连接层 r

$$f_{FC}(x) = \frac{1}{\sqrt{N}} \sum_{i\in[N]} a_{ik}\sigma(w_i^\top x),$$

可以线性化为一个点积核： $k_{FC}(x, y) = h(x^\top y/d)$，其中 h 定义见(67)。此外，所得积分算子 K_{FC} 在奇偶基下对角化为

$$K_{FC}\chi_S = \nu_h(d, |S|)\chi_S,$$

其中系数为 $\nu_h(d, \ell) = \langle h, Q_\ell\rangle_{L^2(\tau)}$ 如(64)所示。

注意，特征值 $\nu_h(d, \ell)$ 随 ℓ递减；对于固定的 ℓ，当 d 较大时，它们满足 $\nu_h(d, \ell) = \Theta_d(d^{-\ell})$。更一般地，可以证明$\lim_{d\to\infty} d^k\nu_h(d, \ell) = \frac{d^k}{dt^k}h(t)\big|_{t=0}$。 —

证明。 第一部分可直接由前述命题推出，其中 $\omega = 1$ 和 $q = \Delta = d$。第二部分依据引理 50，以及（63）。关于特征值的论述，参见（Yang & Salman，2019年）。 □

命题 53（线性 卷积网络的.....化）。一个卷积层后接一个全连接层

$$f_{CNN}(x) = \frac{1}{\sqrt{Nd}} \sum_{i\in[N]} \sum_{k\in[d]} a_{ik}\sigma\left(w_i^\top x_{(k)}^{(q)}\right),$$

可以通过 h的（67）进行线性化

$$k_{CNN}(x,y) = \frac{1}{d}\sum_{k\in[d]} h\left(\left\langle x_{(k)}^{(q)}, y_{(k)}^{(q)}\right\rangle / q\right).$$

在布尔设置下， 所得积分算子 K_{CNN} 在奇偶基和循环基下均被对角化为

$$K_{CNN}\psi_{m,S} = \begin{cases} \nu_h(q,|S|)\frac{(q+1-\mathrm{diam}(S))_+}{q}\psi_{m,S}, & \textit{if } \mathrm{diam}(S)\le q, \\ 0 & \textit{otherwise.}\end{cases}$$

其中 ν_h（q， ℓ）由命题*52*定义。

证明。 第一部分对应于 $\omega = \Delta = 1$ 的情形。第二部分源于 h 在 $\mathbb{Q}_\ell$ 基上的展开，因此得到（详情见 Misiakiewicz 与 Mei (2022) 中的式 (30)）

$$k_{CNN}(x,y) = \sum_{S\subseteq[d],\mathrm{diam}(S)\le q} q^{-1}\nu_h(q,|S|)(q+1-\mathrm{diam}(S))_+\chi_S(x)\chi_S(y).$$

K使 $\{S\,||S| = a,\mathrm{diam}(S) = b\}$ 保持不变这一事实（因为特征值仅取决于 $|S|$ 和 $\mathrm{diam}(S)$），使得我们能够从奇偶基转换到循环基。 □

当池化被纳入核中，且 $\omega > 1$ 满足式（65）时，架构会强制实现局部平移不变性。一个简单的例子是全局平均池化 $\omega = d$，它强制实现对平移的严格不变性，并将奇偶函数投影到其轨道中各元素之和上，从而形成特征基。在这种情况下， K 不再在奇偶基下对角化，而是在循环奇偶函数基下为对角形式。

D.4. 数据增强与架构之间的相互作用

在均匀布尔设置中，数据增强与架构之间的相互作用之所以容易分析，是因为许多算子 K 和 T彼此可交换。

引理 54. 在均匀布尔设置中，与点积核相关的算子 K 与所有由逐位噪声、裁剪、平移或索引翻转构成的算子 T 均可交换。

证明。 对于均匀设定中的点积核，其空间 $V_{d,\ell}$ 是 K的特征子空间。这些空间在所有通过常见数据增强定义的 T 作用下保持不变，因为平移和索引翻转操作会保持子集的基数。因此， K 和 T可以在同一组基下对角化，故二者可交换。 □

根据前述引理，与全连接层线性模型相关的积分算子 K 与所有为常见数据增强定义的算子 T均可交换。对于卷积层，若其 T来源于随机噪声、裁剪或平移，情况也同样如此。[6] 因此，借助命题 4，可以轻松研究架构与数据增强之间的相互作用。

示例 *10*（全连接核与平移增强之间的相互作用）。回顾示例 *9*，从局部窗口中采样平移时， T的特征值具有 $|\hat{}p(m)|^2$ 的形式，并且通常随频率索引 m 在 $\psi_{m,S} = \frac{1}{\sqrt{k_S}}\sum_{k\in[k_S]} e^{2i\pi km/k_S}\chi_{S+k}$ 中递减，其中任意集合 S 都不具有周期性。相比之下，对于特征函数 $\psi_{m,S}$，K 的特征值 $\nu_h(d,|S|)$ 按 $\Theta_d(d^{-|S|})$ 递减，且与 m 无关。因此，使用参数 λ 进行正则化后，预训练会将特征值缩减至 $|\hat{}p(m)|^2 - \lambda\nu_h(d,|S|)^{-1}$。这尤其会消除来自特征函数$\psi_{m,S}$ 的贡献，其中 m 较小（即接近不变），而 $|S|$ 较大。图 *3* 和图 *5* 展示了示意图。

[6]因为它使平移轨道保持不变。

示例 *11*（卷积神经网络的核与平移增强之间的相互作用）。考虑与示例 *10* 相同的情形，即从局部窗口中采样平移。对于补丁宽度为 q 的单层卷积神经网络，特征函数对应于奇偶函数 χ_S 或循环奇偶函数 $\psi_{m,S}$，其中 $\mathit{diam}(S) \le q$，相应的特征值为 $\nu_h(q,\ell)^{\frac{q+1-\mathrm{diam}(S)}{d}}$。这里，对于直径大于 q 的 T 上的特征函数 $\psi_{m,S}$，无论正则化强度 λ 为何，其贡献都会被完全消除，。对于 $\mathit{diam}(S) \le q$ 的特征函数 $\psi_{m,S}$，卷积神经网络会将其贡献缩减至 $|\hat{p}(m)|^2 - \lambda(\nu_h(q,\ell)^{\frac{q+1-\mathrm{diam}(S)}{d}})^{-1}$；$\mathit{diam}(S)$ 越大，缩减幅度越大。

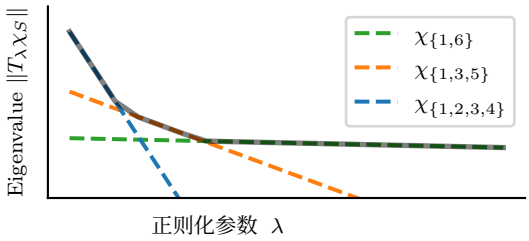

图 *8*。展示相互作用的示意图 随 λ 变化的 T 与 K，其中 K 是一个两层 ReLU 网络的神经切线核，而 T 在 12 位输入上执行窗口大小为 8 的裁剪。这里，我们绘制了两个算子的特征基中三种不同奇偶函数的特征值。直径较大的奇偶函数在 T 中的特征值较小（此处，直径最大的奇偶函数是 $\chi_{\{1,6\}}(X) = X_1X_6$）。相比之下，$K$ 的特征值偏向于支撑在较少比特上的奇偶函数。因此，较小的正则化会偏向直径较小的奇偶函数，而增加正则化则会惩罚基数较大的奇偶函数。

D.5. 关于球面设置的说明

在实验中，我们还考虑了球面 $\mathcal{X} = \mathbb{S}^{d-1}$ 上的均匀数据设置，增强方式为置换，所用点积核为 $\varphi(x)^\top\varphi(y) = h(x^\top y)$。在这种情况下，$L^2(\mathcal{X})$ 的一个自然基函数选择是球谐函数（Efthimiou & Frye，2014）。这些函数由齐次调和多项式构成；与奇偶函数的情形类似，它们可以按次数分组，从而得到任意次数 $\ell \ge 0$ 的球谐函数的正交空间 $V_{d,\ell}$，其中

$$N(d,\ell) := \dim V_{d,\ell} = \frac{2\ell+d-2}{\ell}\binom{\ell+d-3}{d-2}.$$

众所周知，对于点积核，K 在这样的基下是对角的（Smola 等人，2000年；Bach，2017年），其特征值会衰减，且仅取决于次数 ℓ。这些特征值与超立方体情形类似，可表示为

$$\nu_h(d,\ell) = \mathbb{E}_{t\sim\tau}[h(t)Q_{\ell,d}(t)],$$

其中 $Q_{\ell,d}$ 现在是次数为 ℓ 的勒让德（Gegenbauer）多项式，它们关于不同的测度 $d\tau(t) = (1-t^2)^{\frac{d-3}{2}}dt$ 在 $[-1,1]$ 上正交。

由于空间 $V_{d,\ell}$ 在算子 $T = A^\top A$ 作用下保持不变，因此可以证明，存在一组球谐函数的选取，使其同样将 T 对角化（例如参见 Bietti 等人，2021年，引理 12）。于是，我们可以将该基下的特征值 $\lambda_{\ell,j}$ 视为 T 对应谐波 $Y_{\ell,j}$ 不变性的体现；特别地，当 $\lambda_{\ell,j} = 1$ 时，$Y_{\ell,j}$ 对所有数据增强均不变，而当 $\lambda_{\ell,j} < 1$ 时，则不变或仅部分不变。

按 $\lambda_{k,j}$ 固定 k 递减排序 递减 j，T 与 K 之间的相互作用与文中所述的情况相似，例如图 3 .

E. 实验

E.1. 实现细节

此前，我们深入研究了 $\mathcal{H}$ 在 L^2 中的嵌入，其定义为 $S:\ \mathcal{H} \mapsto L^2;\ \theta \to \varphi(\cdot)^\top\theta$。给定样本 $(\xi_{ij})_{i\le n,\ j\le m}$，关于 $\mathcal{H}$ 的所有操作都可以简化为 $\varphi(\xi_{ij})$ 的张成空间（这称为表示定理），而 S 则可以简化为嵌入 $\hat{S}:\ \mathcal{H} \mapsto \mathbb{R}^{nm};\ \theta \to (\frac{1}{nm}\varphi(X_{ij})^\top\theta)_{ij}$。这便引出了实现

$$\hat{T}_\lambda = \beta + (1-\beta)\hat{T} - \lambda\hat{K}.$$

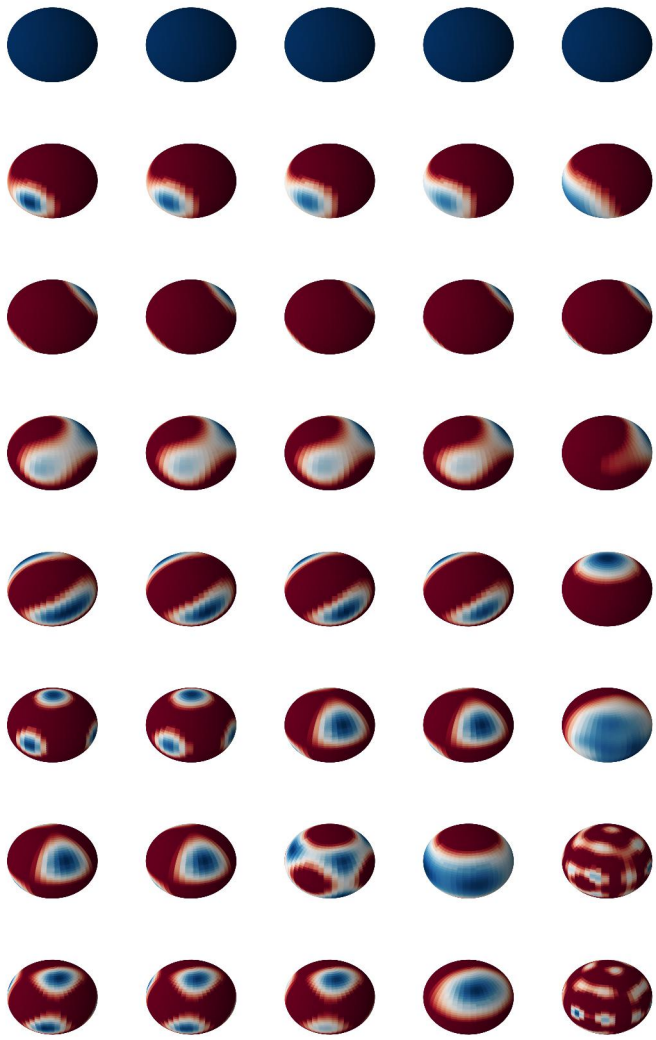

图 9。 图 4 的扩展。 第 i 行表示 i 阶特征函数 T_λ（按特征值降序排列）。正则化程度 λ 沿列方向逐渐增加，增幅为 $\lambda \in \{0, .1, 1, 10, 100\}$。较小的 λ 偏向于选择对这里所选平移增强保持不变的函数，而较大的 λ 则偏向于选择球面上更平滑的函数，在此设置下对应于低阶球谐函数。最右侧的两个结果是 K 的伪逆不稳定性所导致的伪影（从而得到实现 $\varphi K^{-1}\varphi = 0$，而我们定义了 $\varphi^\top K^{-1}\varphi = +\infty$ 当 $K\varphi = 0$时）。

其中，$T \in \mathbb{R}^{nm}$ 是　矩阵，其值如下所示，其中我们用 ij 对 $\mathbb{R}^{nm}$ 中的元素进行索引，且 $i \in [n]$ 和 $j \in [m\ ]$，

$$\hat{T} = I + \sum_{ijk} e_{ij} e_{ik}^{\top},$$

而 K 是定义如下的 Gram 矩阵

$$nm \cdot e_{ij}^{\top} K e_{kl} = k(\xi_{ij}, \xi_{kl}) = \varphi(\xi_{ij})^{\top} \varphi(\xi_{kl}).$$

ˆ注意，矩阵 $T - I$ 可以视为一个图的邻接矩阵，当且仅当两个数据增强来自同一个输入时，它们之间才会连接。等价地，T 可以视为一个拉普拉斯矩阵。$\mathbb{R}^{nm}$ 中 T_λ 的一个特征向量会借助 $S\hat{S}^{-1} = S\hat{S}(\hat{S}\hat{S}^{\top})^{-1} = K_x^{\top} K^{-1}$ 投影回 L^2，其中

$$nmK_x = \left(\varphi(x)^{\top} \varphi(\xi_{ij})\right)_{ij} \in \mathbb{R}^{nm}.$$

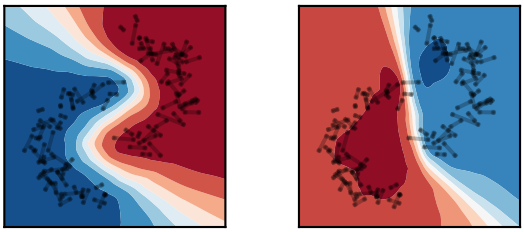

图*10*。使用神经网络的 *VCReg*。最小化器 ψ 的等高线图：在 $\beta = 1$ 时，$\mathcal{L}$ 关于 $\mathcal{X} \to \mathbb{R}$ 的等高线图，其中左图对应 $\beta = 1$，右图对应 $\beta = 0$；使用双层全连接神经网络，且 $k = 1$、$\mathcal{X} = \mathbb{R}^2$、$X$ 服从半月结构分布，$\xi = X + \varepsilon$ 对应较小噪声 ε。数据增强以黑点表示；若它们来自同一个输入 X，则用线连接。

E.2. 图5的实验细节

我们考虑在球面 $\mathbb{S}^{d-1}$ 上均匀分布的数据，其中 $d = 8$；增强由 $\{-1, 0, 1\}$的循环移位构成；点积核的形式为 $k(x, y) = (1 + x^{\top} y)\kappa(x^{\top} y)$，其中 $\kappa(u) = 1 - \arccos(u)/\pi$。
目标函数 f_ℓ^*定义如下：

$$f_1^*(x) = \frac{1}{3} \sum_{j=1}^{3} Q_{1,d}(x_j)$$

$$f_3^*(x) = \frac{1}{d} \sum_{j=1}^{d} Q_{3,d}(x_j),$$

其中 $Q_{\ell,\ d}$ 是附录 D.5中介绍的 Gegenbauer 多项式。注意，f_3^* 是次数为 3 的循环不变球谐函数，而 f_1^* 是次数为 1 的非不变球谐函数（尽管它具有一定的局部移位稳定性）。下游任务的标签由 f_ℓ^*生成，不含噪声。

图5 展示了下游相对超额风险 $\|\hat{f}_n - f_\ell^*\|_{L^2}^2 / \|f_\ell^*\|_{L^2}^2$，该值基于 1500 个测试数据点近似计算，并以预训练中使用的正则化参数 λ 为自变量。我们使用相同的 $n = 300$ 样本进行预训练和下游线性预测。预训练时，每个样本使用全部 3 种数据增强，表征维度为 $k = 20$。下游问题通过核岭回归求解，使用预训练得到的诱导核，并在测试样本上调整岭参数，以避免处理模型选择问题。

E.3. 图6 的实验细节

图 6 考察了一个涉及四个类别的分类问题，并专门构造了一个预训练任务，以设计出表示 ψ：$\mathcal{X} \to \mathbb{R}^k$，用于 $k = 4$，从而解决这一特定分类问题。我们考虑的数据集是半月数据集，其中 $X = Z + \mathbf{1}_{\langle Z,\ e_1 \rangle > 0} e_2 + U$、$Z \sim \mathcal{U}\left(\mathbb{S}^2\right)$和 $U \sim \mathcal{N}(0,\ \sigma^2 I)$，适用于 $\sigma = 0.1$。数据增强施加高斯噪声，$\xi = X + V$，适用于 $V \sim \mathcal{N}(0,\ \sigma^2 I)$，其中 $\sigma = 0.1$。这一设定对应于拉普拉斯算子满足以下条件的情形：

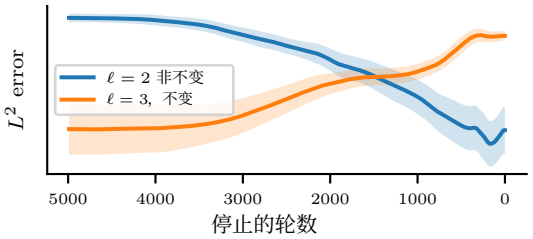

图 *11*。 图 *5* 使用神经网络时的行为。正则化参数 λ 由 SGD 的提前停止所取代。我们考虑一个具有两个隐藏层的神经网络，每个隐藏层均由 200 个神经元组成。优化采用具有固定步长的梯度下降进行。权重初始化带来的随机性在 100 次试验中取平均值，图中显示了标准差。

$\mathcal{L}(\psi) \simeq \|\nabla\psi\|^2_{L^2(\rho_X)}$。因此，理想的 ψ 将对应于拉普拉斯算子的顶层特征值。也就是说，前两个张成空间是两个新月上的常函数，接下来的两个是在给定月牙上具有一次振荡的波，依此类推。从本质上说，可以将 $L^2([0,1])$ 上的谐波视为 $x \to \cos(2i\pi\omega x + \chi)$，其中 $\chi \in \{0, \pi/2\}$ 和 $\omega \in \mathbb{N}$，将线段 $[0,1]$ 变形以匹配一个月牙，并在其他月牙上复制这一组基。在这种设定下，特征函数不是解析函数，因为解析函数无法在两个不同的流形上彼此分离（例如，局部为常数的解析函数必然处处为常数）。因此，使用径向基函数核搜索特征函数（在此情况下，Ψ 仅包含解析函数（Sun & Zhou, 2008））时，需要根据样本数量适当调整正则化参数。这解释了我们在本实验中选择指数核的原因：它对应于$\varphi(x)^\top\varphi(y) = \exp(-\|x-y\|/\sigma)$，并与一个更宽松的索伯列夫空间相关联，而该空间仍是再生核希尔伯特空间（在 $\mathbb{R}^2$ 中，这就是 H^1）。这有助于学习 T 的顶层特征函数，且无需改变 λ，从而更清晰地展示定理 1、2 和 3 的收敛速率。

在我们的实验中，我们固定了 $\lambda = 10^{-3}$ 并将指数核的尺度 σ 设为约等于问题直径的五分之一。我们在图 T 中绘制了通过 $n_{pre} = 2000$ 个样本经验得到的特征函数，见图 13。分类任务旨在学习图 12左侧所描述的四个类别。类别标签包含一定噪声，如条件概率的等高线所示，其中条件概率 Y 是 X 的函数，见图 12中部。该图右侧展示了一个训练集样本示例，其中 $n_{down} = 100$。在实验中，我们固定 $k = 5$，这确保了预训练任务与下游任务的性能之间具有较强的相关性。下游任务采用最小二乘替代目标进行优化：我们学习 $g:\ \mathcal{X} \to \mathbb{R}^4$ 使最小二乘误差达到最小的

$\mathbb{E}[\|g(X) - e_Y\|^2]$ 在解码为$f(X) = \arg\max_{i\in[4]} g_i(X)$ 以获得理想映射 $f^*:\ \mathcal{X} \to \mathcal{Y}$。我们在图14中报告了最小二乘（替代）损失和 0-1 损失上的下游泛化误差。该误差是预训练任务上 100 次试验和下游任务上 200 次试验的平均值。

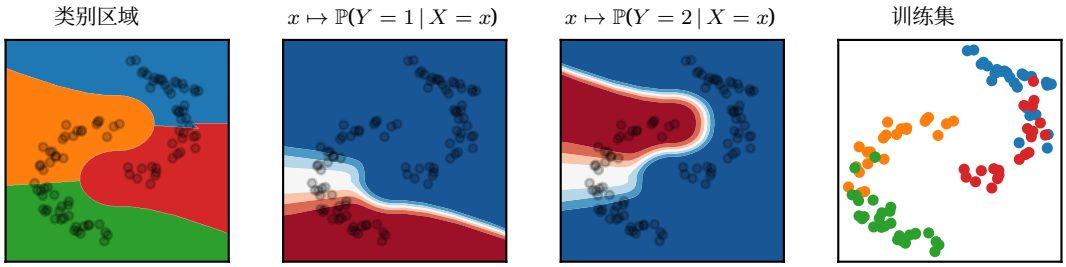

图 *12*。图*6*的设置。下游任务是在 $\mathcal{X} = \mathbb{R}^2$ 中学习四个类别，这些类别显示在左侧。它们由噪声生成。左侧月牙的条件分布 Y在给定 X 时的等高线显示在中间；右侧月牙遵循相同的结构。右侧展示了一个训练集示例。

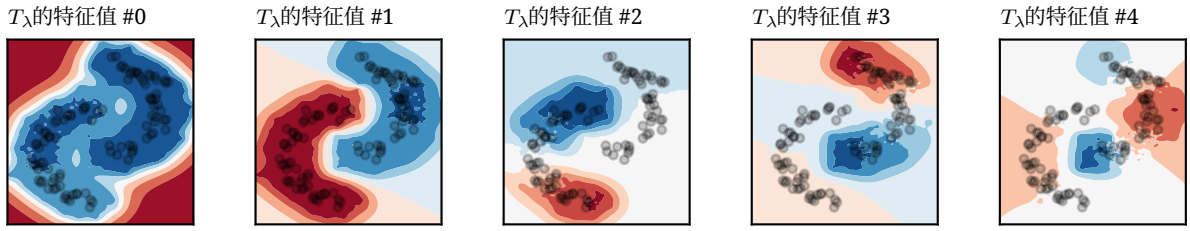

图 *13*。特征值：T_λ 在该问题上使用 2000 个预训练样本进行经验估计，所得经验速率如图 6所示。

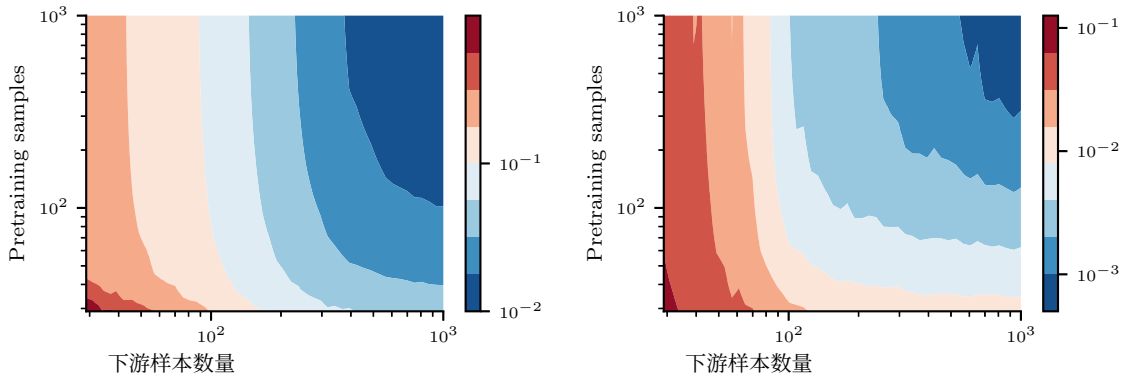

图 *14*。平均下游误差：在预训练任务上进行 100 次试验、在下游任务上进行 200 次试验后计算得出，分别采用最小二乘损失（右）和 0-1 损失（左）。