

理解神经网络的泛化能力： 从神经切线核到自适应特征

(申请清华大学理学博士学位论文)

培 养 单 位： 统计与数据科学系

学 科： 统计学

研 究 生： 李 易 诚

指 导 教 师： 林 乾 副教授

二〇二六年六月

**Understanding the Generalization of
Neural Networks:
From Neural Tangent Kernel Regression
to Adaptive Feature Models**

Dissertation submitted to
Tsinghua University
in partial fulfillment of the requirement
for the degree of
Doctor of Philosophy
in
Statistics

by
Li Yicheng

Dissertation Supervisor: Associate Professor Lin Qian

June, 2026

摘要

在过去的十年中，以神经网络为代表的深度学习技术在计算机视觉、生成式任务以及基础科学研究等领域取得了史无前例的成功，彻底重塑了人工智能的研究范式。然而，神经网络在高度过参数化环境下展现出的优异泛化性能，挑战了传统统计学习理论中关于“偏差-方差权衡”的认知，其背后的优化机理与泛化本质之间仍存在显著的理论鸿沟。

本文旨在从理论层面深入探讨神经网络的泛化机制，研究路径遵循从“静态核视角”向“动态特征演化”的递进。首先，针对神经切线核理论，本文对神经切线核的谱性质进行了精细分析，证明了多层全连接神经网络所对应的神经切线核在一般区域上的特征值衰减速率，这为理解核回归的收敛速度与泛化界奠定了基础。然后，通过建立一大类核回归方法与其所对应序列模型中的收缩估计量之间的等价性，本文完整刻画了核方法的泛化表现，既精确描述了其泛化误差曲线，也系统地揭示了核方法在捕获数据特有结构方面的局限性。

随后，为克服神经切线核框架下“惰性训练”无法解释神经网络特征学习这一缺陷，本文提出了一个通用的自适应特征模型框架。该框架引入了参数化、可训练的特征映射，用以模拟真实网络中的特征学习过程。为了量化特征学习的质量，本文创新性地提出了一种特征误差度量，通过衡量所学特征与目标函数结构之间的对齐程度，建立了特征自适应性与泛化性能之间的数学联系。

在此框架下，本文深入分析了两种具有代表性的自适应模型：对角参数化模型与方向参数化模型。研究表明，前者能够根据目标函数的稀疏性结构动态调整特征权重，而后者则通过学习低维子空间来捕捉数据的内在几何结构。两者在训练过程中对特征映射的改进都可以通过特征误差度量得到精确定量。这说明了由梯度下降驱动的隐式正则化效应能使模型自适应地对齐目标函数，从而显著提升泛化性能。本文的研究工作不仅深化了对神经网络从线性机制向非线性特征学习过渡的理解，也为设计更高效、具备可解释性的深度学习架构提供了理论指导。

关键词：神经网络；神经切线核；核方法；自适应特征模型；特征学习

Abstract

Over the past decade, deep learning technologies, represented by neural networks, have achieved unprecedented success across various fields, including computer vision, generative tasks, and fundamental scientific research, fundamentally reshaping the research paradigm of artificial intelligence. However, the superior generalization performance exhibited by neural networks in highly over-parameterized regimes challenges the classical understanding of the "bias-variance tradeoff" in statistical learning theory. A significant theoretical gap remains between this empirical success and our understanding of the underlying optimization mechanisms and the essence of generalization.

This dissertation aims to explore the generalization mechanisms of neural networks from a theoretical perspective, following a research trajectory that transitions from a "static kernel perspective" toward "dynamic feature evolution." First, regarding Neural Tangent Kernel (NTK) theory, this work provides a refined analysis of the spectral properties of the NTK. Specifically, we prove the eigenvalue decay rates of the NTK corresponding to multi-layer fully connected neural networks over general domains, laying the foundation for understanding the convergence rates and generalization bounds of kernel regression. Subsequently, by establishing the equivalence between a broad class of kernel regression methods and the corresponding shrinkage estimators in sequence models, this dissertation completely characterizes the generalization performance of kernel methods. This approach not only accurately describes the generalization error curves but also systematically reveals the limitations of kernel methods in capturing data-specific structures.

To overcome the limitation of the "lazy training" regime within the NTK framework in explaining feature learning in neural networks, this dissertation proposes a general Adaptive Feature Model framework. This framework introduces parameterized, trainable feature maps to simulate the feature learning dynamics in practical networks. To quantify the quality of feature learning, we introduce a Feature Error Measure (FEM). By measuring the degree of alignment between learned features and the structure of the target function, we establish a mathematical connection between feature adaptivity and generalization performance.

Under this framework, we conduct an in-depth analysis of two representative adaptive models: the diagonally parameterized model and the directionally parameterized

model. Our research demonstrates that the former can dynamically adjust feature weights based on the sparsity structure of the target function, while the latter captures the intrinsic geometric structure of the data by learning low-dimensional subspaces. We show that the improvements in feature maps during training for both models can be precisely quantified through the FEM. These results illustrate that the implicit regularization effects driven by gradient descent enable the models to adaptively align with the target function, thereby significantly enhancing generalization performance. The work presented in this dissertation not only deepens the understanding of the transition from linear mechanisms to non-linear feature learning in neural networks but also provides theoretical guidance for designing more efficient and interpretable deep learning architectures.

Keywords: Neural Networks; Neural Tangent Kernel; Kernel Methods; Adaptive Feature Model; Feature Learning

目 录

摘 要.....	I
Abstract.....	II
目 录.....	IV
插图和附表清单.....	VIII
第 1 章 引言	1
1.1 神经网络的理论视角	1
1.2 相关工作综述	3
1.2.1 神经切线核理论.....	3
1.2.2 核回归理论.....	3
1.2.3 特征学习理论.....	4
1.3 本文的主要工作	5
1.4 基础设定与记号约定	6
第 2 章 神经切线核的特征值衰减率	9
2.1 神经切线核理论	9
2.1.1 神经切线核的简要推导.....	11
2.1.2 ReLU 神经切线核的显式公式.....	13
2.1.3 一致收敛性.....	13
2.2 神经切线核的特征值	14
2.2.1 定理 2.2 的证明.....	16
2.2.2 核函数的变换与特征值衰减速率.....	17
2.3 限制在子区域上的球面内积核的特征值衰减率	20
2.3.1 球谐函数.....	22
2.3.2 球面上的内积核.....	23
2.3.3 Cesàro 求和.....	25
2.3.4 主要引理.....	27
2.3.5 定理 2.3 的证明.....	30
2.3.6 关于假设 2.1 的讨论.....	32
2.3.7 球面子区域的分解.....	33
2.3.8 紧自伴算子特征值的估计.....	36

第 3 章 核回归的泛化误差分析	37
3.1 核方法简介	37
3.1.1 正定核与再生核希尔伯特空间	37
3.1.2 积分算子	38
3.1.3 插值空间	41
3.1.4 正则再生核希尔伯特空间	41
3.1.5 谱算法	43
3.2 谱算法的泛化误差与序列模型的等价性	47
3.2.1 序列模型中的收缩估计量	47
3.2.2 额外假设	48
3.2.3 主定理	49
3.3 结果的讨论与应用	50
3.3.1 极小极大速率	50
3.3.2 精确的泛化误差曲线	51
3.3.3 插值情形	51
3.3.4 饱和效应	52
3.3.5 与 Le Cam 等价性的比较	53
3.3.6 解析算子演算论证	53
3.4 主定理的证明	54
3.4.1 证明思路	54
3.4.2 基本量控制	56
3.4.3 集中性质	58
3.4.4 解析算子演算	60
3.4.5 方差项	63
3.4.6 偏差项	67
3.5 辅助结论	71
3.5.1 滤谱函数的性质	73
3.5.2 正则 RKHS	74
3.5.3 解析滤谱函数	76
3.5.4 回归函数的源条件	77
3.5.5 引理 3.8 的证明	78
3.5.6 集中不等式	81

第 4 章 神经网络的自适应特征模型	83
4.1 自适应特征模型	83
4.2 特征误差度量	84
4.3 序列模型简化	86
4.3.1 核方法下的例证	87
4.3.2 数值验证	90
第 5 章 对角参数化模型的自适应特征理论	92
5.1 固定核回归的局限性	92
5.1.1 不同衰减速率的特征值	93
5.1.2 低维结构	95
5.1.3 错位结构	95
5.2 对角参数化模型	96
5.2.1 更深层的过参数化	97
5.3 泛化能力的结果	97
5.3.1 讨论	99
5.4 特征映射的改进	101
5.4.1 序列损失下的结果	101
5.4.2 样本损失下的结果	102
5.5 数值实验	104
5.5.1 低维结构的设定	104
5.5.2 错位结构的设定	104
5.6 单层参数化的泛化能力证明	107
5.6.1 梯度流的显式形式	107
5.6.2 分量划分	109
5.6.3 集中不等式	110
5.6.4 定理 5.1 的证明	111
5.6.5 一维动力学	117
5.7 多层参数化的泛化能力证明	123
5.7.1 多层模型的方程分析	123
5.7.2 定理 5.2 的证明	130
5.8 特征改善的证明	135
5.8.1 特征误差度量的基本性质	135
5.8.2 一维动力学补充结果	137

5.8.3 序列损失下的单层参数化.....	138
5.8.4 序列损失下的更深层过参数化.....	141
5.8.5 样本损失的情况, 定理 5.5 的证明.....	145
5.9 补充内容.....	147
5.9.1 集中不等式.....	147
5.9.2 第 5.1 节中例子的分析.....	148
5.9.3 杂项结果.....	150
第 6 章 方向参数化模型的自适应特征理论	152
6.1 Hermite 多项式与高斯分布	152
6.2 单指标模型	154
6.3 多指标模型	157
6.4 单指标模型的证明	160
6.4.1 特征误差度量的基本性质	160
6.4.2 总体动力学.....	162
6.4.3 序列模型.....	165
6.5 多指标模型的证明	171
6.5.1 预备知识.....	172
6.5.2 特征误差度量.....	181
6.5.3 总体动力学.....	183
6.5.4 序列模型的分析.....	190
6.6 辅助结果.....	203
6.6.1 随机过程的控制.....	203
6.6.2 高斯分布.....	204
6.6.3 Hermite 多项式	206
6.6.4 序列模型.....	209
6.6.5 级数.....	209
6.6.6 一些初等函数.....	210
第 7 章 总结与展望	212
参考文献.....	214

插图和附表清单

图 3.1	滤谱函数的示意图	45
图 3.2	围道 Γ_λ 的示意图.....	62
图 4.1	训练过程中特征误差度量的下降	90
图 4.2	经验损失和序列损失下训练曲线的相似性	91
图 4.3	经验损失和序列损失下特征误差度量之间的能量距离	91
图 5.1	固定核方法与对角自适应核方法在具有低维结构的示例上的对比	105
图 5.2	固定核方法与对角自适应核方法在各种设定下的对比	106
图 5.3	泛化误差以及特征值项 $a_j(t)b_j^D(t)$ 随时间 t 的演化	108
表 5.1	不同回归函数下对角自适应特征模型的泛化误差收敛速率	105
表 5.2	不同深度下对角自适应特征模型的泛化误差的收敛速率	106

第1章 引言

过去十余年中，以神经网络为代表的深度学习技术迅速发展，深刻改变了人工智能的研究范式。这场变革通常被追溯到2012年的ImageNet图像识别竞赛：以AlexNet^[1]为代表的深度卷积神经网络（CNN）显著降低了分类错误率，展示了深层架构在自动提取层级特征方面的强大能力，推动计算机视觉从人工特征工程转向端到端学习。随后，残差网络（ResNet）^[2]等结构进一步缓解了深层网络的优化困难，使神经网络在目标检测、人脸识别和自动驾驶视觉感知等任务中取得了突破性进展。

在感知任务之外，神经网络在生成式建模与语言建模中同样展现出强大能力。从变分自编码器（VAE）^[3]和生成对抗网络（GAN）^[4]，到近年来快速发展的扩散模型（Diffusion Models）^[5]，神经网络逐步显示出对高维复杂概率分布的刻画能力。与此同时，以注意力机制和Transformer架构^[6]为基础的大语言模型（LLMs）进一步将深度学习推向新的高度。通过在超大规模语料库上进行自监督预训练，以GPT系列^[7]为代表的模型表现出复杂推理、上下文学习和跨领域知识整合等能力，显著拓展了神经网络的应用边界。

神经网络的影响也早已超出传统互联网应用，深入到基础科学研究之中。在生命科学中，AlphaFold^[8]显著推动了蛋白质结构预测；在化学与材料科学中，神经网络被用于加速量子化学计算和新材料设计；在物理学中，算子神经网络等方法为求解复杂偏微分方程提供了新的工具。这些跨领域的成功应用表明，神经网络在处理高维复杂系统时具有独特优势。然而，与这种经验上的成功相比，其理论基础仍显不足。如何从理论上解释神经网络的优化行为、特征学习机制以及泛化能力，仍是当前机器学习理论中的核心问题之一。

1.1 神经网络的理论视角

尽管神经网络在实践中非常成功，但其理论基础在很长一段时间内却相对薄弱。在传统的统计学习理论框架下^[9]，模型复杂度与泛化误差通常遵循“偏差-方差权衡”。但是，神经网络的参数量往往十分巨大。根据经典的Vapnik-Chervonenkis维（VC维）理论^[10]，当模型的参数量远大于样本量时，模型极易陷入严重的过拟合，从而导致泛化性能的急剧下降。然而，现代神经网络却表现出一种令人费解的现象：即使在参数量多到足以完全拟合训练样本及其中噪声的情况下，模型在未知数据上的泛化表现依然能够持续提升^[11]。这一现象挑战了统计学的核心认

知，产生了一系列深刻的科学疑问：

- **优化之谜：**在高度非凸且充满局部极小值的损失曲面上，梯度下降方法是否能收敛到全局极小值附近？
- **泛化之谜：**在无数个能够完美拟合训练数据的解中，为什么神经网络总能“隐式地”选择那些具有优良泛化性能的解？

为了破解神经网络的奥秘，学术界在过去几年中形成了两个主要的研究范式：神经切线核理论与特征学习理论。

神经切线核理论：固定核的视角 神经切线核（Neural Tangent Kernel, NTK）^① 理论^[12]是理解过参数化神经网络训练动力学的奠基性工具。该理论指出，在网络宽度趋于无穷大的极限情况下，神经网络的训练过程可以被一个核回归估计量很好地近似，称之为神经切线核回归。它的核心见解是，在无限宽极限下，网络的特征映射在训练过程中保持静态（即所谓的“惰性训练”机制），其行为类似于具有固定表示的再生核希尔伯特空间（RKHS）中的核方法^[13-14]。由于神经网络的训练过程可以由对应的核梯度下降方法所近似，神经网络的梯度下降优化过程与泛化性能可以通过分析相应的核方法来理解^[15-16]：核方法的训练损失是凸的，因此梯度下降能够有效地收敛到全局最优解，而核回归的泛化能力则可以通过成熟的统计学习理论进行分析。通过利用成熟的核回归理论，神经切线核框架能够解释过参数化神经网络在某些情况下的泛化能力。即便如此，虽然神经切线核框架利用成熟的核回归理论解释了网络的部分泛化能力，但它未能捕捉到真实神经网络在有限宽度下的动态特征演化。它本质上将神经网络简化为传统的统计方法，无法解释网络如何从数据中自适应地提取特征。

特征学习理论：动态演化的视角 与神经切线核的静态观点不同，特征学习（Feature Learning）被认为是神经网络成功的关键因素^[17-18]。它描述了神经网络在训练过程中，通过动态调整其内部表示来发现任务相关模式的过程。在真实神经网络中，特征映射会随着训练的进程而不断演化，从而使网络能够更好地适应数据的内在结构。这种动态适应性使得神经网络能够超越传统的非参数回归方法（如核回归或样条平滑^[13,19]），后者依赖于静态、预定义的特征映射，难以捕捉数据的复杂模式。特征学习理论试图揭示神经网络如何通过这种动态特征调整来实现卓越的泛化性能。然而，尽管特征学习被广泛认为是神经网络成功的关键，但对其理论理解仍然较为零散和初步。一些研究关注随机特征模型在单步梯度下降后的表现，

^① Neural Tangent Kernel 也常译为“神经正切核”。本文采用“神经切线核”的译法，以强调“tangent”在此处与一阶线性化和切空间相关，而非三角函数意义上的正切。

另一些则探讨了过参数化带来的隐式正则化效应。虽然学术界已经意识到特征学习的重要性,但如何定量地描述“学到了什么特征”以及“特征如何影响泛化”依然缺乏统一的数学语言。如何在一个统一的数学框架下量化特征学习的质量,并建立其与泛化性能之间的精确联系,依然是该领域的前沿问题。

1.2 相关工作综述

1.2.1 神经切线核理论

神经网络的神经切线核近似 神经切线核最早由 Jacot et al.^[12] 提出,也有许多后续工作进一步深化理论。例如, Du et al.^[20], Arora et al.^[21] 表明两层神经网络的训练轨迹逐点收敛到神经切线核回归估计量的训练轨迹; Du et al.^[22], Allen-Zhu et al.^[23], Lee et al.^[24] 进一步将结果扩展到了多层网络和 ResNet。然而,如果想要利用神经切线核回归估计量的泛化误差来逼近过参数化神经网络的泛化误差,神经网络对核回归估计量的逼近必须是一致的。遗憾的是,现有的研究过参数化神经网络泛化误差的两项工作^[25-26] 忽视了上述逐点收敛与一致收敛之间的微妙区别,因此其主张中可能存在一些漏洞。之后, Lai et al.^[27] 证明了两层宽 ReLU 神经网络一致收敛到相应神经切线核回归估计量,而 Li et al.^[28] 则将结果扩展到了多层网络。

神经切线核的谱性质 自从 Jacot et al.^[12] 引入神经切线核以来,神经切线核的谱性质一直受到理论界学者的特别关注。例如,注意到与全连接 ReLU 网络相关的神经切线核是球面上的内积核,一些研究利用球面调和函数理论^[29] 来研究神经切线核的特征分解^[30-33]。特别地, Bietti et al.^[30] 和 Bietti et al.^[33] 表明,如果输入在 $\mathbb{S}^d$ 上均匀分布,则与两层和多层神经网络相关的神经切线核的特征值衰减率为 $i^{-(d+1)/d}$ 。然而,他们的分析依赖于球面上的球面调和函数理论来推导特征值的显式表达式,这无法推广到一般的输入域和分布。最近,考虑到区间上的两层 ReLU 神经网络, Lai et al.^[27] 证明了相应神经切线核的特征值衰减率为 i^{-2} 。然而,他们的技术严重依赖于神经切线核在 $\mathbb{R}$ 上的显式表达式,很难扩展到定义在 $\mathbb{R}^d$ 上的神经切线核或与多层宽网络相关的神经切线核。

1.2.2 核回归理论

核方法的最优性 有大量工作研究了核岭回归以及谱算法的最优速率。经典工作^[14] 证明了当回归函数位于再生核希尔伯特空间中时,核岭回归的最小极大最优性;后续工作^[34-36] 进一步将结果扩展到误指定情况,即回归函数不位于再生核

希尔伯特空间中。Zhang et al.^[37] 和 Yao et al.^[38] 考虑了核梯度方法，并分别证明了一致性和快速收敛速率。一般谱算法首先由 Bauer et al.^[39] 引入，然后在后续工作^[40-41] 中得到了广泛研究，但未考虑核的特征值衰减（见假设 3.1），因此速率并非最优。在某些限制性条件下，Caponnetto^[42] 证明了谱算法的最小极大最优性。最近，一系列工作进一步将结果扩展到更一般的情况（例如 Blanchard et al.^[43], Lin et al.^[44]）。此外，近期工作^[36] 展示了即使回归函数无界时错置情况下的最优性。读者还可以参考 Zhang et al.^[36] 中的表 1，以了解相关结果的总结。然而，如上所述，这些结果仅关注上界，不足以提供精确的泛化误差曲线。

核岭回归的近期进展 特别关注核岭回归，一系列近期工作提供了其泛化的进一步结果。一些工作^[45-47] 研究了核插值方法，这是核岭回归正则化趋于零的极限情况，并证明了它无法泛化。使用限制性的高斯设计假设以及一些非严格论证，Bordelon et al.^[48], Cui et al.^[49] 推导了核岭回归的泛化误差曲线，而 Mallinar et al.^[50] 进一步研究了插值与非插值解之间的转变。

高维极限下的核回归 还有一系列工作研究了高维极限下核回归的泛化，此时输入空间的维度 d 随 n 发散，如 $d/n \rightarrow \gamma \in (0, \infty)$ 或者 $d \asymp n^\alpha$ 。例如，在高维设置中，Liang et al.^[51] 展示了核插值可以泛化；Ghorbani et al.^[52], Ghosh et al.^[53], Liu et al.^[54], Lu et al.^[55], Zhang et al.^[56], Lu et al.^[57] 研究了核岭回归和核梯度方法的泛化能力。此外，他们设置中的高维渐进设定实际上使问题更容易，因为在高维渐进下非线性的核函数可以被其前几项的基函数展开所近似。由此，他们可以应用成熟的随机矩阵理论，但这在本文的设置中并非如此。

1.2.3 特征学习理论

一步梯度下降的特征学习 以 Ba et al.^[58] 为代表的一系列工作大多将浅层神经网络视为一种随机特征模型，其中特征权重首先通过一步梯度下降进行训练，随后利用岭回归计算输出权重，从而揭示特征学习行为的泛化性质^[18,59-64]。这些研究表明，特征学习体现为特征矩阵的自适应调整，通常表现为其谱中出现与目标函数结构相一致的秩为 1 的“尖峰”。除了随机特征模型外，自适应核的观点也在各种特定的形式下得到了研究^[17,65]。

过参数化作为隐式正则化 理解过参数化在深度学习中的作用也引发了研究热潮。一种观点认为，通过基于梯度的算法训练的过参数化模型会表现出对简单解的某种隐式偏好，这些解包括线性模型^[66]、矩阵分解^[67-70]、线性网络^[71-72] 以及神

神经网络^[17,73]。此外，随机梯度下降等梯度下降的变体也被证明具有隐式正则化效应^[74-76]。虽然这些工作大多仅关注优化过程和最终解，但只有少数研究为带有早停的过参数化模型提供了泛化保证。与本研究最相关的工作包括线性回归^[77-79]和单指标模型^[80]中的过参数化研究。

1.3 本文的主要工作

本文的研究位于神经切线核理论与特征学习理论的交汇处。本文关注的核心问题是：在固定核视角下，神经切线核及其对应核方法能够解释何种泛化行为；当固定特征难以捕捉数据中的特定结构时，训练过程中自适应演化的特征又如何改善模型对目标函数的表达与泛化性能。围绕这一问题，本文先分析固定核方法的谱结构与泛化能力，再进一步研究可训练特征映射带来的自适应改进。

前半部分聚焦于神经切线核及其对应核方法的精细分析。首先，本文研究多层全连接神经网络所对应神经切线核的谱性质，证明在一般输入区域与输入分布下，其特征值衰减速率为 $(d+1)/d$ ，其中 d 为输入空间维度。该结果突破了现有文献主要局限于球面均匀分布的分析框架，为理解神经切线核回归的泛化能力提供了谱理论基础。

在此基础上，本文进一步研究一般核回归方法的泛化能力，特别是与神经切线核回归相对应的核梯度下降方法。本文利用解析算子演算与精细的集中不等式分析，给出核回归泛化误差的 $(1 + o_p(1))$ 精确刻画，并证明其与对应序列模型中收缩估计量的泛化误差等价，从而较为完整地揭示核方法在非参数回归中的优势与局限。这部分结果对应本文第3章，其作用不仅是深化对固定核方法的理解，也为第4章和第5章中自适应核与自适应特征模型的序列化分析提供理论前提。

后半部分转向神经网络中的特征学习问题。为此，本文提出一个通用的自适应特征模型框架，引入参数化、可训练的特征映射，以刻画训练过程中不断演化的表示结构。为了定量描述这种演化是否真正改善了模型对目标函数的表达，本文进一步提出特征误差度量（Feature Error Measure, FEM），用以衡量所学特征与目标函数之间的对齐程度，并将特征学习质量与泛化性能联系起来。

在这一框架下，本文重点分析两类具有代表性的模型：对角参数化自适应特征模型与方向参数化自适应特征模型。对于前者，本文证明可训练特征值能够自适应地调整各个特征分量的权重，从而对齐目标函数的稀疏结构，并达到近乎最优的泛化能力；相比之下，固定核方法即使经过最优调参也无法达到同样效果。对于后者，本文在高斯单指标与多指标模型中引入可学习子空间方向的特征映射参数化，并证明梯度下降能够在训练过程中同时学习方向结构与未知链接函数，从

而持续提升特征映射的表达能力。这两类模型都涉及非线性、非凸且通常没有显式解的训练动力学，其分析也体现了本文方法上的主要理论挑战。

通过对自适应特征模型实例的深入分析，本文展示了自适应特征学习如何通过调整特征映射来对齐目标函数的结构，从而显著提升泛化性能。建立这一理论框架不仅有助于理解现有模型的泛化机理，更能为设计更高效、可解释的结构提供数学指导。

本文的主要结构如下：第2章将简要介绍神经切线核理论，并深入研究神经切线核的谱性质，特别是其特征值的衰减行为；第3章将回顾核回归的经典设定，并给出解析谱算法的序列模型等价性，从而完全确定核回归的泛化能力；第4章将提出自适应特征模型的一般形式，并利用前面所得到的序列模型等价性，提出特征误差度量来量化所学特征的有效性，并引入序列模型损失以简化分析；随后，第5章引入对角参数化自适应特征模型，证明对角参数化的自适应特征可以对齐信号的坐标结构；而第6章引入方向参数化自适应特征模型，证明方向参数化的自适应特征可以对齐信号的低维子空间结构；最后，第7章将总结本文的主要贡献，并讨论未来的研究方向。

1.4 基础设定与记号约定

本文主要研究非参数回归问题。假设输入空间为 $\mathcal{X}$ ，输出空间为 $\mathcal{Y} \subseteq \mathbb{R}$ 。假设有一个定义在 $\mathcal{X} \times \mathcal{Y}$ 上的未知的联合分布 ρ 。记 μ 为 ρ 在 $\mathcal{X}$ 上的边缘分布。目标是从样本中学习一个函数 $f : \mathcal{X} \rightarrow \mathcal{Y}$ 来近似描述输入 x 与输出 y 之间的关系。为了衡量估计函数的性能，通常使用均方误差（Mean Squared Error, MSE）作为损失函数，即

$$\text{Risk}(f) = \mathbb{E}_{(x,y) \sim \rho} [(f(x) - y)^2]. \quad (1.1)$$

最小化该风险函数的最优解称为回归函数，记为 f^* ，它就是条件期望

$$f^*(x) = \mathbb{E}[y | x]. \quad (1.2)$$

由此，也可以将模型写为

$$y = f^*(x) + \epsilon, \quad (1.3)$$

其中 $\epsilon := y - f^*(x)$ 是噪声项，满足 $\mathbb{E}[\epsilon | x] = 0$ 。给定一个 f^* 的估计量，本文使用超额风险（Excess Risk）来衡量估计量的泛化能力，也称之为泛化误差，定义为：

$$\|\hat{f} - f^*\|_{L^2(\mathcal{X}, \mu)}^2 = \mathbb{E}_{x \sim \mu} [\hat{f}(x) - f^*(x)]^2 = \text{Risk}(\hat{f}) - \text{Risk}(f^*). \quad (1.4)$$

这里的期望仅对独立测试点 $x \sim \mu$ 计算；当 $\hat{f}$ 由训练样本产生并因此具有随机性时，上式表示条件于训练样本后的泛化误差，必要时还可以进一步对训练样本取期望。

假设从未知分布 ρ 中采样得到 n 个独立同分布的样本 $(x_i, y_i) \sim \rho, i = 1, 2, \dots, n$ 。引入均方误差 (1.1) 的样本版本，即经验损失（经验风险），定义如下：

$$\hat{\mathcal{L}}_n(f) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2. \quad (1.5)$$

在本文中，一个估计量往往是通过在经验损失下的某种训练过程得到的。

本文使用 $C, c, \dots$ 表示某个正常数，其值在证明的不同位置可能不同。一般地，本文使用 C 表示较大的正常数， c 表示较小的正常数。对于两个正数 a 和 b ，本文使用 $a \lesssim b$ 来表示存在一个正常数 C 使得 $a \leq Cb$ 。本文也使用 $a \gtrsim b$ 来表示 $b \lesssim a$ ，并使用 $a \asymp b$ 来表示同时有 $a \lesssim b$ 和 $a \gtrsim b$ 。此外，本文也使用标准的大 O 符号来表示渐近关系，即 $a = O(b)$ 表示存在正常数 C 使得 $a \lesssim b$ ， $a = o(b)$ 表示 $\lim_{n \rightarrow \infty} a/b = 0$ ， $a = \Omega(b)$ 表示 $b = O(a)$ ， $a = \Theta(b)$ 表示同时有 $a = O(b)$ 和 $a = \Omega(b)$ 。本文还会使用依概率的记号 $a_n = O_{\mathbb{P}}(b_n)$ ，表示对于任意 $\epsilon > 0$ ，存在正常数 C_ϵ 和 n_ϵ 使得当 $n \geq n_\epsilon$ 时，有 $\mathbb{P}(a_n \leq C_\epsilon b_n) \geq 1 - \epsilon$ ；以及 $a_n = o_{\mathbb{P}}(b_n)$ ，表示对于任意 $\epsilon > 0$ 和 $\delta > 0$ ，存在 $n_{\epsilon, \delta}$ 使得当 $n \geq n_{\epsilon, \delta}$ 时，有 $\mathbb{P}(a_n \leq \delta b_n) \geq 1 - \epsilon$ 。

对于一个集合 A ，本文使用 $|A|$ 或者 $\#A$ 表示其基数。记 $x^+ = \max(x, 0)$ 表示 x 的正部， $x^- = \max(-x, 0)$ 表示 x 的负部。也使用 $x \wedge y = \min(x, y)$ 和 $x \vee y = \max(x, y)$ 分别表示 x 和 y 的最小值与最大值。

对于一般的希尔伯特空间 H ，本文使用 $\|\cdot\|_H$ 表示其范数， $\langle \cdot, \cdot \rangle_H$ 表示其内积。在没有歧义的情况下，本文会省略下标 H 。如果 H 是复值希尔伯特空间，则本文约定内积 $\langle \cdot, \cdot \rangle_H$ 是线性于第二个变量、共轭线性于第一个变量的。对于一个集合 $\mathcal{X}$ 以及其上测度 μ ，本文使用 $L^p(\mathcal{X}, \mu)$ 表示其上的可积函数空间，配备范数 $\|\cdot\|_{L^p}$ 。平方可积函数空间 $L^2(\mathcal{X}, \mu)$ 的内积由 $\langle \cdot, \cdot \rangle_{L^2}$ 表示。通常本文默认它们为实值函数空间，但本文也会考虑复值函数空间，此时的内积由 $\langle f, g \rangle_{L^2} = \int_{\mathcal{X}} \overline{f(x)} g(x) d\mu(x)$ 定义，其中 $\overline{f(x)}$ 表示 $f(x)$ 的复共轭，这样定义与前面的约定一致。在没有歧义的情况下，本文会直接使用 L^p 或 L^2 。

对于希尔伯特空间 H_1 和 H_2 之间的有界线性算子 $A : H_1 \rightarrow H_2$ ，本文使用 $\|A\|_{\mathcal{B}(H_1, H_2)}$ 表示其算子范数。如果 $H_1 = H_2$ ，本文简记为 $\|A\|_{\mathcal{B}(H)}$ 。在没有歧义的情况下，本文也会直接使用 $\|A\|$ 来表示算子范数。本文使用 $\|A\|_2$ 表示算子的希尔伯特-施密特范数。当 $H_1 = H_2 = H$ 时，本文使用 $\|A\|_1$ 表示算子的迹类范数，以及 $\text{Tr}(A)$ 表示算子的迹。

本文使用 $\mathbf{a} \odot \mathbf{b}$ 表示两个向量 $\mathbf{a}, \mathbf{b}$ 的逐元素 (Hadamard) 乘积。令 $S \subseteq \mathbb{N}^+$ 为一个指标集。对于向量 $\mathbf{v}$, 本文用 $\mathbf{v}_S$ 表示下标在 S 中的子向量。令 A 为一个矩阵, $R \subseteq \mathbb{N}^+$ 为另一个指标集。本文对子矩阵使用类似的符号 A_{SR} 、 $A_{\cdot R}$ 和 A_{kR} 。此外, 本文用 $D_{\mathbf{v}}$ 表示以 $\mathbf{v}$ 为对角元素的对角矩阵。

第2章 神经切线核的特征值衰减率

本章以神经切线核 (Neural Tangent Kernel, NTK) 理论^[12]为出发点, 研究过参数化神经网络对应核函数的谱性质。神经切线核理论表明, 在宽度趋于无穷大的极限下, 神经网络的训练过程可以由一个核回归估计量来近似; 因此, 要理解这一近似模型的泛化能力, 首先需要掌握神经切线核本身的特征值结构。本章的主要结论定理 2.2 表明, 当 $d \geq 3$ 时, 对于一般的 d 维输入空间, 神经切线核的特征值以大约 $(d+1)/d$ 的幂律速率衰减。为建立这一结论, 本文进一步证明了定理 2.3, 说明球面内积核在球面子区域上的特征值衰减率与其在整个球面上的衰减率一致; 这一结果本身也具有独立意义。这部分分析不仅补充了神经切线核在一般输入区域和分布下的谱理论, 也为后续章节研究神经切线核回归的泛化误差奠定了基础^①。

2.1 神经切线核理论

作为神经切线核的具体例子, 本章主要研究具有 L 个隐藏层的全连接神经网络 $f(x; \theta)$ 。在这里, 本章特别考虑输入空间 $x = (\tilde{x}, 1) \in \mathcal{X} = \Omega \times \{1\} \subset \mathbb{R}^{d+1}$, 其中 $\Omega \subset \mathbb{R}^d$ 为实质输入空间, 额外的维度 1 用于表示偏置项 (bias term), 从而将偏置项包含在权重矩阵中以简化表示。设网络各层宽度分别为 $m_1, m_2, \dots, m_L$ ($L \geq 2$ 固定), 再额外定义输入层宽度 $m_0 = d+1$ 和输出层宽度 $m_{L+1} = 1$ 。本章所考虑的网络模型定义如下:

$$\begin{aligned} \mathbf{u}^{(1)}(x) &= \mathbf{W}^{(1)}x \in \mathbb{R}^{m_1}, \\ \mathbf{u}^{(l)}(x) &= \sqrt{\frac{2}{m_{l-1}}} \mathbf{W}^{(l)} \sigma(\mathbf{u}^{(l-1)}(x)) \in \mathbb{R}^{m_l}, \quad l \in \{2, \dots, L+1\}, \\ f(x; \theta) &= \mathbf{u}^{(L+1)}(x). \end{aligned} \quad (2.1)$$

此处, $\mathbf{u}^{(l)}(x)$ 表示第 l 层的预激活值 (pre-activation) 向量; $\sigma(x)$ 为逐元素施加的激活函数。网络参数集合 θ 包含了所有的权重矩阵 $\mathbf{W}^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$ 。

神经网络的参数初始化通常采用随机初始化策略。在此, 本章假设所有网络参数均独立同分布地从零均值、单位方差的高斯分布中随机抽取, 即:

$$\mathbf{W}_{i,j}^{(l)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1), \quad \text{for } l = 1, \dots, L+1.$$

① 本章内容主要源自笔者已发表的论文 Li et al.^[28]。

神经网络通常使用梯度下降法（或其变体）在训练数据集上进行训练。回顾经验损失函数为

$$\hat{\mathcal{L}}_n(\theta) = \frac{1}{2n} \sum_{i=1}^n (f(x_i; \theta) - y_i)^2,$$

其中额外的因子 $\frac{1}{2}$ 是为了简化后续的导数表达式。虽然梯度下降法通常以离散时间的形式进行，但为了简化分析，本章考虑其连续时间的极限形式，即梯度流（Gradient Flow）。具体地，梯度流定义为参数 θ_t 随时间 t 的演化满足以下微分方程：

$$\dot{\theta}_t = -\nabla_{\theta} \mathcal{L}(\theta_t) = -\frac{1}{n} \nabla_{\theta} f(x; \theta_t) (f(x; \theta_t) - y), \quad (2.2)$$

将训练得到的网络预测器记为 $\hat{f}_t^{\text{NN}}(x) := f(x; \theta_t)$ 。

显然，梯度流方程（2.2）是一个高度非线性的微分方程，难以求得显式解。但是，著名的神经切线核（NTK）理论^[12]指出，当网络宽度趋于无穷大时，网络训练过程可以被一个核回归估计量很好地近似。这种近似也被称为神经网络的一种惰性训练机制（Lazy Training Regime）。为了说明这一点，可以计算网络输出所满足的微分方程。利用梯度流方程（2.2），链式法则以及平方损失的形式，有：

$$\begin{aligned} \frac{d}{dt} \hat{f}_t^{\text{NN}}(x) &= \langle \nabla_{\theta} f(x; \theta_t), \dot{\theta}_t \rangle \\ &= -\frac{1}{n} \sum_{i=1}^n \langle \nabla_{\theta} f(x; \theta_t), \nabla_{\theta} f(x_i; \theta_t) \rangle (f(x_i; \theta_t) - y_i). \end{aligned}$$

由此，可以引入一个随时间变化的核函数，本章称之为网络切线核：

$$k_t(x, x') = \langle \nabla_{\theta} f(x; \theta_t), \nabla_{\theta} f(x'; \theta_t) \rangle. \quad (2.3)$$

由此，可以将网络输出的演化方程重写为：

$$\frac{d}{dt} \hat{f}_t^{\text{NN}}(x) = -\frac{1}{n} k_t(x, X) (\hat{f}_t^{\text{NN}}(X) - y), \quad (2.4)$$

其中 $k_t(x, X) = (k_t(x, x_1), k_t(x, x_2), \dots, k_t(x, x_n))$ 是一个 $1 \times n$ 的向量， $\hat{f}_t^{\text{NN}}(X) = (\hat{f}_t^{\text{NN}}(x_1), \hat{f}_t^{\text{NN}}(x_2), \dots, \hat{f}_t^{\text{NN}}(x_n))$ 为 $n \times 1$ 的网络在训练集上的预测值向量。

神经切线核^[12]的核心发现是：在大宽度极限下，网络切线核 k_0 收敛于一个确定的核函数 k_{NT} ，即神经切线核（Neural Tangent Kernel, NTK）。由此，可以考虑将（2.4）中的 k_t 替换为 k_{NT} ，从而得到如下的神经切线核回归估计量的演化方程：

$$\frac{d}{dt} \hat{f}_t^{\text{NTK}}(x) = -\frac{1}{n} k_{\text{NT}}(x, X) (\hat{f}_t^{\text{NTK}}(X) - y). \quad (2.5)$$

需要强调的是，（2.4）中的 k_t 是沿训练轨迹变化的网络切线核，而（2.5）使用的是初始化时刻的极限核 k_{NT} 。两者能够相互近似依赖于宽网络的惰性训练机制：在适当

宽度和训练时间尺度下，参数相对初始化点的移动较小，从而 k_t 在训练过程中保持接近 k_0 ，而 k_0 又在宽度趋于无穷时收敛到确定的 k_{NT} 。神经切线核理论的进一步结论是，在大宽度极限下，核函数的收敛不仅发生在初始时刻 $t = 0$ ，而且在整个训练过程中均成立，并且神经网络的输出轨迹 $\hat{f}_t^{\text{NN}}(x)$ 可以被神经切线核回归估计量 $\hat{f}_t^{\text{NTK}}(x)$ 很好地近似。直观来看，比较 (2.4) 和 (2.5)，利用常微分方程的稳定性理论，确实可以预期这个近似成立。

2.1.1 神经切线核的简要推导

为了给神经切线核理论提供一些直观的理解以及文章内容的自洽性，在此只关注初始化时刻，不严格地简要推导神经切线核的收敛性和递归表达式。考虑宽度 $m = \min(m_1, \dots, m_L) \rightarrow \infty$ 。记 $\alpha^{(l)}(x) = \sigma(u^{(l)}(x)) \in \mathbb{R}^{m_l}$ 为第 l 层的激活值向量 (activation)，其元素为 $\alpha_i^{(l)}(x)$ 。根据网络定义，有：

$$u_i^{(l)}(x) = \sqrt{\frac{2}{m_{l-1}}} \sum_{j=1}^{m_{l-1}} W_{ij}^{(l)} \alpha_j^{(l-1)}(x). \quad (2.6)$$

为了推导神经切线核的递归形式，下面将分为两部分：前向传播和反向传播。

前向传播 首先考察前向传播过程，对于层数递归地计算预激活值的分布。固定 $\alpha^{(l-1)}(x)$ ，由于权重 $W_{ij}^{(l)} \sim \mathcal{N}(0, 1)$ 且独立同分布，根据中心极限定理 (CLT)，当 $m_{l-1} \rightarrow \infty$ 时， $u_i^{(l)}(x)$ 服从正态分布，而作为函数 $x \mapsto u_i^{(l)}(x)$ ，它们构成一个高斯过程 (Gaussian Process, GP)：

$$u_i^{(l)}(x) \sim \mathcal{GP}(0, k_{\text{RF},l}(x, x')), \quad k_{\text{RF},l}(x, x') = \lim_{m \rightarrow \infty} \mathbb{E} [u_i^{(l)}(x), u_i^{(l)}(x')],$$

其中均值为零是容易得到的。注意它与神经元指标 i 无关，因为不同神经元的权重是独立同分布的。现在来计算这个核函数 $k_{\text{RF},l}(x, x')$ 的递归形式。首先，容易看出 $k_{\text{RF},0}(x, x') = \langle x, x' \rangle$ 。对于 $l \geq 1$ ，先计算有限宽度下的经验协方差：

$$\begin{aligned} \mathbb{E} u_i^{(l)}(x) u_i^{(l)}(x') &= \frac{2}{m_{l-1}} \mathbb{E} \left[\left(\sum_{j=1}^{m_{l-1}} W_{ij}^{(l)} \alpha_j^{(l-1)}(x) \right) \left(\sum_{k=1}^{m_{l-1}} W_{ik}^{(l)} \alpha_k^{(l-1)}(x') \right) \right] \\ &= \frac{2}{m_{l-1}} \sum_{j=1}^{m_{l-1}} \alpha_j^{(l-1)}(x) \alpha_j^{(l-1)}(x'), \end{aligned}$$

其中利用了权重的独立性消去了交叉项。当 $m_{l-1} \rightarrow \infty$ 时，利用大数定律， $\alpha_j^{(l-1)}(x) = \sigma(u_j^{(l-1)}(x))$ 以及前一层的高斯过程结论，得到：

$$k_{\text{RF},l}(x, x') = 2 \mathbb{E} \sigma(u^{(l-1)}(x)) \sigma(u^{(l-1)}(x')) = 2 \mathbb{E}_{(w,z)} [\sigma(w) \sigma(z)], \quad (2.7)$$

其中 (w, z) 是服从均值为零，协方差矩阵为

$$\begin{pmatrix} k_{\text{RF},l-1}(x, x) & k_{\text{RF},l-1}(x, x') \\ k_{\text{RF},l-1}(x, x') & k_{\text{RF},l-1}(x', x') \end{pmatrix}$$

的二维高斯随机变量。这一核函数也被称为随机特征核 (Random Feature Kernel, RFK)，因为它对应于神经网络的随机特征模型。

反向传播 现在需要计算神经切线核 $k_{\text{NT}}(x, x')$ 的递归形式。根据定义，神经切线核可以写成所有网络参数梯度内积之和。再引入截止到第 l 层的神经切线核：

$$k_{\text{NT},l}(x, x') = \lim_{m \rightarrow \infty} \left\langle \nabla_{\theta_{\leq l}} \mathbf{u}_i^{(l)}(x), \nabla_{\theta_{\leq l}} \mathbf{u}_i^{(l)}(x') \right\rangle,$$

其中 $\theta_{\leq l}$ 表示截至第 l 层的所有参数。

利用网络结构，可以将神经切线核分解为最后一层参数项和前序层传播项之和：

$$k_{\text{NT},l}(x, x') = \lim_{m \rightarrow \infty} \left\langle \nabla_{\mathbf{W}^{(l-1)}} \mathbf{u}_i^{(l)}(x), \nabla_{\mathbf{W}^{(l-1)}} \mathbf{u}_i^{(l)}(x') \right\rangle + \lim_{m \rightarrow \infty} \left\langle \nabla_{\theta_{<l-1}} \mathbf{u}_i^{(l)}(x), \nabla_{\theta_{<l-1}} \mathbf{u}_i^{(l)}(x') \right\rangle.$$

对于第一项，有

$$\left\langle \nabla_{\mathbf{W}^{(l-1)}} \mathbf{u}_i^{(l)}(x), \nabla_{\mathbf{W}^{(l-1)}} \mathbf{u}_i^{(l)}(x') \right\rangle = \frac{2}{m_{l-1}} \sum_{j=1}^{m_{l-1}} \sigma(\mathbf{u}_j^{(l-1)}(x)) \sigma(\mathbf{u}_j^{(l-1)}(x')),$$

所以当 $m_{l-1} \rightarrow \infty$ 时，利用大数定律和前向传播的高斯过程结论，得到：

$$\left\langle \nabla_{\mathbf{W}^{(l-1)}} \mathbf{u}_i^{(l)}(x), \nabla_{\mathbf{W}^{(l-1)}} \mathbf{u}_i^{(l)}(x') \right\rangle \rightarrow 2\mathbb{E}_{(w,z)}[\sigma(w)\sigma(z)] = k_{\text{RF},l}(x, x').$$

对于第二项，利用链式法则，

$$\begin{aligned} \nabla_{\theta_{<l-1}} \mathbf{u}_i^{(l)}(x) &= \sum_{j=1}^{m_{l-1}} \frac{\partial \mathbf{u}_i^{(l)}(x)}{\partial \mathbf{u}_j^{(l-1)}(x)} \nabla_{\theta_{<l-1}} \mathbf{u}_j^{(l-1)}(x) \\ &= \sqrt{\frac{2}{m_{l-1}}} \sum_{j=1}^{m_{l-1}} W_{ij}^{(l)} \dot{\sigma}(\mathbf{u}_j^{(l-1)}(x)) \nabla_{\theta_{<l-1}} \mathbf{u}_j^{(l-1)}(x), \end{aligned}$$

其中 $\dot{\sigma}(x)$ 为激活函数 $\sigma(x)$ 的导数。将其代入第二项，得到：

$$\begin{aligned} \left\langle \nabla_{\theta_{<l-1}} \mathbf{u}_i^{(l)}(x), \nabla_{\theta_{<l-1}} \mathbf{u}_i^{(l)}(x') \right\rangle &= \frac{2}{m_{l-1}} \sum_{j,k=1}^{m_{l-1}} W_{ij}^{(l)} W_{ik}^{(l)} \dot{\sigma}(\mathbf{u}_j^{(l-1)}(x)) \dot{\sigma}(\mathbf{u}_k^{(l-1)}(x')) \\ &\quad \cdot \left\langle \nabla_{\theta_{<l-1}} \mathbf{u}_j^{(l-1)}(x), \nabla_{\theta_{<l-1}} \mathbf{u}_k^{(l-1)}(x') \right\rangle. \end{aligned}$$

其中的后一项收敛到前一层的神经切线核 $k_{\text{NT},l-1}(x, x')$ 。而对于前一项，利用权重

的独立性，有

$$\frac{2}{m_{l-1}} \sum_{j,k=1}^{m_{l-1}} W_{ij}^{(l)} W_{ik}^{(l)} \dot{\sigma}(\mathbf{u}_j^{(l-1)}(x)) \dot{\sigma}(\mathbf{u}_k^{(l-1)}(x')) \rightarrow 2\mathbb{E} \dot{\sigma}(\mathbf{u}_j^{(l-1)}(x)) \dot{\sigma}(\mathbf{u}_k^{(l-1)}(x')) = 2\mathbb{E}_{(w,z)} [\dot{\sigma}(w) \dot{\sigma}(z)],$$

其中 (w, z) 服从与前向传播相同的二维高斯分布。综上所述，得到了神经切线核的递归形式：

$$k_{\text{NT},l}(x, x') = k_{\text{RF},l}(x, x') + k_{\text{NT},l-1}(x, x') \cdot 2\mathbb{E}_{(w,z)} [\dot{\sigma}(w) \dot{\sigma}(z)]. \quad (2.8)$$

2.1.2 ReLU 神经切线核的显式公式

利用表达式 (2.7) 和 (2.8)，可以递归地计算神经切线核的值^[12]。特别地，对于 ReLU 激活函数 $\sigma(x) = \max(0, x)$ ，借助 Cho et al.^[81] 所引入的反余弦核（arccosine kernel）技巧，可以计算出期望项的解析解，得到更为简洁的显式公式。定义

$$\kappa_0(u) = \frac{1}{\pi} (\pi - \theta) = \frac{1}{\pi} (\pi - \arccos u), \quad (2.9)$$

$$\kappa_1(u) = \frac{1}{\pi} [\sin \theta + (\pi - \theta) \cos \theta] = \frac{1}{\pi} \left[\sqrt{1 - u^2} + u(\pi - \arccos u) \right], \quad (2.10)$$

其中 $\theta = \arccos u$ 。那么，对于 $(w, z) \sim \mathcal{N}(0, B)$ ， $B = \begin{pmatrix} a^2 & uab \\ uab & b^2 \end{pmatrix}$ ，有

$$2\mathbb{E}_{(w,z) \sim \mathcal{N}(0,B)} [\dot{\sigma}(w) \dot{\sigma}(z)] = \kappa_0(u), \quad 2\mathbb{E}_{(w,z) \sim \mathcal{N}(0,B)} [\sigma(w) \sigma(z)] = ab\kappa_1(u).$$

由此，通过进一步化简计算，可以得到

$$k_{\text{RF},l}(x, x') = \|x\| \|x'\| \kappa_1^{(l)}(\bar{u}), \quad \bar{u} = \frac{\langle x, x' \rangle}{\|x\| \|x'\|}, \quad (2.11)$$

$$k_{\text{NT},l}(x, x') = \|x\| \|x'\| \sum_{r=0}^l \kappa_1^{(l-r)}(\bar{u}) \prod_{s=1}^r \kappa_0(\kappa_1^{(l-s)}(\bar{u})), \quad (2.12)$$

其中 $\kappa^{(l)}$ 表示对函数 κ 的 l 次复合。从表达式 (2.12) 可以看出，如果输入数据处于单位球面上（即 $\|x\| = 1$ ），那么神经切线核仅依赖于输入数据之间的内积 $\bar{u} = \langle x, x' \rangle$ 。这种只依赖于内积的核函数称为内积核（dot-product kernel），属于径向基核（radial basis kernel）的一种特殊情况。

2.1.3 一致收敛性

尽管已有工作^[15,24]表明神经网络的输出函数 $\hat{f}_t^{\text{NN}}(x)$ 可以被 $\hat{f}_t^{\text{NTK}}(x)$ 逼近，但大多数现有结果建立的是点态（pointwise）收敛。也就是说，这些结果仅表明对于固定的 x ， $\sup_{t \geq 0} |\hat{f}_t^{\text{NTK}}(x) - \hat{f}_t^{\text{NN}}(x)|$ 在高概率下是很小的。

然而，为了保证 $\hat{f}_t^{\text{NN}}$ 的泛化误差能够被 $\hat{f}_t^{\text{NTK}}$ 的泛化误差准确刻画，需要 $\hat{f}_t^{\text{NN}}$

向 $\hat{f}_t^{\text{NTK}}$ 一致收敛 (uniform convergence)。假设 $\hat{f}_t^{\text{NTK}}$ 初始化与 $\hat{f}_t^{\text{NN}}$ 相同。例如额外考虑一个镜像神经网络结构, 可以保证 $\hat{f}_0^{\text{NN}} \equiv 0$, 此时设置 $\hat{f}_0^{\text{NTK}} \equiv 0$ 即可; 或者一般地, $\hat{f}_0^{\text{NN}}$ 服从一个高斯过程分布, 此时可以将 $\hat{f}_0^{\text{NTK}}$ 设置为 $\hat{f}_0^{\text{NN}}$ 的同一实现。Li et al.^[28] 证明了如下的一致收敛性结果:

定理 2.1 (一致收敛性): 假设 $\mathcal{X}$ 有界且经验核矩阵 $(k_{\text{NT}}(x_i, x_j))_{n \times n}$ 的最小特征值 $\lambda_0 > 0$ 。存在一个多项式 $\text{poly}(\cdot)$, 使得对于任意 $\epsilon > 0$ 和 $\delta \in (0, 1)$, 当网络宽度 m 满足 $m \geq \text{poly}(n, \lambda_0^{-1}, \|\mathbf{y}\|, \ln(1/\delta), 1/\epsilon)$ 时, 在随机初始化下以至少 $1 - \delta$ 的概率成立:

$$\sup_{t \geq 0} \sup_{x \in \mathcal{X}} \left| \hat{f}_t^{\text{NTK}}(x) - \hat{f}_t^{\text{NN}}(x) \right| \leq \epsilon.$$

上述定理中的 λ_0 是给定训练样本对应的经验核矩阵的最小特征值, 一般可以依赖于样本和样本量 n 。因此, 宽度下界对 λ_0^{-1} 的多项式依赖应理解为该一致近似结果的一个条件; 后文关于积分算子特征值衰减率的结果并不直接推出经验核矩阵的 λ_0 存在与 n 无关的下界。

作为该定理的直接推论, 可以证明 $\hat{f}^{\text{NN}}$ 的泛化性能可以被 $\hat{f}^{\text{NTK}}$ 很好地近似。于是, 可以将神经网络的泛化性能分析归结为对神经切线核回归的泛化性能分析, 这是本文将在后续的第 3 章中进行的工作。

2.2 神经切线核的特征值

在核方法的框架下, 核函数的谱性质, 特别是特征值衰减率, 对于泛化能力的分析至关重要。具体地, 考虑定义在输入空间 $\mathcal{X}$ 上的核函数 $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 以及 $\mathcal{X}$ 上的一个测度 μ (不一定需要是概率测度), 可以引入一个与核函数 k 相关联的积分算子 $T : L^2(\mathcal{X}, d\mu) \rightarrow L^2(\mathcal{X}, d\mu)$, 其定义为:

$$(Tf)(x) = \int_{\mathcal{X}} k(x, x') f(x') d\mu(x'), \quad f \in L^2(\mathcal{X}, d\mu). \quad (2.13)$$

在一定的正则性条件下 (例如 μ 为有限测度, k 也有界), 这个积分算子是自伴紧正算子, 因此 T 的谱分解给出了一组 $L^2(\mathcal{X}, d\mu)$ 中标准正交的特征函数 $\{e_i\}_{i \in N}$ 以及对应的非负特征值序列 $\{\lambda_i\}_{i \in N}$ (其中 $N \subseteq \mathbb{N}$ 为一个指标集), 满足

$$Tf = \sum_{i=1}^{\infty} \lambda_i \langle f, e_i \rangle_{L^2} e_i.$$

不仅如此, 特征值和特征函数还给出了核函数本身的级数表示, 这就是著名的 Mercer 分解^[82-83]:

$$k(x, x') = \sum_{i \in N} \lambda_i e_i(x) e_i(x'), \quad (2.14)$$

注意到这个分解不仅依赖于核函数 k , 还依赖于定义在 $\mathcal{X}$ 上的测度 μ 。所以, 为了强调特征值对核函数与测度的依赖性, 也将其记为 $\lambda_i(k; \mathcal{X}, d\mu) = \lambda_i$ 。将在第 3 章的第 3.1 节中对核函数理论进行更为详细的介绍。

在核回归的研究中, λ_i 的渐近行为尤为重要, 因为它反映了核函数所对应的再生核希尔伯特空间 (Reproducing Kernel Hilbert Space, RKHS) 的复杂度, 也与相应核回归估计量的收敛速率密切相关 (例如参见 Caponnetto et al.^[14], Lin et al.^[84])。特别地, 文献中常假设特征值呈幂律衰减, 即 $\lambda_i \asymp i^{-\beta}$, 且相应的极小极大最优速率依赖于指数 β 。因此, 确定所研究核函数的衰减率, 尤其是幂律衰减情况下的指数 β , 是非常重要的。

鉴于全连接 ReLU 神经网络对应的神经切线核本质上是球面上的内积核, 已有大量文献致力于利用球谐函数 (spherical harmonics) 理论^[29]来研究神经切线核的特征分解性质^[30-33]。特别地, Bietti et al.^[30]和 Bietti et al.^[33]的研究表明, 当输入数据服从 $\mathbb{S}^d$ 上的均匀分布时, 与两层及多层神经网络相关联的神经切线核的特征值衰减率为 $i^{-(d+1)/d}$ 。然而, 上述分析严重依赖于球面上的球谐函数理论以导出特征值的显式表达式, 这种方法具有明显的局限性, 即难以直接推广到一般的输入区域及一般的输入分布上。近年来, Lai et al.^[27]针对定义在实数区间上的两层 ReLU 神经网络进行了研究, 证明了其对应的神经切线核的特征值以 i^{-2} 的速率衰减。尽管如此, 该工作的分析技术高度依赖于 $\mathbb{R}$ 上神经切线核的特定一维显式表达, 难以扩展至定义在高维空间 $\mathbb{R}^d$ 上的情形, 也无法涵盖多层宽神经网络对应的神经切线核。因此, 确定一般区域上神经切线核的特征值衰减率仍然是一个重要且具有挑战性的问题, 这也是本章的主要目标。

首先, 可以说明神经切线核是严格正定的。对于它的证明, 可以参考 Arora et al.^[15]和 Li et al.^[28]。

命题 2.1: k_{NT} 在 $\mathcal{X} = \mathbb{R}^d \times \{1\}$ 上是严格正定的, 即对于不同的点 $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, 核矩阵的最小特征值 $\lambda_{\min}(\Theta_{\text{NTK}}(\mathbf{x}_i, \mathbf{x}_j))_{n \times n} > 0$ 。

本章的主要结果如下, 给出了多层 ReLU 神经网络对应的神经切线核在一般区域和一般分布下的特征值衰减率。

定理 2.2: 设 $d \geq 3$, 且 $\mu|_{\Omega}$ 为 $\Omega \subseteq \mathbb{R}^d$ 上的概率测度, 且具有黎曼可积密度 $p(x)$, 满足 $p(x) \leq C(1 + \|x\|^2)^{-(d+3)/2}$ (其中 C 为某常数)。那么, k_{NT} 在 $\mathcal{X} = \Omega \times \{1\}$ 上

关于 μ 的特征值衰减速率为:

$$\lambda_i(k_{\text{NT}}; \mathcal{X}, d\mu) \asymp i^{-\frac{d+1}{d}}. \quad (2.15)$$

以下对于定理 2.2 进行一些讨论。

首先,关于密度 $p(x)$ 的条件被许多常见分布所满足,例如次高斯(sub-Gaussian)分布或具有有界支集的分布。不仅如此,可以看到定理 2.2 给出的衰减速率与神经网络的层数 L 无关,只与输入空间的维度 d 有关,这与 Bietti et al.^[30] 和 Bietti et al.^[33] 在球面上得到的结果一致。这说明,如果仅仅关注特征值的渐近衰减速率,网络的深度并不会影响神经切线核的谱性质。

定理 2.2 的结论和证明方法可以推广到其他类型的激活函数和网络架构。例如,也可以考虑残差神经网络(ResNet)对应的神经切线核,或者是其他激活函数,特别是齐次激活函数,例如 $\text{ReLU}^\alpha(x) = \max(x, 0)^\alpha$ 和 Leaky ReLU $\text{LeakyReLU}(x) = \max(x, 0) + a \min(x, 0)$ (其中 $a \in (0, 1)$ 为一个常数)。

2.2.1 定理 2.2 的证明

为了证明定理 2.2, 将利用核函数的变换性质, 将定义在 $\mathbb{R}^d$ 上的神经切线核问题转化为定义在球面 $\mathbb{S}^d$ 上的齐次神经切线核问题。然后, 可以利用 Bietti et al.^[33] 已经建立的 $\mathbb{S}^d$ 上神经切线核的特征值衰减率结果。具体地, 本章在第 2.2.2 小节中给出基本的核函数变换理论, 然后在第 2.3 节给出本章这一部分的主要定理, 说明限制在球面子区域上的内积核与在整个球面上的内积核的特征值衰减率相同。为保持从一般区域到球面内积核结果的叙述顺序, 下面的证明将先引用后文的定理 2.3, 其完整陈述和证明见第 2.3 节。利用这些结果, 本文给出定理 2.2 的证明如下。

记 $\mathbb{S}_+^d = \{y = (y_1, \dots, y_{d+1}) \in \mathbb{S}^d : y_{d+1} > 0\}$, 并引入同胚映射 $\Phi : \mathbb{R}^d \rightarrow \mathbb{S}_+^d$, 定义为 $x \mapsto \tilde{x}/\|\tilde{x}\|$, 其中 $\tilde{x} = (x, 1) \in \mathbb{R}^{d+1}$ 。易证雅可比矩阵和 Gram 矩阵分别为:

$$J\Phi = \frac{1}{\|\tilde{x}\|} \begin{pmatrix} I_d \\ 0 \end{pmatrix} - \frac{\tilde{x}x^T}{\|\tilde{x}\|^3}, \quad G = (J\Phi)^T J\Phi = \frac{1}{\|\tilde{x}\|^2} I_d - \frac{xx^T}{\|\tilde{x}\|^4}, \quad \det G = \|\tilde{x}\|^{-2(d+1)}.$$

在 $\mathbb{S}^d$ 上定义齐次 NTK $\Theta_{\text{NTK},0}$:

$$\Theta_{\text{NTK},0}(y, y') := \sum_{r=0}^L \kappa_1^{(r)}(u) \prod_{s=r}^{L-1} \kappa_0(\kappa_1^{(s)}(u)), \quad u = \langle y, y' \rangle. \quad (2.16)$$

易验证:

$$K_1(x, x') := \Phi^* \Theta_{\text{NTK},0} = \sum_{r=0}^L \kappa_1^{(r)}(\bar{u}) \prod_{s=r}^{L-1} \kappa_0(\kappa_1^{(s)}(\bar{u})), \quad \Theta_{\text{NTK}} = \|\tilde{x}\| \odot K_1 + 1.$$

因此, 由命题 2.2 和命题 2.3 可得:

$$\lambda_i(\|\tilde{x}\| \odot K_1; \mathcal{X}, d\mu) = \lambda_i(K_1; \mathcal{X}, \|\tilde{x}\|^2 d\mu) = \lambda_i(\Theta_{\text{NTK},0}; \mathbb{S}^d, \Phi_*(\|\tilde{x}\|^2 d\mu)).$$

此外, 记 $\tilde{\sigma} = \Phi_*(\|\tilde{x}\|^2 d\mu)$ 以及 $p(x) = \frac{d\mu}{dx}$, 有 $d\tilde{\sigma} = p(x)\|\tilde{x}\|^2 \Phi_*(dx)$. 另一方面, $\mathbb{S}_+^d$ 上的典范均匀测度 σ 由 $d\sigma = |\det G|^{\frac{1}{2}} \Phi_*(dx)$ 给出, 因此有:

$$q(y) := \frac{d\tilde{\sigma}}{d\sigma} = |\det G|^{-\frac{1}{2}} \|\tilde{x}\|^2 p(x) = \|\tilde{x}\|^{d+3} p(x), \quad y \in \mathbb{S}_+^d.$$

因此, 关于 $p(x)$ 的条件意味着 $q(y)$ 是黎曼可积且有上界的. 现在, Bietti et al.^[33] 已经建立了 $\mathbb{S}^d$ 上内积核 $\Theta_{\text{NTK},0}$ 关于 $d\sigma$ 的 EDR, 即 $\lambda_i(\Theta_{\text{NTK},0}; \mathbb{S}^d, d\sigma) \asymp i^{-\frac{d+1}{d}}$. 由定理 2.3 可知 $\lambda_i(\Theta_{\text{NTK},0}; \mathbb{S}^d, d\tilde{\sigma}) \asymp i^{-\frac{d+1}{d}}$.

2.2.2 核函数的变换与特征值衰减速率

在本小节中介绍关于算子 T 特征值的一些初步结果, 这些结果允许通过代数变换处理核函数, 从而简化分析. 首先定义缩放核 $(\rho \odot k)(x, x') = \rho(x)k(x, x')\rho(x')$, 其中 $\rho: \mathcal{X} \rightarrow \mathbb{R}$ 为某一函数. 易得如下命题:

命题 2.2: 设 $\rho: \mathcal{X} \rightarrow \mathbb{R}$ 为可测函数. 则有:

$$\lambda_i(\rho \odot k; \mathcal{X}, d\mu) = \lambda_i(k; \mathcal{X}, \rho^2 d\mu). \quad (2.17)$$

证明 证明如下的特征值和特征函数的对应关系: (λ_i, e_i) 为 k 关于测度 $\rho^2 d\mu$ 的特征值和特征函数, 则 $(\lambda_i, \rho e_i)$ 为 $\rho \odot k$ 关于测度 μ 的特征值和特征函数, 反之亦然. 按照前者的定义, 有

$$\lambda_i e_i(x) = \int_{\mathcal{X}} k(x, x') e_i(x') \rho^2(x') d\mu(x').$$

两边乘以 $\rho(x)$, 得到

$$\begin{aligned} \lambda_i \rho(x) e_i(x) &= \int_{\mathcal{X}} \rho(x) k(x, x') \rho(x') (\rho(x') e_i(x')) d\mu(x') \\ &= \int_{\mathcal{X}} (\rho \odot k)(x, x') (\rho(x') e_i(x')) d\mu(x'), \end{aligned}$$

这就说明 ρe_i 是 $\rho \odot k$ 关于测度 μ 的特征函数, 且对应的特征值为 λ_i . 反之亦然, 因此完成证明. $\blacksquare$

此外, 如果 ρ 是有界的, 可以利用紧自伴算子特征值的极小极大原理 (Minimax Principle) 进一步估计特征值.

引理 2.1: 设可测函数 $\rho: \mathcal{X} \rightarrow \mathbb{R}$ 满足 $0 \leq c \leq \rho^2(x) \leq C$. 则对于所有 $i = 1, 2, \dots$, 有:

$$c \lambda_i(k; \mathcal{X}, d\mu) \leq \lambda_i(\rho \odot k; \mathcal{X}, d\mu) \leq C \lambda_i(k; \mathcal{X}, d\mu).$$

等价地, 若 ν 是 $\mathcal{X}$ 上的另一个测度, 且满足 $0 \leq c \leq \frac{d\nu}{d\mu} \leq C$, 则对于所有 $i = 1, 2, \dots$, 有:

$$c\lambda_i(k; \mathcal{X}, d\mu) \leq \lambda_i(k; \mathcal{X}, d\nu) \leq C\lambda_i(k; \mathcal{X}, d\mu).$$

证明 利用命题 2.2, 只需要证明关于测度变换的部分。不妨假设 $c > 0$, 不然不等式的前一半是平凡成立的。而由对称性, 只需证明后一半。对于 k 的再生核希尔伯特空间 $\mathcal{H}_k$, 可以引入关于测度 μ (或者 ν) 的协方差算子 $C_\mu : \mathcal{H}_k \rightarrow \mathcal{H}_k$ (或者 C_ν), 定义为

$$\langle C_\mu f, g \rangle_{\mathcal{H}_k} = \int_{\mathcal{X}} f(x)g(x) d\mu(x), \quad f, g \in \mathcal{H}_k.$$

这个协方差算子的特征值与积分算子 T 的特征值相同 (参见第 3.1 节)。使用紧自伴算子特征值的极小极大原理, 引理 2.9, 有

$$\lambda_i(C_\nu) = \inf_{\substack{V \subset \mathcal{H}_k \\ \dim(V)=i-1}} \sup_{\substack{f \in V^\perp \\ \|f\|_{\mathcal{H}_k}=1}} \langle C_\nu f, f \rangle_{\mathcal{H}_k}.$$

利用测度之间的界限关系, 有

$$\langle C_\nu f, f \rangle_{\mathcal{H}_k} = \int_{\mathcal{X}} f^2(x) d\nu(x) \leq C \int_{\mathcal{X}} f^2(x) d\mu(x) = C \langle C_\mu f, f \rangle_{\mathcal{H}_k}.$$

这就说明了

$$\lambda_i(k; \mathcal{X}, d\nu) = \lambda_i(C_\nu) \leq C\lambda_i(C_\mu) = C\lambda_i(k; \mathcal{X}, d\mu).$$

■

注 2.1: 引理 2.1 的证明巧妙地利用了积分算子与再生核希尔伯特空间中协方差算子之间的联系, 从而始终处理的是同一个空间, 即 $\mathcal{H}_k$, 而不需要涉及到不同测度下的 L^2 空间之间的映射关系, 后者往往较为复杂且难以处理。

推论 2.1 (微分同胚变换): 令 $\mathcal{X}_1, \mathcal{X}_2$ 为两个黎曼流形, k 为 $\mathcal{X}_1$ 上的核。假设 $\varphi : \mathcal{X}_1 \rightarrow \mathcal{X}_2$ 为 C^1 -微分同胚, 即 φ 和 φ^{-1} 均为 C^1 , 且 $0 < c \leq |\det J\varphi| \leq C$, 其中 $J\varphi$ 表示 φ 的雅可比矩阵。则有:

$$c\lambda_i(k; \mathcal{X}_1, dx_1) \leq \lambda_i(\varphi_* k; \mathcal{X}_2, dx_2) \leq C\lambda_i(k; \mathcal{X}_1, dx_1). \quad (2.18)$$

证明 由变量替换公式, 有:

$$\lambda_i(k; \mathcal{X}_1, |\det J\varphi| dx_1) = \lambda_i(\varphi_* k; \mathcal{X}_2, dx_2).$$

然后, 应用引理 2.1, 即可完成证明。 ■

接下来, 考虑核函数的变换。设 $\mathcal{X}_1, \mathcal{X}_2$ 为两个集合, $\varphi : \mathcal{X}_1 \rightarrow \mathcal{X}_2$ 为双射, k_2

为 $\mathcal{X}_2$ 上的核。定义 $\mathcal{X}_1$ 上的拉回核 (pull-back kernel) $\varphi^* k_2$ 为:

$$(\varphi^* k_2)(x_1, x'_1) = k_2(\varphi(x_1), \varphi(x'_1)).$$

此外, 假设 $\mathcal{X}_1$ 是具有测度 μ_1 的可测空间, 定义 $\mathcal{X}_2$ 上的前推测度 (push-forward measure) $\mu_2 = \varphi_* \mu_1$, 即 $\mu_2(A) = \mu_1(\varphi^{-1}(A))$ 。此时, 易见:

命题 2.3: 设 $\mathcal{X}_1, \mathcal{X}_2$ 为两个可测空间, $\varphi: \mathcal{X}_1 \rightarrow \mathcal{X}_2$ 为可测单射, μ_1 为 $\mathcal{X}_1$ 上的测度, 且 $\mu_2 = \varphi_* \mu_1$ 。假设 k_2 为 $\mathcal{X}_2$ 上的核, 且 $k_1 = \varphi^* k_2$ 。则有:

$$\lambda_i(k_1; \mathcal{X}_1, d\mu_1) = \lambda_i(k_2; \mathcal{X}_2, d\mu_2). \quad (2.19)$$

证明 假设 (e, λ) 为 k_2 关于测度 μ_2 的特征函数和特征值, 则

$$\lambda e(x_2) = \int_{\mathcal{X}_2} k_2(x_2, x'_2) e(x'_2) d\mu_2(x'_2).$$

将 $x_2 = \varphi(x_1)$ 代入上式, 并利用测度的前推定义, 得到:

$$\begin{aligned} \lambda e(\varphi(x_1)) &= \int_{\mathcal{X}_1} k_2(\varphi(x_1), \varphi(x'_1)) e(\varphi(x'_1)) d\mu_1(x'_1) \\ &= \int_{\mathcal{X}_1} (\varphi^* k_2)(x_1, x'_1) e(\varphi(x'_1)) d\mu_1(x'_1). \end{aligned}$$

这就说明 $e \circ \varphi$ 是 k_1 关于测度 μ_1 的特征函数, 且对应的特征值为 λ 。反之亦然, 因此完成证明。 ■

最后, 再考虑两个核函数之和的情形。

引理 2.2: 设 k_1, k_2 为 $\mathcal{X}$ 上的两个正定核。则

$$\lambda_i(k_1 + k_2; \mathcal{X}, d\mu) \geq \lambda_i(k_1; \mathcal{X}, d\mu).$$

进一步地, 假设当 $i \rightarrow \infty$ 时, $\lambda_i(k_1; \mathcal{X}, d\mu) \asymp \lambda_{2i}(k_1; \mathcal{X}, d\mu)$ 且 $\lambda_i(k_2; \mathcal{X}, d\mu) = O(\lambda_i(k_1; \mathcal{X}, d\mu))$ 。则

$$\lambda_i(k_1 + k_2; \mathcal{X}, d\mu) \asymp \lambda_i(k_1; \mathcal{X}, d\mu), \quad i \rightarrow \infty.$$

证明 对于前一部分, 有

$$T_{k_1+k_2} = T_{k_1} + T_{k_2} \geq T_{k_1},$$

其中 T_k 表示核 k 对应的积分算子, 且 $\geq$ 表示算子半正定性。因此直接得到特征值的不等式。对于后一部分, 利用算子特征值的 Weyl 不等式 (引理 2.10) 得到

$$\begin{aligned} \lambda_{2i-1}(k_1 + k_2; \mathcal{X}, d\mu) &\leq \lambda_i(k_1; \mathcal{X}, d\mu) + \lambda_i(k_2; \mathcal{X}, d\mu) \\ &\lesssim C \lambda_{2i}(k_1; \mathcal{X}, d\mu) + \lambda_i(k_1; \mathcal{X}, d\mu) \\ &\lesssim \lambda_{2i}(k_1; \mathcal{X}, d\mu) \leq \lambda_{2i-1}(k_1; \mathcal{X}, d\mu), \end{aligned}$$

其中第二个和第三个不等式利用了假设条件，最后一个不等式利用了特征值的单调性。对于 λ_{2i} 也可以类似地得到同样的结论，因此完成证明。 ■

注 2.2: 引理 2.2 中的条件 $\lambda_i(k_1; \mathcal{X}, d\mu) \asymp \lambda_{2i}(k_1; \mathcal{X}, d\mu)$ 确保了 k_1 的特征值衰减不会太快，从而避免了 k_2 对总特征值衰减率产生显著影响。例如，当 k_1 的特征值呈幂律衰减时 $\lambda_i \asymp i^{-\beta}$ ，该条件是满足的。但如果 k_1 的特征值衰减非常快，例如指数衰减 $\lambda_i \asymp e^{-\alpha i}$ ，则该条件不成立。将在后续分析中证明，对于神经切线核而言，其主导部分的特征值衰减速率为幂律衰减，因此该条件是满足的。

2.3 限制在子区域上的球面内积核的特征值衰减率

前一小节的结果为分析核函数的特征值提供了有力的工具，但是仍然有许多局限性。例如，即便知道某个核函数 k 在 $\mathcal{X}$ 上的特征值衰减速率，也只能够利用引理 2.1 得到 k 在 $\mathcal{X}$ 的一个子集 $\Omega \subseteq \mathcal{X}$ 上的特征值衰减速率的上界，因为它等价于考虑 μ 在 Ω 上的限制所对应的测度。但是无法利用该引理得到下界，因为此时的 Radon 导数下界为零。事实上，如果没有额外的条件，无法得到这样的下界。例如考虑仅定义在子区域 Ω 上的核函数，通过零延拓，可以将其定义在更大的区域上，此时特征值衰减速率当然不变，但若再将它限制到 Ω 之外的区域上，其特征值将全部为零。

另一方面，假设要研究 $\lambda_i(k_1; \mathcal{X}_1, d\mu_1)$ ，一种自然的想法是寻找某个变换 $\varphi : \mathcal{X}_1 \rightarrow \mathcal{X}_2$ ，使得 k_1 可以表示为某个 $\mathcal{X}_2$ 上核函数 k_2 的拉回核，即 $k_1 = \varphi^* k_2$ 。如果 k_2 关于 $\mathcal{X}_2$ 上某测度 σ 的特征值衰减速率已知或易于获得，那么一种直观的想法是结合命题 2.3 和引理 2.1 来获得 k_1 关于 μ_1 的特征值衰减率。然而，在许多情况下， $\varphi(\mathcal{X}_1)$ 是 $\mathcal{X}_2$ 的真子集，或者前推测度 $\mu_2 = \varphi_* \mu_1$ 并不与 σ 等价，例如 Radon 导数 $\frac{d\mu_2}{d\sigma}$ 可能在某些区域上为零。这样，也无法利用引理 2.1 得到 k_2 关于测度 μ_2 的特征值衰减速率的下界，从而也无法得到 k_1 关于 μ_1 的特征值衰减速率的下界。这正是本章所关心的神经切线核所面临的主要挑战之一。

幸运的是，如果核函数满足适当的不变性性质，仍然能够说明限制在子区域上的核函数的特征值衰减速率与原核函数的特征值衰减速率相同。一种情况是平移不变核函数，即 $k(x, x') = g(x - x')$ 。此时，Widom^[85] 的如下结果极具启发性。这个结果说明了，只要平移不变核的傅里叶系数满足适当的正则性条件，那么限制在子区域上的核函数的特征值衰减速率与原核函数的特征值衰减速率相同。

命题 2.4 (Widom^[85]): 设 $\mathbb{T}^d = [-\pi, \pi)^d$ 为 d 维环面，且

$$k(x, x') = \sum_{n \in \mathbb{Z}^d} c_n e^{in \cdot x} e^{-in \cdot x'}$$

为 $\mathbb{T}^d$ 上的平移不变核。进一步假设 c_n 满足: (i) $c_n \geq 0$; (ii) 固定除 n_{i_0} 外的所有 n_i , c_n 作为 n_{i_0} 的函数, 在 $-\infty$ 到某 $\bar{n} = \bar{n}(i_0)$ 之间非减, 在 $\bar{n}$ 到 ∞ 之间非增; (iii) 若 $|n|, |m| \rightarrow \infty$ 且 $|n| = O(|m|)$, 则 $c_m = O(c_n)$; (iv) 若 $|n|, |m| \rightarrow \infty$ 且 $|n| = o(|m|)$, 则 $c_m = o(c_n)$ 。那么, 对于有界非零黎曼可积函数 ρ , 有:

$$\lambda_i(k; \mathbb{T}^d, \rho^2 dx) \asymp \lambda_i(k; \mathbb{T}^d, dx).$$

然而, 上述结果并不适用于本章的情况, 因为本章关注的神经切线核在环面上不是平移不变的, 而是在球面上旋转不变的, 即内积核。尽管如此, 受此结果启发, 本章的主要贡献之一是为球面上的内积核建立了一个类似的结果。设 $d \geq 3$, $\mathbb{S}^d \subset \mathbb{R}^{d+1}$ 为 d 维单位球面, σ 为 $\mathbb{S}^d$ 上的勒贝格测度。回顾一下, 内积核 $k(x, x')$ 是指仅依赖于输入内积 $u = \langle x, x' \rangle$ 的核。

得益于球谐函数理论^[29], 积分算子 T 的特征函数以及 k 的 Mercer 分解可以显式给出:

$$k(x, x') = \sum_{n=0}^{\infty} \mu_n \sum_{l=1}^{a_n} Y_{n,l}(x) Y_{n,l}(x'), \quad (2.20)$$

其中 $\{Y_{n,l}, n \geq 0, l = 1, \dots, a_n\}$ 是由球谐函数构成的正交基, $a_n = \binom{n+d}{n} - \binom{n-2+d}{n-2}$ 是 n 阶球谐函数空间的维数, μ_n 是算子 T 重数为 a_n 的特征值。为了陈述本章的结果, 首先引入关于特征值渐近衰减率的如下条件。

假设 2.1: 设 $(\mu_n)_{n \geq 0}$ 为正数递减序列。假设如下条件成立:

- (a) 定义 $N(\epsilon) = \max\{n : \mu_n > \epsilon\}$ 。对于任意固定常数 $c > 0$, 当 $\epsilon \rightarrow 0$ 时, $N(c\epsilon) = \Theta(N(\epsilon))$; 假设 $\epsilon, \delta \rightarrow 0$ 且 $\epsilon = o(\delta)$, 则 $N(\delta) = o(N(\epsilon))$ 。
- (b) 对所有 n , 有 $\Delta^{d+1} \mu_n \geq 0$, 其中 Δ 为定义 2.2 中定义的前向差分算子。
- (c) 存在常数 $q \in \mathbb{N}_+$ 和 $D > 0$, 使得对任意 $n \geq 0$, 有

$$\sum_{l=0}^d \binom{\tilde{n}+l}{l} \Delta^l \mu_{\tilde{n}} \leq D \mu_n, \quad \text{其中 } \tilde{n} = qn. \quad (2.21)$$

注 2.3: 假设 2.1 是一个关于衰减率的温和条件, 且定理 2.3 仅要求假设 2.1 在渐近意义下成立, 因此这一要求相当普遍。例如, 若满足以下任一情况, 则该要求成立:

- $\mu_n \asymp n^{-\beta}$, 其中 $\beta > d$ 。
- $\mu_n \asymp n^{-\beta} (\ln n)^p$, 其中 $\beta > d, p \in \mathbb{R}$, 或者 $\beta = d, p > 1$ 。

指数型衰减 $\mu_n \asymp e^{-c_1 n^\beta}$ 满足后文差分和导数控制中所需的正则性, 但它不满足条件 (a) 中关于 $N(\epsilon)$ 的尺度分离要求: 例如取 $\delta = \sqrt{\epsilon}$ 时, $N(\delta)/N(\epsilon)$ 不趋于零。因此不作为假设 2.1 的例子。此外, 这些衰减率条件与现有理论相符, 因为 Widom^[85] 中也需要类似的条件 (ii)-(iv)。

现在，可以陈述关于球面内积核在一般测度下的特征值衰减速率的主要结果。

定理 2.3: 设 $d \geq 3$ ，且 $k(x, x')$ 为 $\mathbb{S}^d$ 上的内积核，其在分解 (2.20) 中对应的特征值为 $(\mu_n)_{n \geq 0}$ 。假设存在一个满足假设 2.1 的序列 $(\tilde{\mu}_n)_{n \geq 0}$ ，使得 $\mu_n \asymp \tilde{\mu}_n$ 。则对于 $\mathbb{S}^d$ 上有界非零的黎曼可积函数 ρ ，有：

$$\lambda_i(k; \mathbb{S}^d, \rho^2 d\sigma) \asymp \lambda_i(k; \mathbb{S}^d, d\sigma). \quad (2.22)$$

作为本章的主要技术贡献，该定理是 Widom^[85] 中结果的一个非平凡推广，将其从环面适配到了球面。本章遵循 Widom^[85] 的基本思想，通过首先证明主要引理（引理 2.5）来建立该定理，但由于特征系统差异巨大，Widom^[85] 的方法不再适用。为了证明主要引理，本章利用了球面上精细的调和分析，结合 Cesàro 求和技术和特征值的左外推，这需要假设 2.1 中微妙的条件。

定理 2.3 表明，内积核关于一般测度的特征值衰减速率与该核关于均匀测度的特征值衰减速率相同。结合第 2.2.2 小节中的结果，这提供了一种确定一般域上核特征值衰减速率的新方法。可以首先将核变换为关于某种测度的球面内积核；然后利用定理 2.3 表明，若考虑球面上均匀测度，所得内积核的衰减率保持不变；最后通过分析工具确定内积核在整个球面上的衰减率，因为可以解析地计算核函数的球谐函数展开式。这种方法能够确定一般域上多层神经网络对应的神经切线核的特征值衰减速率。

在本节中，对于紧自伴算子 A 和 $\epsilon > 0$ ，记 $N^\pm(\epsilon, A)$ 为 A 严格大于（小于） ϵ （ $-\epsilon$ ）的特征值的个数。由于 A 为紧自伴算子，其非零特征值仅有有限个正值和有限个负值，因此 $N^\pm(\epsilon, A)$ 总是有限的。对于集合 Ω ，记 P_Ω 为示性函数（indicator function） $\mathbf{1}_\Omega$ 的乘法算子。

2.3.1 球谐函数

首先，介绍球谐函数及后续将用到的一些性质。更多细节可参考 Dai et al.^[29]。设 σ 为 $\mathbb{S}^d$ 上的勒贝格测度， $L^2(\mathbb{S}^d)$ 为配备了如下内积的（实）希尔伯特空间：

$$\langle f, g \rangle_{L^2(\mathbb{S}^d)} = \frac{1}{\omega_d} \int_{\mathbb{S}^d} f g \, d\sigma.$$

根据球谐函数理论，拉普拉斯-贝尔特拉米（Laplace-Beltrami）算子 $\Delta_{\mathbb{S}^d}$ （即球面拉普拉斯算子）的特征系统给出了如下的正交直和分解： $L^2(\mathbb{S}^d) = \bigoplus_{n=0}^{\infty} \mathcal{H}_n^d(\mathbb{S}^d)$ ，其中 $\mathcal{H}_n^d(\mathbb{S}^d)$ 是 $d+1$ 元 n 次齐次调和多项式在 $\mathbb{S}^d$ 上的限制，且 $\mathcal{H}_n^d(\mathbb{S}^d)$ 中的每个元素都是 $\Delta_{\mathbb{S}^d}$ 的特征函数，对应的特征值为 $-n(n+d-1)$ 。这给出了 $L^2(\mathbb{S}^d)$ 的一组标准正交基：

$$\{Y_{n,l}, l = 1, \dots, a_n, n = 1, 2, \dots\},$$

其中 $a_n = \binom{n+d}{n} - \binom{n-2+d}{n-2} \asymp n^{d-1}$ 是空间 $\mathcal{H}_n^d(\mathbb{S}^d)$ 的维数, $Y_{n,l} \in \mathcal{H}_n^d(\mathbb{S}^d)$ 。注意到:

$$\sum_{n \leq N} a_n = C_{N+d}^N + C_{N-1+d}^{N-1} \asymp N^d. \quad (2.23)$$

此外, 求和项

$$Z_n(\mathbf{x}, \mathbf{y}) = \sum_{l=1}^{a_n} Y_{n,l}(\mathbf{x}) Y_{n,l}(\mathbf{y}) \quad (2.24)$$

不依赖于标准正交基 $Y_{n,l}$ 的选取, Z_n 被称为带球多项式 (Zonal Polynomials)。当 $d \geq 2$ 时, 有:

$$Z_n(\mathbf{x}, \mathbf{y}) = \frac{n+\lambda}{\lambda} C_n^\lambda(u), \quad u = \langle \mathbf{x}, \mathbf{y} \rangle, \quad \lambda = \frac{d-1}{2}, \quad (2.25)$$

其中 C_n^λ 为盖根堡 (Gegenbauer) 多项式。

球谐函数的一个关键性质是如下的 Funk-Hecke 公式, 见 Dai et al.^[29] 的定理 1.2.9。

命题 2.5 (Funk-Hecke 公式): 设 $d \geq 3$, f 为可积函数且满足 $\int_{-1}^1 |f(t)|(1-t^2)^{d/2-1} dt < \infty$ 。则对于任意 $Y_n \in \mathcal{H}_n^d(\mathbb{S}^d)$, 有

$$\frac{1}{\omega_d} \int_{\mathbb{S}^d} f(\langle \mathbf{x}, \mathbf{y} \rangle) Y_n(\mathbf{y}) d\sigma(\mathbf{y}) = \mu_n(f) Y_n(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{S}^d, \quad (2.26)$$

其中常数 $\mu_n(f)$ 定义为 $\mu_n(f) = \omega_d \int_{-1}^1 f(t) \frac{C_n^\lambda(t)}{C_n^\lambda(1)} (1-t^2)^{\frac{d-2}{2}} dt$ 。

还需要以下关于带球多项式 Cesàro 和的定理。关于 Cesàro 和以及系数 A_{n-k}^d 的定义可参见第 2.3.3 小节。

命题 2.6: 设

$$K_n = \frac{1}{A_n^d} \sum_{k=0}^n A_{n-k}^d \frac{k+\lambda}{\lambda} C_k^\lambda(u), \quad (2.27)$$

为 $\frac{k+\lambda}{\lambda} C_k^\lambda$ 的 d -阶 Cesàro 和。则存在正常数 C , 使得对于所有 $n \geq 1$, 有

$$0 \leq K_n(u) \leq C n^{-1} (1-u+n^{-2})^{-(\lambda+1)}. \quad (2.28)$$

证明 参见 Dai et al.^[29] 的定理 2.4.3 和引理 2.4.6。 ■

2.3.2 球面上的内积核

对比 (2.26) 与积分算子的定义 (2.13), Funk-Hecke 公式表明 Y_n 是球面上任意内积核 $k(\mathbf{x}, \mathbf{y}) = f(\langle \mathbf{x}, \mathbf{y} \rangle)$ 的特征函数。因此, 任意的内积核 $k(\mathbf{x}, \mathbf{y})$ 总是有如下的 Mercer 分解和谱分解:

$$k(\mathbf{x}, \mathbf{y}) = \sum_{n=0}^{\infty} \mu_n \sum_{l=1}^{a_n} Y_{n,l}(\mathbf{x}) Y_{n,l}(\mathbf{y}), \quad T = \sum_{n=0}^{\infty} \mu_n \sum_{l=1}^{a_n} Y_{n,l} \otimes Y_{n,l}. \quad (2.29)$$

在此需要注意, μ_n 是重数为 a_n 的特征值, 不应将其与计入重数的特征值 λ_i 混淆。鉴于 (2.29), 可以将内积核及其对应的积分算子与序列 $(\mu_n)_{n \geq 0}$ 联系起来。

注 2.4: 可以将 (2.29) 与环面上的平移不变核的傅里叶展开式进行对比。容易证明, 对于平移不变核, 即 $k(x, x') = g(x - x')$, 在环面的均匀测度下, 其 Mercer 分解为

$$k(x, x') = \sum_{n \in \mathbb{Z}^d} \lambda_n e^{i\langle n, x \rangle} e^{-i\langle n, x' \rangle}.$$

此外, 由于每个 μ_n 的重数为 a_n , 由 (2.23) 可得

$$N^+(\epsilon, T) = \sum_{n \leq N(\epsilon)} a_n \asymp N(\epsilon)^d, \quad (2.30)$$

其中 $N(\epsilon) = \max\{n : \mu_n > \epsilon\}$ 定义同假设 2.1(a)。这给出了渐近速率 λ_i 与 μ_n 之间的简单关系。

命题 2.7: 假设 T 由 (2.29) 给定, 且对于某 $\beta > d$, 有 $\mu_n \asymp n^{-\beta}$ 。则有: (a) $N^+(\epsilon, T) \asymp \epsilon^{-d/\beta}$; (b) $\lambda_i(T) \asymp i^{-\beta/d}$; (c) 设 A 为满足 $N^+(\epsilon, A) \asymp N^+(\epsilon, T)$ 的自伴紧正算子, 则 $\lambda_i(A) \asymp i^{-\beta/d}$ 。

证明 对于 (a), 由于 d 是固定的, 有

$$\sum_{n \leq N} a_n = C_{N+d}^N + C_{N-1+d}^{N-1} \asymp N^d.$$

假设对于 $C_1, C_2 > 0$, 有 $C_1 n^{-\beta} \leq \mu_n \leq C_2 n^{-\beta}$ 。那么,

$$N^+(\epsilon, T) = \#\{i : \lambda_i > \epsilon\} = \sum_{n: \mu_n > \epsilon} a_n \leq \sum_{n: C_2 n^{-\beta} > \epsilon} a_n = \sum_{n \leq C_2^{1/\beta} \epsilon^{-1/\beta}} a_n \lesssim \epsilon^{-\frac{d}{\beta}},$$

并且

$$N^+(\epsilon, T) = \sum_{n: \mu_n > \epsilon} a_n \geq \sum_{n: C_1 n^{-\beta} > \epsilon} a_n = \sum_{n \leq C_1^{1/\beta} \epsilon^{-1/\beta}} a_n \gtrsim \tilde{C}_1 \epsilon^{-\frac{d}{\beta}}.$$

由于 (b) 是 (c) 的特例, 只需证明 (c)。设 $\{\lambda_i\}$ 为 A 的降序排列的特征值集合。取 $\epsilon = \lambda_i$, 得到

$$i \geq N^+(\lambda_i, A) \gtrsim N^+(\lambda_i, T) \gtrsim \lambda_i^{-d/\beta},$$

因此 $\lambda_i \lesssim i^{-\beta/d}$ 。另一方面, 对于某个较小的 $\delta > 0$, 取 $\epsilon = (1 - \delta)\lambda_i < \lambda_i$, 得到

$$i \leq N^+((1 - \delta)\lambda_i, A) \lesssim N^+((1 - \delta)\lambda_i, T) \lesssim \lambda_i^{-d/\beta},$$

因此 $\lambda_i \gtrsim i^{-\beta/d}$, 证明完毕。 ■

命题 2.8: 序列 $\mu_n = c_0(n+1)^{-\beta}$ (其中 $\beta > d$) 满足假设 2.1。

证明 首先, 条件 (a) 成立, 因为由命题 2.7 可知 $\Psi(\delta) \asymp \delta^{-d/\beta}$ 。现在, 令 $N = \min\{n :$

$\mu_n \leq \delta\} \asymp \delta^{-1/\beta}$, 条件 (b) 可由下式推得:

$$\sum_{n: \mu_n \leq \delta} a_n \mu_n^2 \leq C \sum_{n \geq N} a_n \mu_n^2 \leq C \sum_{n \geq N} n^{d-1} n^{-2\beta} \leq C N^{d-2\beta} \leq C \delta^2 N^d \leq C \delta^2 \Psi(\delta).$$

最后, 对于条件 (c), 命题 2.11 表明 $0 < \Delta^p \mu_n \leq c_0(\beta)_p (n+1)^{-(\beta+p)}$, 因此

$$\begin{aligned} \sum_{l=0}^d \binom{n+l}{n} \Delta^l \mu_n &\leq c_0 \sum_{l=0}^d \frac{(n+1)_l}{l!} (\beta)_l (n+1)^{-(\beta+l)} \\ &\leq c_0 \sum_{l=0}^d \frac{(\beta)_l (n+1)_l}{(n+1)^l l!} (n+1)^{-\beta} \leq C(n+1)^{-\beta}. \end{aligned}$$

■

2.3.3 Cesàro 求和

在分析内积核时, 需要使用 Cesàro 求和, 相关内容也可参考 Dai et al.^[29] 的第 A.4 节。

定义 2.1: 设 $p \geq 0$ 。序列 a_k 的 p -阶 Cesàro 和 s_n 定义为

$$s_n^p = \frac{1}{A_n^p} \sum_{k=0}^n A_{n-k}^p a_k, \quad A_k^p := \binom{k+p}{k}. \quad (2.31)$$

注 2.5: 根据定义, 有

$$\sum_{k=0}^n A_{n-k}^p = \sum_{k=0}^n A_k^p = \sum_{k=0}^n \binom{k+p}{k} = \binom{n+p+1}{n} = A_n^{p+1}.$$

如果 $p = 0$, 则 $A_k^p = 1$, 此时 $s_n^0 = \sum_{k=0}^n a_k$ 为普通和。

如果 $p = 1$, 则 $A_k^p = k+1$, 此时

$$s_n^1 = \frac{1}{n+1} \sum_{k=0}^n (n-k+1) a_k = \frac{1}{n+1} \sum_{k=0}^n \left(\sum_{r=0}^k a_k \right)$$

为前缀和的平均值。

定义 2.2 (差分): 设 $\mathbf{a} = (a_k)_{k \geq 0}$ 为一序列。定义序列上的差分算子如下:

$$\left(\Delta^0 \mathbf{a} \right)_k = a_k, \quad (\Delta \mathbf{a})_k = a_k - a_{k+1}, \quad \Delta^{p+1} \mathbf{a} = \Delta(\Delta^p \mathbf{a}). \quad (2.32)$$

常记 $(\Delta^p \mathbf{a})_k = \Delta^p a_k$ 。易证 $\Delta^p a_k = \sum_{r=0}^p \binom{p}{r} (-1)^r a_{k+r}$ 。

定义 2.3 (尾和): 设 $\mathbf{a} = (a_k)_{k \geq 0}$ 为一序列。假设以下所有求和均绝对收敛, 定义序列上的尾和算子如下:

$$(S^0 \mathbf{a})_k = a_k, \quad (S \mathbf{a})_k = \sum_{r \geq k} a_r, \quad S^{p+1} \mathbf{a} = S(S^p \mathbf{a}). \quad (2.33)$$

常记 $(S^p \mathbf{a})_k = S^p a_k$ 。

以下命题给出了尾和与差分之间的基本关系。

命题 2.9: 有如下结果: (a) $S^p a_n = \sum_{k=0}^{\infty} A_k^{p-1} a_{n+k}$; (b) $S \Delta a_n = \Delta S a_n = a_n$; (c) 于是, $\sum_{k=0}^{\infty} A_k^p \Delta^{p+1} a_{n+k} = (S^{p+1} \Delta^{p+1} a)_n = a_n$ 。

证明 对于 (a), 计算

$$S^p a_n = \sum_{i_1 \geq n} \sum_{i_2 \geq i_1} \cdots \sum_{i_p \geq i_{p-1}} a_{i_p} = \sum_{i_p \geq n} \sum_{n \leq i_1 \leq i_2 \leq \cdots \leq i_{p-1} \leq i_p} a_{i_p} = \sum_{i_p \geq n} \binom{i_p - n + p - 1}{p - 1} a_{i_p} = \sum_{k=0}^{\infty} A_k^{p-1} a_{n+k}.$$

对于 (b), 类似地, 有

$$\begin{aligned} (S \Delta a)_n &= \sum_{k \geq n} \Delta a_k = \sum_{k \geq n} (a_k - a_{k+1}) = a_n, \\ (\Delta S a)_n &= S a_n - S a_{n+1} = \sum_{k \geq n} a_k - \sum_{k \geq n+1} a_k = a_n. \end{aligned}$$

■

有如下的分部求和公式, 另见 Dai et al.^{[29](A.4.8)}。

命题 2.10 (分部求和): 设 a_k, b_k 为两个序列, $p \in \mathbb{N}$ 。则

$$\sum_{k=0}^{\infty} a_k b_k = \sum_{k=0}^{\infty} \Delta^{p+1} b_k \sum_{j=0}^k A_{k-j}^p a_j = \sum_{k=0}^{\infty} \Delta^{p+1} b_k A_k^p s_k^p,$$

其中 s_k^p 是 a_k 的 p -Cesàro 均值。

对于函数 $f : [a, b] \rightarrow \mathbb{R}$, 可以类似地定义 $\Delta f(x) = f(x) - f(x+1)$ 以及 $\Delta^{p+1} f(x) = \Delta^p f(x) - \Delta^p f(x+1)$ 。以下基本引理建立了函数差分与导数之间的联系。

引理 2.3: 设 $p \in \mathbb{N}$ 。假设 $f \in C^p([a, a+p])$, 则

$$\Delta^p f(a) = (-1)^p \int_{[0,1]^p} f^{(p)}(a + t_1 + \cdots + t_p) dt_1 \cdots dt_p. \quad (2.34)$$

证明 用归纳法证明。当 $p = 1$ 时,

$$\Delta f(a) = f(a) - f(a+1) = - \int_{[0,1]} f'(a + t_1) dt_1.$$

假设 (2.34) 对 p 成立。则

$$\begin{aligned} \Delta^{p+1} f(a) &= \Delta^p f(a) - \Delta^p f(a+1) \\ &= (-1)^p \left[\int_{[0,1]^p} f^{(p)}(a + \sum_{i=1}^p t_i) dt_1 \cdots dt_p - \int_{[0,1]^p} f^{(p)}(a+1 + \sum_{i=1}^p t_i) dt_1 \cdots dt_p \right] \\ &= (-1)^p \int_{[0,1]^p} \left[f^{(p)}(a + \sum_{i=1}^p t_i) - f^{(p)}(a+1 + \sum_{i=1}^p t_i) \right] dt_1 \cdots dt_p \end{aligned}$$

$$\begin{aligned}
 &= (-1)^{p+1} \int_{[0,1]^p} \left[\int_{[0,1]} f^{(p+1)}(a + \sum_{i=1}^p t_i + t_{p+1}) dt_{p+1} \right] dt_1 \cdots dt_p \\
 &= (-1)^{p+1} \int_{[0,1]^{p+1}} f^{(p+1)}(a + \sum_{i=1}^{p+1} t_i) dt_1 \cdots dt_{p+1},
 \end{aligned}$$

因此 (2.34) 对 $p+1$ 成立。 ■

命题 2.11: 假设 $\mu_k = c_0(k+1)^{-\beta}$, $k \geq 0$ 。令 $(\beta)_p := \beta(\beta+1)\cdots(\beta+p-1)$, 则

$$0 < \Delta^p \mu_k \leq c_0(\beta)_p (k+1)^{-(\beta+p)}.$$

证明 应用前一个引理, 取 $f(x) = c(x+1)^{-\beta}$ 以及 $f^{(p)}(x) = (-1)^p c_0(\beta)_p (x+1)^{-(\beta+p)}$ 。 ■

引理 2.4: 设 $\mu = (\mu_k)_{k \geq 0}$ 为一序列, 使得对于某个 $p \geq 0$, 有 $\Delta^p \mu_k \geq 0, \forall k \geq 0$ 。

给定 $N \geq 0$, 可以构造一个左外推序列 $(\tilde{\mu}_k)_{k \geq 0}$ 使得:

- 当 $k \geq N$ 时 $\tilde{\mu}_k = \mu_k$, 当 $k < N$ 时 $\tilde{\mu}_k \leq \mu_k$;
- 对于所有 $k \geq 0$, 有 $\Delta^p \tilde{\mu}_k \geq 0$;
- 令 $\bar{\mu}_k = \mu_k - \tilde{\mu}_k$ 为残差序列, 则对于所有 $k \geq 0$, 有 $\Delta^p \bar{\mu}_k \geq 0$;
- 首项满足

$$\tilde{\mu}_0 = \mathcal{L}_N^p \mu := \sum_{l=0}^{p-1} A_N^l \Delta^l \mu_N.$$

注: $\mathcal{L}_N^p$ 的形式与假设 2.1 中的 (2.21) 的左端项相同。

证明 通过其 p -阶差分递归地定义 $\tilde{\mu}_k$ 。令 $\Delta^p \tilde{\mu}_k = 0$ ($k < N$) 以及 $\Delta^p \tilde{\mu}_k = \Delta^p \mu_k$ ($k \geq N$)。然后, 通过迭代求和即可得出 (1)、(2) 以及递归公式:

$$\Delta^{p-s} \tilde{\mu}_{N-r} = \sum_{l=0}^{s-1} A_r^l \Delta^{p-s+l} \mu_N, \quad s = 1, \dots, p, \quad r = 0, \dots, N-1,$$

这给出了 (4)。最后一个陈述 (3) 源于 $\Delta^p \bar{\mu}_k = \Delta^p \mu_k - \Delta^p \tilde{\mu}_k \geq 0$ 这一事实。 ■

2.3.4 主要引理

本章关键结果——定理 2.3 的证明思路沿用了 Widom^[85] 的路线: 首先建立下面的关键引理 (引理 2.5), 然后利用球面上区域的分解来证明定理 2.3。

这里的主要技术贡献在于引理 2.5 的证明, 它是 Widom^[85] 中主要引理的球面版本。由于现在的特征系统由球谐函数给出, Widom^[85] 中的方法不再适用。本章中的证明需要对球谐函数进行更精细的分析, 并结合了 Cesàro 求和技术以及特征值的左外推 (Left Extrapolation)。先来看该引理的陈述。

引理 2.5: 设 T 由 (2.29) 给定, 其降序特征值为 $\mu = (\mu_n)_{n \geq 0}$ 。进一步假设 $(\mu_n)_{n \geq 0}$

满足假设 2.1。设 Ω_1, Ω_2 为两个具有分段光滑边界的不交区域。则当 $\epsilon \rightarrow 0$ 时, 有

$$N^\pm(\epsilon, P_{\Omega_1} T P_{\Omega_2} + P_{\Omega_2} T P_{\Omega_1}) = o(N^+(\epsilon, T)). \quad (2.35)$$

证明 设 $\delta > \epsilon > 0$, 其中 δ 稍后确定。取 $M_\delta = \min\{n : \mu_n \leq \delta\} \leq M_\epsilon = \min\{n : \mu_n \leq \epsilon\}$ 。利用引理 2.4 (取 $p = d + 1$ 并结合假设 2.1), 首先构造序列 $\mu^{(1)}$ 作为 μ 在 qM_ϵ 处的左外推; 然后构造序列 $\mu^{(2)}$ 作为残差序列 $\mu - \mu^{(1)}$ 在 qM_δ 处的左外推; 并记 $\mu^{(3)} = \mu - \mu^{(1)} - \mu^{(2)}$, 其中 q 为假设 2.1 中指定的整数。这三个序列满足:

$$\begin{aligned} \mu_n &= \mu_n^{(1)} + \mu_n^{(2)} + \mu_n^{(3)}, \quad \Delta^{d+1} \mu_n^{(i)} \geq 0, \quad i = 1, 2, 3; \\ \mu_0^{(1)} &= \mathcal{L}_{qM_\epsilon}^{d+1} \mu \leq D\mu_{M_\epsilon} \leq D\epsilon; \\ \mu_n^{(2)} &= 0, \quad \forall n \geq qM_\epsilon, \quad \mu_0^{(2)} = \mathcal{L}_{qM_\delta}^{d+1} (\mu - \mu^{(1)}) \leq \mathcal{L}_{qM_\delta}^{d+1} \mu \leq D\delta; \\ \mu_n^{(3)} &= 0, \quad \forall n \geq qM_\delta, \end{aligned} \quad (2.36)$$

其中控制项 $\mathcal{L}_{qM_\delta}^{d+1} \mu \leq C\mu_M$ 来自 (2.21)。现在, 定义 T_i 为与 $\mu^{(i)}$ 关联的积分算子, 即 $T_i = \sum_{n=0}^{\infty} \mu_n^{(i)} \sum_{l=1}^{a_n} Y_{n,l} \otimes Y_{n,l}$ 。令 $N_i^+(\epsilon)$ 为算子 $P_{\Omega_1} T_i P_{\Omega_2} + P_{\Omega_2} T_i P_{\Omega_1}$ 大于 ϵ 的特征值个数。由引理 2.6 有:

$$N^+((2D+1)\epsilon, P_{\Omega_1} T P_{\Omega_2} + P_{\Omega_2} T P_{\Omega_1}) \leq N_1^+(2D\epsilon) + N_2^+(\epsilon) + N_3^+(0). \quad (2.37)$$

对于 $N_1^+(2D\epsilon)$, 注意到 $\|T_1\| \leq D\epsilon$, 因此

$$\|P_{\Omega_1} T_1 P_{\Omega_2} + P_{\Omega_2} T_1 P_{\Omega_1}\| \leq 2D\epsilon,$$

这意味着:

$$N_1^+(2D\epsilon) = 0. \quad (2.38)$$

对于 $N_3^+(0)$, 由于 $\mu_n^{(3)} \neq 0$ 仅当 $n < qM_\delta$ 时成立, 所以

$$N_3^+(0) \leq 2\text{Rank}(T_3) \leq 2 \sum_{n < qM_\delta} a_n \leq 2Cq^d \sum_{n < M_\delta} a_n = 2CN^+(\delta, T), \quad (2.39)$$

其中第三个不等式使用了 (2.23), 且 $\sum_{n < M_\delta} a_n = N^+(\delta, T)$ (注意 μ_n 的重数为 a_n)。

剩下的任务是估计 $N_2^+(\epsilon)$ 。首先, 根据 HS-范数 (Hilbert-Schmidt norm) 的定义及 Simon^[86] 的定理 3.8.5, 有:

$$\begin{aligned} \epsilon^2 N_2^+(\epsilon) &\leq \|P_{\Omega_1} T_2 P_{\Omega_2} + P_{\Omega_2} T_2 P_{\Omega_1}\|_{\text{HS}}^2 \\ &= 2 \int_{\Omega_1} \int_{\Omega_2} \left| \sum_n \mu_n^{(2)} Z_n(\mathbf{x}, \mathbf{y}) \right|^2 d\mathbf{y} d\mathbf{x} \\ &=: 2I. \end{aligned} \quad (2.40)$$

固定 Ω_2 的一个内点 $\mathbf{e}$, 引入一个等距变换 $R_{\mathbf{e}, \mathbf{x}}$ 使得 $R_{\mathbf{e}, \mathbf{x}} \mathbf{e} = \mathbf{x}$ 。如果 $\mathbf{e}, \mathbf{x}$ 不平行,

可取其为由 $\mathbf{e}, \mathbf{x}$ 张成平面上的旋转；若 $\mathbf{x} = \mathbf{e}$ 则取恒等映射；若 $\mathbf{x} = -\mathbf{e}$ 则取反射。由于 $R_{\mathbf{e}, \mathbf{x}}$ 是等距的且 $Z_n(\mathbf{x}, \mathbf{y})$ 仅依赖于 $\langle \mathbf{x}, \mathbf{y} \rangle$ ，有：

$$\begin{aligned}
 I &= \int_{\Omega_1} \int_{\Omega_2} \left| \sum_n \mu_n^{(2)} Z_n(\mathbf{x}, \mathbf{y}) \right|^2 d\mathbf{y} d\mathbf{x} = \int_{\Omega_1} \int_{\Omega_2} \left| \sum_n \mu_n^{(2)} Z_n(R_{\mathbf{e}, \mathbf{x}}^{-1} \mathbf{x}, R_{\mathbf{e}, \mathbf{x}}^{-1} \mathbf{y}) \right|^2 d\mathbf{y} d\mathbf{x} \\
 &= \int_{\Omega_1} \int_{\Omega_2} \left| \sum_n \mu_n^{(2)} Z_n(\mathbf{e}, R_{\mathbf{e}, \mathbf{x}}^{-1} \mathbf{y}) \right|^2 d\mathbf{y} d\mathbf{x} = \int_{\Omega_1} \int_{R_{\mathbf{e}, \mathbf{x}}^{-1} \Omega_2} \left| \sum_n \mu_n^{(2)} Z_n(\mathbf{e}, \mathbf{y}) \right|^2 d\mathbf{y} d\mathbf{x} \\
 &= \iint_{\mathbb{S}^d \times \mathbb{S}^d} \mathbf{1}_{\{\mathbf{x} \in \Omega_1, \mathbf{y} \in R_{\mathbf{e}, \mathbf{x}}^{-1} \Omega_2\}} \left| \sum_n \mu_n^{(2)} Z_n(\mathbf{e}, \mathbf{y}) \right|^2 d\mathbf{x} d\mathbf{y} \\
 &= \int_{\mathbb{S}^d} \left(\int_{\mathbb{S}^d} \mathbf{1}_{\{\mathbf{x} \in \Omega_1, R_{\mathbf{e}, \mathbf{x}} \mathbf{y} \in \Omega_2\}} d\mathbf{x} \right) \left| \sum_n \mu_n^{(2)} Z_n(\mathbf{e}, \mathbf{y}) \right|^2 d\mathbf{y} \\
 &= \int_{\mathbb{S}^d} |\{\mathbf{x} \in \Omega_1 : R_{\mathbf{e}, \mathbf{x}} \mathbf{y} \in \Omega_2\}| \left| \sum_n \mu_n^{(2)} Z_n(\mathbf{e}, \mathbf{y}) \right|^2 d\mathbf{y} \\
 &\leq C \int_{\mathbb{S}^d} \arccos \langle \mathbf{y}, \mathbf{e} \rangle \left| \sum_n \mu_n^{(2)} Z_n(\mathbf{e}, \mathbf{y}) \right|^2 d\mathbf{y},
 \end{aligned}$$

其中最后一个不等式来自命题 2.14。令 $\eta > 0$ （稍后确定），将最后的积分分解为两部分：

$$I = I_1 + I_2 = \left(\int_{\langle \mathbf{y}, \mathbf{e} \rangle > 1-\eta} + \int_{\langle \mathbf{y}, \mathbf{e} \rangle < 1-\eta} \right) \arccos \langle \mathbf{y}, \mathbf{e} \rangle \left| \sum_n \mu_n^{(2)} Z_n(\mathbf{e}, \mathbf{y}) \right|^2 d\mathbf{y}.$$

对于 I_1 ，使用估计 $\arccos u \leq C \sqrt{1-u}$ ，得到：

$$\begin{aligned}
 I_1 &\leq \int_{\langle \mathbf{y}, \mathbf{e} \rangle > 1-\eta} C \eta^{\frac{1}{2}} \left| \sum_n \mu_n^{(2)} Z_n(\mathbf{e}, \mathbf{y}) \right|^2 d\mathbf{y} \leq C \eta^{\frac{1}{2}} \int_{\mathbb{S}^d} \left| \sum_n \mu_n^{(2)} Z_n(\mathbf{e}, \mathbf{y}) \right|^2 d\mathbf{y} \\
 &= C \eta^{\frac{1}{2}} \sum_n a_n (\mu_n^{(2)})^2 \leq C \eta^{\frac{1}{2}} \sum_{n < q M_\epsilon} a_n (\mu_0^{(2)})^2 \\
 &\leq C \eta^{\frac{1}{2}} q^d \sum_{n < M_\epsilon} a_n (\mu_0^{(2)})^2 \leq C \eta^{\frac{1}{2}} N^+(\epsilon, T) \delta^2,
 \end{aligned} \tag{2.41}$$

其中在推导中多次使用了 (2.36) 和 (2.23)。

对于 I_2 ，回顾 (2.25) 并记 $u = \langle \mathbf{y}, \mathbf{e} \rangle$ ，有

$$I_2 = \int_{-1}^{1-\eta} \left| \sum_n \mu_n^{(2)} \frac{n+\lambda}{\lambda} C_n^\lambda(u) \right|^2 (1-u^2)^{\frac{d-2}{2}} \arccos u du,$$

其中 $\lambda = \frac{d-1}{2}$ 。利用分部求和（命题 2.10），可得

$$\sum_n \mu_n^{(2)} \frac{n+\lambda}{\lambda} C_n^\lambda(u) = \sum_n \Delta^{d+1} \mu_n^{(2)} A_n^d K_n(u),$$

其中 K_n 是 (2.27) 中定义的 $C_k^\lambda(u)$ 的 d -阶 Cesàro 和。此外，由命题 2.6 中的 (2.28) 可得：

$$\begin{aligned} \left| \sum_n \mu_n^{(2)} \frac{n+\lambda}{\lambda} C_n^\lambda(u) \right| &= \sum_n \Delta^{d+1} \mu_n^{(2)} A_n^d K_n(u) \leq C(1-u)^{-(\lambda+1)} \sum_n A_n^d \Delta^{d+1} \mu_n^{(2)} \\ &= C(1-u)^{-(\lambda+1)} \mu_0^{(2)} \leq C(1-u)^{-(\lambda+1)} \delta, \end{aligned}$$

其中倒数第二个等式来自命题 2.9。将此估计代回 I_2 ，得到：

$$I_2 \leq C\delta^2 \int_{-1}^{1-\eta} (1-u)^{-2(\lambda+1)} (1-u^2)^{\frac{d-2}{2}} (1-u)^{1/2} du \leq C\delta^2 \eta^{-\frac{d-1}{2}}. \quad (2.42)$$

结合估计 (2.41) 和 (2.42)。取 $\eta = N^+(\delta, T)^{-\frac{2}{d}}$ ，有

$$I \leq I_1 + I_2 \leq C\delta^2 N^+(\epsilon, T)^{\frac{d-1}{d}},$$

因此由 (2.40) 可得：

$$N_2^+(\epsilon) \leq C \left(\frac{\delta}{\epsilon} \right)^2 N^+(\epsilon, T)^{\frac{d-1}{d}}. \quad (2.43)$$

最后，将 (2.38), (2.43) 和 (2.39) 代入 (2.37)，得到：

$$N^+((2D+1)\epsilon, P_{\Omega_1} T P_{\Omega_2} + P_{\Omega_2} T P_{\Omega_1}) \leq 2N^+(\delta, T) + C \left(\frac{\delta}{\epsilon} \right)^2 N^+(\epsilon, T)^{\frac{d-1}{d}}.$$

现在，(2.30) 能够推导出关于 $N^+(\epsilon, T)$ 的类似于假设 2.1(a) 的性质。因此，取 $\delta = \epsilon N^+(\epsilon, T)^{\frac{1}{4d}}$ 使得 $\epsilon = o(\delta)$ ，有 $N^+(\delta, T) = o(N^+(\epsilon, T))$ ，从而

$$N^+((2D+1)\epsilon, P_{\Omega_1} T P_{\Omega_2} + P_{\Omega_2} T P_{\Omega_1}) \leq o(N^+(\epsilon, T)) + C N^+(\epsilon, T)^{\frac{2d-1}{2d}} = o(N^+(\epsilon, T)).$$

由于 D 是固定常数且 $N^+(\epsilon/(2D+1), T) = \Theta(N^+(\epsilon, T))$ ，用 $\epsilon/(2D+1)$ 替换 ϵ 即得证。对于 $N^-(\epsilon, P_{\Omega_1} T P_{\Omega_2} + P_{\Omega_2} T P_{\Omega_1})$ 的情况，证明类似。 ■

如下是一个辅助引理，用于估计多个算子之和的特征值个数。

引理 2.6： 令 $A_1, \dots, A_k$ 为自伴紧算子。设 $\epsilon = \sum_{i=1}^k \epsilon_i$ 。记 $N^\pm(\epsilon, T)$ 为算子 T 中严格大于（小于） ϵ （ $-\epsilon$ ）的特征值个数。则有

$$N^\pm(\epsilon, \sum_{i=1}^k A_i) \leq \sum_{i=1}^k N^\pm(\epsilon_i, A_i), \quad (2.44)$$

证明 见 Widom^[85] 的引理 5。 ■

2.3.5 定理 2.3 的证明

以下是引理 2.1 的一个直接推论。在此列出是因为后续将频繁使用。

推论 2.2： 假设 $\Omega_1 \subseteq \Omega_2$ 。则 $N^+(\epsilon, P_{\Omega_1} T P_{\Omega_1}) \leq N^+(\epsilon, P_{\Omega_2} T P_{\Omega_2})$ 。

首先证明以下关于将区域分割为等距子域的引理，这将在后续证明中递归使

用。

引理 2.7: 设 T 与引理 2.5 中相同。设 $S \subseteq \mathbb{S}^d$, 假设当 $\epsilon \rightarrow 0$ 时 $N^+(\epsilon, P_S T P_S) \asymp N^+(\epsilon, T)$ 。进一步假设 $\Omega \subseteq \mathbb{S}^d$ 是具有分段光滑边界的子区域, 且存在 Ω 的 m 个等距副本 $\Omega_1, \dots, \Omega_m$, 使得它们的不交并 (相差一个零测集) 即为 S 。那么, 存在常数 $c > 0$, 使得对于较小的 ϵ , 有

$$cN^+(\epsilon, T) \leq N^+(\epsilon, P_\Omega T P_\Omega) \leq N^+(\epsilon, T). \quad (2.45)$$

证明 上界由推论 2.2 直接得出。现在考虑下界。由于 $\Omega_1, \dots, \Omega_m$ 构成了 S 的一个不交覆盖, 有:

$$P_S T P_S = \left(\sum_{i=1}^m P_{\Omega_i} \right) T \left(\sum_{j=1}^m P_{\Omega_j} \right) = \sum_i P_{\Omega_i} T P_{\Omega_i} + \sum_{i < j} \left(P_{\Omega_i} T P_{\Omega_j} + P_{\Omega_j} T P_{\Omega_i} \right).$$

利用引理 2.6, 可得

$$N^+(2\epsilon, T) \leq N^+(\epsilon, \sum_i P_{\Omega_i} T P_{\Omega_i}) + \sum_{i < j} N^+ \left(\frac{1}{C_m^2} \epsilon, P_{\Omega_i} T P_{\Omega_j} + P_{\Omega_j} T P_{\Omega_i} \right),$$

因此

$$N^+(\epsilon, \sum_i P_{\Omega_i} T P_{\Omega_i}) \geq N^+(2\epsilon, P_S T P_S) - \sum_{i < j} N^+ \left(\frac{1}{C_m^2} \epsilon, P_{\Omega_i} T P_{\Omega_j} + P_{\Omega_j} T P_{\Omega_i} \right).$$

注意到 Ω_i 是不交的且与 Ω 等距, 对于左边有:

$$N^+(\epsilon, \sum_i P_{\Omega_i} T P_{\Omega_i}) = \sum_i N^+(\epsilon, P_{\Omega_i} T P_{\Omega_i}) = mN^+(\epsilon, P_\Omega T P_\Omega).$$

另一方面, 由引理 2.5,

$$N^+ \left(\frac{1}{C_m^2} \epsilon, P_{\Omega_i} T P_{\Omega_j} + P_{\Omega_j} T P_{\Omega_i} \right) = o \left(N^+ \left(\frac{1}{C_m^2} \epsilon, T \right) \right) = o(N^+(\epsilon, T)),$$

其中注意到由假设 2.1(a) 可知对于固定 $c > 0$, 有 $N^+(c\epsilon, T) \asymp N^+(\epsilon, T)$ 。代入这两个估计并利用 $N^+(\epsilon, T) \asymp N^+(\epsilon, P_S T P_S)$, 得到

$$\begin{aligned} N^+(\epsilon, P_\Omega T P_\Omega) &\geq \frac{1}{m} N^+(2\epsilon, P_S T P_S) - o(N^+(\epsilon, T)) \geq cN^+(2\epsilon, T) - o(N^+(\epsilon, T)) \\ &\geq cN^+(\epsilon, T) - o(N^+(\epsilon, T)), \end{aligned}$$

从而证得下界。 ■

经过上述准备, 可以证明定理 2.3:

定理 2.3 的证明 设 T 为与 k 关联的积分算子。首先考虑 $\rho = \mathbf{1}_S$ 为开集 S 的示性函数且 $\mu_n = \tilde{\mu}_n$ 的情形。由命题 2.2 知, 只需考虑 $P_S T P_S$ 。由于 $N^+(\epsilon, A)$ 的渐近行为唯一确定了 $\lambda_i(A)$, 只需证明 $N^+(\epsilon, P_S T P_S) \asymp N^+(\epsilon, T)$ 。

取命题 2.15 中给出的子区域序列 $U_0, V_0, U_1, V_1, \dots \subseteq \mathbb{S}^d$ ，并通过归纳法证明 $N^+(\epsilon, P_{U_i} T P_{U_i}) \asymp N^+(\epsilon, T)$ 。初始情况由 $U_0 = \mathbb{S}^d$ 显然成立。假设 $N^+(\epsilon, U_i) \asymp N^+(\epsilon, T)$ ，由引理 2.7 以及 U_i 是 V_i 的若干等距副本的不交并这一事实，得到 $N^+(\epsilon, P_{V_i} T P_{V_i}) \asymp N^+(\epsilon, T)$ 。此外，由于 $V_i \subseteq U_{i+1}$ ，再次由推论 2.2，有

$$N^+(\epsilon, P_{V_i} T P_{V_i}) \leq N^+(\epsilon, P_{U_{i+1}} T P_{U_{i+1}}) \leq N^+(\epsilon, T),$$

因此 $N^+(\epsilon, P_{U_{i+1}} T P_{U_{i+1}}) \asymp N^+(\epsilon, T)$ 。

至此证明了 $N^+(\epsilon, P_{U_i} T P_{U_i}) \asymp N^+(\epsilon, T)$ 。由于 S 是开集且 $\text{diam } U_i \rightarrow 0$ ，总能找到某个 $U_i \subseteq S$ ，因此由推论 2.2 可得 $N^+(\epsilon, P_S T P_S) \asymp N^+(\epsilon, T)$ 。

对于一般的 μ_n ，设 T_- 和 T_+ 分别为由序列 $c_1 \tilde{\mu}_n$ 和 $c_2 \tilde{\mu}_n$ 定义的积分算子（类似于 (2.29)）。则 $T_- \leq T \leq T_+$ ，从而 $P_\Omega T_- P_\Omega \leq P_\Omega T P_\Omega \leq P_\Omega T_+ P_\Omega$ ，这意味着

$$\lambda_i(P_S T_- P_S) \leq \lambda_i(P_S T P_S) \leq \lambda_i(P_S T_+ P_S),$$

结果可由前述情形直接得出。

最后，假设 ρ 是一个有界非零黎曼可积函数。上界由引理 2.1 及 ρ 的有界性证得。对于下界，断言存在一个开集 Ω ，使得在 Ω 上 $\rho(\mathbf{x})^2 \geq c > 0$ 。再次利用引理 2.1，得出结论：

$$\lambda_i(k; \mathbb{S}^d, \rho^2 d\sigma) \geq \lambda_i(k; \Omega, \rho^2 d\sigma) \geq c \lambda_i(k; \Omega, d\sigma) \asymp \lambda_i(k; \mathbb{S}^d, d\sigma).$$

现在证明该断言。由于 ρ 黎曼可积，其间断点集为零测集。如果 $\rho(\mathbf{x})$ 在所有连续点处均为 0，则 $\rho(\mathbf{x}) = 0$ 几乎处处成立，这与 ρ 非零的假设矛盾。因此，必然存在一个连续点 $\mathbf{x}_0$ 使得 $\rho(\mathbf{x}_0) > 0$ ，而 Ω 可取为 $\mathbf{x}_0$ 的一个小邻域。

2.3.6 关于假设 2.1 的讨论

在本小节中讨论假设 2.1 成立的一些充分条件。第一个命题展示了当 μ_n 由函数 f 给定时，差分与导数之间的基本关系，这是 (2.34) 的直接推论。

命题 2.12： 假设 $\mu_n = f(n)$ ，其中 $f(x)$ 为定义在 $\mathbb{R}_{\geq 0}$ 上的函数。则：

- 若 $(-1)^p f^{(p)}(x) \geq 0, \forall x \geq N_0$ ，则 $\Delta^p \mu_n \geq 0, \forall n \geq N_0$ 。
- 若 $(-1)^{p+1} f^{(p+1)}(x) \geq 0, \forall x \geq N_0$ ，则 $\Delta^p \mu_n \leq f^{(p)}(n), \forall n \geq N_0$ 。

下一个引理表明，界定 (2.21) 中的最高阶项是充分的。

引理 2.8： 如果对于递减序列 B_n ，不等式 $\binom{n+d}{d} \Delta^d \mu_n \leq B_n$ 对所有 $n \geq 0$ 成立。那么，不等式 $\binom{n+l}{l} \Delta^l \mu_n \leq \frac{d}{l} B_n$ 对所有 $1 \leq l \leq d$ 成立。因此，如果存在 $q \in \mathbb{N}_+$ 和 $D' > 0$ 使得 $B_{qn} \leq D' \mu_n$ ，则 (2.21) 成立。

证明 用归纳法证明。假设结论对 $l+1$ 成立，则

$$\begin{aligned} \binom{n+l}{l} \Delta^l \mu_n &= \binom{n+l}{l} \sum_{k \geq n} \Delta^{l+1} \mu_k \leq \binom{n+l}{l} \sum_{k \geq n} \frac{d}{l+1} B_k \binom{k+l+1}{l+1}^{-1} \\ &\leq \binom{n+l}{l} \frac{d}{l+1} B_n \sum_{k \geq n} \binom{k+l+1}{l+1}^{-1} \\ &= \binom{n+l}{l} \frac{d}{l+1} B_n \frac{n!(l+1)!}{l(n+l)!} = \frac{d}{l} B_n. \end{aligned}$$

■

结合前两个结果可得如下推论。

推论 2.3: 假设 $\mu_n = f(n)$ ，其中 $f(x)$ 为定义在 $\mathbb{R}_{\geq 0}$ 上的函数。那么，假设 2.1 中的 (b,c) 对 $n \geq N_0$ 成立的一个充分条件是：存在 $q \in \mathbb{N}_+$ 和 $D' > 0$ ，使得对于所有 $x \geq N_0$ ，有

$$(-1)^{d+1} f^{(d+1)}(x) \geq 0 \quad \text{且} \quad (-1)^d x^d f^{(d)}(qx) \leq D' f(x). \quad (2.46)$$

命题 2.13: 对于以下每种形式的 μ_n ，都存在一个序列 $(\mu_n)_{n \geq 0}$ 满足假设 2.1，且当 n 足够大时该形式成立。

- $\mu_n = c_0 n^{-\beta}$ ，其中 $c_0 > 0$ 且 $\beta > d$ ；
- $\mu_n = c_0 n^{-\beta} (\ln n)^p$ ，其中 $c_0 > 0, \beta > d, p \in \mathbb{R}$ ，或者 $\beta = d, p > 1$ 。

证明 条件 (a) 对于这些渐近速率显然满足。在 $n \geq N_0$ (N_0 足够大) 时利用推论 2.3 验证 (b,c)，并按照引理 2.4 对 μ_n 进行左外推，使得条件对所有 $n \geq 0$ 成立。

- 对于 $\mu_n = f(n) = c_0 n^{-\beta}$ ，有 $(-1)^p f^{(p)}(x) = c_0 (\beta)_p x^{-(\beta+p)}$ ，其中 $(\beta)_p = \beta(\beta+1) \cdots (\beta+p-1)$ ，因此取 $q = 1$ 和 $D' = (\beta)_p$ 时 (2.46) 成立。
- 对于 $\mu_n = f(n) = c_0 n^{-\beta} (\ln n)^p$ ，当 $x \rightarrow \infty$ 时有

$$(-1)^p f^{(p)}(x) \asymp c_0 (\beta)_p x^{-(\beta+p)} (\ln x)^p,$$

因此取 $q = 1$ 时 (2.46) 仍然成立。

■

2.3.7 球面子区域的分解

令 $d(x, y) = \arccos \langle x, y \rangle$ 为球面上的测地距离 (geodesic distance)。第一个命题处理两个子区域旋转后的“重叠面积”问题。

命题 2.14: 设 $\Omega_1, \Omega_2 \subset \mathbb{S}^d$ 为两个具有分段光滑边界的不交区域。固定两点 $e, y \in \mathbb{S}^d$ 。假设对于任意 $x \in \mathbb{S}^d$ ， $R_{e,x}$ 是一个使得 $R_{e,x} e = x$ 的等距变换。那么，存在常

数 M 使得

$$|\{x \in \Omega_1 : R_{e,x}y \in \Omega_2\}| \leq M d(y, e) = M \arccos \langle y, e \rangle. \quad (2.47)$$

证明 令 $r = d(y, e)$ 。由于 $R_{e,x}$ 是等距变换，有

$$r = d(y, e) = d(R_{e,x}y, R_{e,x}e) = d(R_{e,x}y, x).$$

因此，如果 $r < d(x, \partial\Omega_1)$ ，注意到 $x \in \Omega_1$ ，则有 $R_{e,x}y \in \overline{B}_x(r) \subset \Omega_1$ ，从而 $R_{e,x}y \notin \Omega_2$ 。所以，

$$\{x \in \Omega_1 : R_{e,x}y \in \Omega_2\} \subseteq \{x \in \Omega_1 : d(x, \partial\Omega_1) \leq r\}.$$

后者即为 Weyl^[87]中定义的 $\partial\Omega_1$ 的半径为 r 的管状邻域 (tube)。此外，由于 $\partial\Omega_1$ 是分段光滑的，Weyl^[87]的结果表明存在常数 M ，使得

$$|\{x \in \Omega_1 : d(x, \partial\Omega_1) \leq r\}| \leq Mr,$$

从而得到所需的估计。 ■

接下来的命题给出了球面的一个分解。

命题 2.15: 存在一系列具有分段光滑边界的子区域 $U_0, V_0, U_1, V_1, \dots \subseteq \mathbb{S}^d$ ，满足：

- $U_0 = \mathbb{S}^d$;
- 存在 V_i 的不交等距副本 $V_{i,1}, \dots, V_{i,n_i}$ ，使得 $U_i = \bigcup_{j=1}^{n_i} V_{i,j} \cup Z$ ，其中 Z 为零测集；
- 经过某种等距变换后， $V_i \subseteq U_{i+1}$ ；
- $\text{diam } V_i \rightarrow 0$ 。

证明 记 $S_{p,r} = \{x \in \mathbb{S}^d \mid \langle x, p \rangle > \cos r\}$ 为以 p 为中心、半径为 r 的球冠 (spherical cap)，并记 $S_r = S_{e_{d+1},r}$ ，其中 e_{d+1} 为最后一个坐标方向的单位向量。

首先，令 $V_0 = \{x = (x_1, \dots, x_{d+1}) \in \mathbb{S}^d \mid x_i > 0, i = 1, \dots, d+1\}$ 。通过反射变换，存在 2^{d+1} 个 V_0 的等距副本，它们的不交并构成了整个球面减去赤道部分，后者为零测集。

接着，取 $p = \frac{1}{\sqrt{d+1}}(1, \dots, 1)$ ，对于任意点 $x \in V_0$ ，有

$$\langle x, p \rangle = \frac{1}{\sqrt{d+1}}(x_1 + \dots + x_{d+1}) \geq \frac{1}{\sqrt{d+1}}.$$

因此，对于 $r_1 = \arccos \frac{1}{\sqrt{d+1}} < \frac{\pi}{2}$ ，有 $V_0 \subset S_{p,r_1}$ ，可以取 $U_1 = S_{r_1}$ 。

现在假设已有 $U_i = S_{r_i}$ 且 $r_i < \frac{\pi}{2}$ 。利用极坐标，有参数化表示：

$$x_1 = \sin \theta_1 \cdots \sin \theta_d, \quad x_2 = \sin \theta_1 \cdots \cos \theta_d, \quad \dots, \quad x_d = \sin \theta_1 \cos \theta_2, \quad x_{d+1} = \cos \theta_1, \quad (2.48)$$

其中 $\theta_d \in [0, 2\pi]$ 且 $\theta_j \in [0, \pi]$ ($j = 1, \dots, d-1$)。此时球冠由 $S_r = \{x \mid \theta_1 < r\}$ 给定。考虑切片区域

$$V_i = \left\{ x \in \mathbb{S}^d \mid \theta_1 < r, \theta_j \in (0, \frac{\pi}{2}), j = 2, \dots, d-1, \theta_d \in \left(-\frac{\pi}{4}, \frac{\pi}{4}\right) \right\} \subset S_r.$$

通过关于 θ_d 的旋转和关于 θ_j ($j = 2, \dots, d-1$) 的反射, 可以找到 2^d 个 V_i 的等距副本, 它们的不交并与 S_{r_i} 仅相差 V_i 的边界并集, 这又是一个零测集。

现在, 要寻找 $r_{i+1} < r_i$ 使得 $V_i \subset U_{i+1} = S_{p, r_{i+1}}$ 。取点 $p = (p_1, \dots, p_{d+1})$ 如下:

$$p_{d+1} = \cos \eta, \quad p_d = \dots = p_2 = \frac{1}{\sqrt{d-1}} \sin \eta, \quad p_1 = 0,$$

其中 $\eta \in (0, r_i)$ 稍后确定。假设 $x \in V_i$ 由 (2.48) 给定, 则有 $\langle p, x \rangle = \cos \eta \cos \theta_1 + \frac{\sin \eta}{\sqrt{d-1}} (x_d + \dots + x_2)$ 。注意到 $\theta_j \in [0, \frac{\pi}{2}]$ 且 $|\theta_d| \leq \frac{\pi}{4}$, 有 $x_d + \dots + x_2 \geq \sin \theta_1 \cos \theta_d \geq \frac{1}{\sqrt{2}} \sin \theta_1$ 。因此,

$$\begin{aligned} \langle p, x \rangle &\geq \cos \eta \cos \theta_1 + a \sin \eta \sin \theta_1, \quad a = \frac{1}{\sqrt{2(d-1)}}, \\ &\geq \min(\cos \eta, \cos \eta \cos r_i + a \sin \eta \sin r_i), \quad \text{因为 } \theta_1 \in (0, r_i), \\ &= \begin{cases} \cos \eta, & \tan \eta > \frac{1 - \cos r_i}{a \sin r_i}, \\ \cos r_i \cos \eta + (a \sin r_i) \sin \eta, & \text{其他.} \end{cases} \end{aligned}$$

可知第二项在 $\tan \eta_0 = a \tan r_i$ 时取最大值, 且

$$\cos r_i \cos \eta_0 + a \sin r_i \sin \eta_0 = \sqrt{\cos^2 r_i + a^2 \sin^2 r_i}.$$

一方面, 如果 $\tan \eta_0 \leq \frac{1 - \cos r_i}{a \sin r_i}$, 取 $\eta = \eta_0$, 此时最小值由第二项取得, 故 $\langle p, x \rangle \geq \sqrt{\cos^2 r_i + a^2 \sin^2 r_i}$, 并且对于 $r_{i+1} = \arccos \sqrt{\cos^2 r_i + a^2 \sin^2 r_i}$ 有 $V_i \subset S_{p, r_{i+1}}$ 。此时, 有

$$\sin^2 r_{i+1} = 1 - (\cos^2 r_i + a^2 \sin^2 r_i) = (1 - a^2) \sin^2 r_i. \quad (2.49)$$

另一方面, 如果 $\tan \eta_0 = a \tan r_i > \frac{1 - \cos r_i}{a \sin r_i}$, 取 $\eta = \arctan \frac{1 - \cos r_i}{a \sin r_i}$ 。此时最小值由第一项取得, 且 $\langle p, x \rangle \geq \cos \eta$, 这意味着对于 $r_{i+1} = \eta$ 有 $V_i \subset S_{p, r_{i+1}}$ 。此时, 有

$$\tan r_{i+1} = \tan \eta = \frac{1 - \cos r_i}{a \sin r_i} < a \tan r_i. \quad (2.50)$$

综合考虑 (2.49) 和 (2.50) 两种情况, 并注意到 $a = \frac{1}{\sqrt{2(d-1)}} \in (0, 1)$, 可以得出结论: $r_{i+1} < r_i$ 且 $r_i \rightarrow 0$ 。 ■

2.3.8 紧自伴算子特征值的估计

对于一个希尔伯特空间上的紧自伴正算子 A ，用 $\lambda_n(A)$ 表示 A 的第 n 大（记重数）的特征值。如下的极小极大原理是泛函分析中的经典结果^[86]。

引理 2.9 (极小极大原理): 设 A 是希尔伯特空间上的紧自伴正算子。则

$$\lambda_n(A) = \inf_{\substack{V \subseteq H \\ \dim V = n-1}} \sup_{\substack{x \in V^\perp \\ \|x\|=1}} \langle Ax, x \rangle.$$

不仅如此，还有如下的 Weyl 不等式，用于估计多个算子之和的特征值。

引理 2.10 (Weyl 不等式): 令 A, B 为紧自伴正算子。则

$$\lambda_{i+j-1}(A+B) \leq \lambda_i(A) + \lambda_j(B), \quad i, j \geq 1. \quad (2.51)$$

第3章 核回归的泛化误差分析

本章从经典核回归框架出发，系统研究谱算法的泛化误差。核心目标是建立核方法与序列模型之间的精确联系，并据此给出谱算法泛化误差的统一刻画。在这一框架下，极小极大速率、饱和效应以及超有效性等现象都可以作为直接推论得到，从而较为完整地展示核方法在非参数回归中的优势与局限。

3.1 核方法简介

本章假设输入空间为 $\mathcal{X} \subset \mathbb{R}^d$ ，输出空间为 $\mathcal{Y} \subseteq \mathbb{R}$ ，未知的 $\mathcal{X} \times \mathcal{Y}$ 上的联合分布 ρ 在 $\mathcal{X}$ 上的边缘分布为 μ ，假设 μ 的支撑集为整个 $\mathcal{X}$ 。

3.1.1 正定核与再生核希尔伯特空间

首先引入正定核函数与再生核希尔伯特空间的概念，更加详细的介绍可以参考 Steinwart et al.^[13] 和 Wendland^[19]。

定义 3.1 (核函数): 一个二元函数 $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ 称为 (复值) 正定核函数，如果：

- 对任意的 $x, x' \in \mathcal{X}$ ，都有 $k(x, x') = \overline{k(x', x)}$ ；
- 对于任意的 $n \in \mathbb{N}$ ，任意的 $\{x_i\}_{i=1}^n \subseteq \mathcal{X}$ 以及任意的 $\{c_i\}_{i=1}^n \subseteq \mathbb{C}$ ，都有

$$\sum_{i=1}^n \sum_{j=1}^n \overline{c_i} c_j k(x_i, x_j) \geq 0, \quad (3.1)$$

或者等价地，矩阵 $(k(x_i, x_j))_{i,j=1}^n$ 是半正定的。

定义 3.2 (再生核希尔伯特空间): 给定一个正定核函数 $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ ，存在唯一的 (复值) 希尔伯特空间 $\mathcal{H}_k$ ，其元素为从 $\mathcal{X}$ 到 $\mathbb{C}$ 的函数，并且满足以下两个性质：

- 对于任意的 $x \in \mathcal{X}$ ，函数 $k(x, \cdot)$ 属于 $\mathcal{H}_k$ ；
- 对于任意的 $f \in \mathcal{H}_k$ 和任意的 $x \in \mathcal{X}$ ，都有再生性质

$$f(x) = \langle k(x, \cdot), f \rangle_{\mathcal{H}_k}. \quad (3.2)$$

该希尔伯特空间称为由核函数 k 所生成的再生核希尔伯特空间 (Reproducing Kernel Hilbert Space, RKHS)。

注 3.1 (实值核与复值核): 机器学习领域通常只考虑实值核函数，即 $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 。然而，为了后文技术处理中所涉及到的复分析工具，本文需要考虑更一般的复值核函数。实值核函数是复值核函数的一个特例，因此大多数实值核函数的性质和结果也适用于复值核函数。不仅如此，如果 k 是一个复值核函数，它所对应的再生

核希尔伯特空间为 $\mathcal{H}_k$ ，如果 k 实际上取值于实数域，即 $k(x, x') \in \mathbb{R}$ ，那么将 $\mathcal{H}_k$ 中的函数取实部所得到的

$$\mathcal{H}_{k,0} := \{\operatorname{Re} f \mid f \in \mathcal{H}_k\}, \quad \|g\|_{\mathcal{H}_{k,0}} := \inf_{f \in \mathcal{H}_k, \operatorname{Re} f = g} \|f\|_{\mathcal{H}_k}$$

构成一个实值再生核希尔伯特空间，其对应的核函数仍然是 k ，见 Steinwart et al.^[13] 的推论 4.22。

本文将引入关于核函数的如下记号。由于复值核函数涉及到共轭对称性，需要特别注意内积的线性与共轭线性性质，因此本文将一些常见的结论也一并列出。定义单点核函数 k_x 为

$$k_x := k(x, \cdot) \in \mathcal{H}_k,$$

于是它满足 $f(x) = \langle k_x, f \rangle_{\mathcal{H}_k}$ ，特别有

$$\langle k_{x'}, k_x \rangle_{\mathcal{H}_k} = k(x, x') = \overline{k(x', x)} = \overline{\langle k_x, k_{x'} \rangle_{\mathcal{H}_k}},$$

需要注意这里的前后顺序和共轭关系。本文定义核函数向量以及矩阵为

$$k(x, X) := (k(x, x_1), k(x, x_2), \dots, k(x, x_n))_{1 \times n}, \quad k(X, X) := (k(x_i, x_j))_{i,j=1}^n \in \mathbb{C}^{n \times n}.$$

此外，本文也定义归一化的核矩阵为

$$K := \frac{1}{n} k(X, X) \in \mathbb{C}^{n \times n}.$$

3.1.2 积分算子

为了辅助分析核方法的泛化误差，需要引入与核函数相关的积分算子。以下，本文假设核函数 k 连续，并且满足

$$\sup_{x \in \mathcal{X}} k(x, x) = \kappa^2 < \infty. \quad (3.3)$$

本文也引入核函数的 L^2 范数

$$\|k\|_{L^2}^2 := \int_{\mathcal{X}} k(x, x) d\mu(x) \leq \kappa^2 < \infty.$$

此时，再生核希尔伯特空间 $\mathcal{H}_k$ 中的函数均为 μ -平方可积的，由包含关系自然诱导出嵌入映射

$$S_k^* : \mathcal{H}_k \rightarrow L^2(\mathcal{X}, \mu).$$

它的伴随算子 S_k 满足

$$\langle f, S_k g \rangle_{\mathcal{H}_k} = \langle S_k^* f, g \rangle_{L^2}, \quad \forall f \in \mathcal{H}_k, g \in L^2.$$

特别地,

$$\begin{aligned} (S_k g)(x) &= \langle k_x, S_k g \rangle_{\mathcal{H}_k} = \langle S_k^* k_x, g \rangle_{L^2} \\ &= \int_{\mathcal{X}} \overline{k(x, x')} g(x') d\mu(x'), \quad \forall g \in L^2. \end{aligned} \quad (3.4)$$

由于 μ 的支撑集为 $\mathcal{X}$, 故 S_k^* 是单射。于是, 由伴随算子的性质, 有

$$\overline{\text{Ran } S_k} = (\text{Ker } S_k^*)^\perp = \mathcal{H}_k, \quad \overline{\text{Ran } S_k^*} = \text{Ker } S_k^\perp \subseteq L^2.$$

不仅如此, 由 (3.3), 还可以说明^[13] S_k^*, S_k 是希尔伯特-施密特算子, 并且它们的希尔伯特-施密特范数满足

$$\|S_k\|_2 = \|S_k^*\|_2 = \|k\|_{L^2} \leq \kappa. \quad (3.5)$$

现在, 定义积分算子以及协方差算子为

$$L_k = S_k^* S_k : L^2 \rightarrow L^2, \quad T_k = S_k S_k^* : \mathcal{H}_k \rightarrow \mathcal{H}_k.$$

它们的表达式如下:

$$(T_k f)(x) = \int_{\mathcal{X}} \overline{k(x, x')} f(x') d\mu(x'). \quad (3.6)$$

由 (3.5), 可以说明 L_k 和 T_k 是紧自伴正定算子, 并且它们的迹类范数满足

$$\text{Tr } L_k = \text{Tr } T_k = \|k\|_{L^2}^2 \leq \kappa^2.$$

由此, 算子的谱分解或奇异值分解得到

$$L_k = \sum_{m=1}^{\infty} \mu_m P_{V_m}, \quad (3.7)$$

其中 $(\mu_m)_{m=1}^{\infty}$ 是 L_k 的互异的非零特征值, 按照递减排列, $(V_m)_{m=1}^{\infty}$ 是对应的特征子空间, P_{V_m} 是 L^2 上投影到 V_m 的正交投影算子。这里, 不妨只考虑无穷维的情况, 因为有限维的情况更为简单。令 $d_m = \dim V_m$ 为特征值 μ_m 的重数。对于每个 V_m , 可以进一步选择一组标准正交基 $\{e_{m,l}\}_{l=1}^{d_m}$, 其中每个 $e_{m,l} \in \mathcal{H}_k$ 是 μ -等价类中的代表元。那么, $\{e_{m,l}\}$ 构成了 $\overline{\text{Ran } S_k^*} \subseteq L^2$ 的一组标准正交基, 且 $\{\mu_m^{1/2} e_{m,l}\}$ 构成了 $\overline{\text{Ran } S_k} = \mathcal{H}_k$ 的一组标准正交基。

不仅如此, 由 Mercer 定理^[13,83] 也可以得到核函数的展开:

$$k(x, x') = \sum_{m=1}^{\infty} \mu_m \sum_{l=1}^{d_m} \overline{e_{m,l}(x)} e_{m,l}(x'), \quad (3.8)$$

其中级数绝对且一致收敛。为了与现有文献保持一致, 用 $(\lambda_j)_{j \geq 1}$ 表示 T_k 计重数的非零特征值, 并略微滥用符号, 用 e_j 表示相应的特征函数, 注意这些特征函数的选取并不唯一。

注 3.2 (L_k 与 T_k): 此处澄清一下积分算子 L_k 和协方差算子 T_k 之间的关系, 以及特征函数的选取问题。首先, 虽然它们具有相同的计重数的特征值, 但是 L_k 是定义在 L^2 上的算子, 而 T_k 是定义在再生核希尔伯特空间 $\mathcal{H}_k$ 上的算子, 它们的特征函数所在的空间也不同。现在先用 $[e_j]_{\sim} \in L^2$ 表示 L_k 的特征函数, 它实际是函数的等价类, 并不属于 $\mathcal{H}_k$ 。虽然如此, 由于

$$\lambda_j [e_j]_{\sim} = L_k [e_j]_{\sim} = S_k^*(S_k [e_j]_{\sim}),$$

S_k^* 是自然嵌入映射, 以及 $S_k [e_j]_{\sim} \in \mathcal{H}_k$, 得到

$$\lambda_j e_j =_{\mu} S_k [e_j]_{\sim} \in \mathcal{H}_k,$$

其中 $=_{\mu}$ 表示 μ -几乎处处相等。由于 $\lambda_j \neq 0$, 所以可以取 $e_j = \lambda_j^{-1} S_k [e_j]_{\sim} \in \mathcal{H}_k$, 即选取 e_j 属于 $\mathcal{H}_k$, 那么上述等式就变成了严格相等。于是, 有

$$T_k e_j = S_k(S_k^* e_j) = S_k [e_j]_{\sim} = \lambda_j e_j,$$

这就说明了 e_j 也是 T_k 的特征函数。由此, 为了简单起见, 可以不再区分 L^2 中的等价类 $[e_j]_{\sim}$ 和它的代表元 e_j , 而 e_j 就成为了 L_k 和 T_k 的共同特征函数。不太严谨地, 有如下的展开表达:

$$L_k f = T_k f = \sum_{j=1}^{\infty} \lambda_j f_j e_j, \quad f = \sum_{j=1}^{\infty} f_j e_j,$$

读者还需要注意, $(e_j)_{j \geq 1}$ 在 L^2 中标准正交, 而 $(\lambda_j^{-1/2} e_j)_{j \geq 1}$ 才是在 $\mathcal{H}_k$ 中标准正交, 上述关于 T_k 的展开并不是 $\mathcal{H}_k$ 中的标准谱分解。

本文引入关于 L_k (或者 T_k) 特征值的如下假设, 这在现有文献中被广泛采用 [14, 35, 88]。

假设 3.1 (特征值衰减): 存在 $\beta > 1$ 以及常数 $c_{\text{eig}}, C_{\text{eig}} > 0$ 使得

$$c_{\text{eig}} j^{-\beta} \leq \lambda_j \leq C_{\text{eig}} j^{-\beta}, \quad j \geq 1, \quad (3.9)$$

或者等价地, 也可以由 μ_m 表示:

$$|\{i : \lambda_i \geq \lambda\}| = \sum_{m: \mu_m \geq \lambda} d_m = \Theta(\lambda^{-1/\beta}), \quad \lambda \rightarrow 0. \quad (3.10)$$

其中 $(\lambda_j)_{j \geq 1}$ 是积分算子 T 计及重数的特征值, 而 $\{\mu_m\}_{m \geq 1}$ 是 (3.8) 中定义的互异特征值。

注 3.3: (3.9) 与 (3.10) 的等价性是初等的, 其证明见命题 3.8。(3.10) 使本文能够处理一些不够显式的情况。例如, 若对于某些 $\gamma \geq 1$ 有 $\sum_{k=1}^m d_k \asymp m^{\gamma}$, 则 $\lambda_j \asymp j^{-\beta}$ 仅等价于 $\mu_m \asymp m^{-\gamma\beta}$ 。

注 3.4: 假设 3.1 中的特征值衰减率是现有文献中常见的假设^[14,35,88]。由于本章的目标是推导精确的泛化误差，因此 (3.9) 中的下界部分也是必要的。许多常用核函数都满足假设 3.1，例如拉普拉斯核、Matérn 核以及神经切线核 (NTK) 等，特别是本文在定理 2.2 中就证明了 NTK 满足该假设。这一关于特征值的假设刻画了再生核希尔伯特空间中函数的平滑度，较大的 β 意味着更好的平滑性。还注意到，由于 T 是迹类算子， $(\lambda_j)_{j \geq 1}$ 是可求和的，因此要求 $\beta > 1$ 是必要的。假设 3.1 与现有文献中再生核希尔伯特空间的有效维度 (effective dimension) 或容量条件 (capacity condition) 也有着密切联系^[14]。事实上，本章的结论也容易推广到更一般的特征值衰减情形，只要它们符合某种较弱的有效维度条件即可。但是为了简化讨论，本文在此仅考虑多项式衰减的情况。

3.1.3 插值空间

为了阐述本章的结果，需要进一步引入插值空间^[83]。对于 $p \geq 0$ ，通过下式定义分数次幂算子 $L_k^p : L^2 \rightarrow L^2$ ：

$$L_k^p = \sum_{m=1}^{\infty} \mu_m^p P_{V_m} = \sum_{m=1}^{\infty} \mu_m^p \sum_{l=1}^{d_m} \langle e_{m,l}, \cdot \rangle_{L^2} e_{m,l}. \quad (3.11)$$

随后，可以引入插值空间 $[\mathcal{H}_k]^s$ ：

$$[\mathcal{H}_k]^s = \text{Ran } L_k^{s/2} = \left\{ f = \sum_{m=1}^{\infty} \mu_m^{s/2} \sum_{l=1}^{d_m} a_{m,l} e_{m,l} \mid \sum_{m=1}^{\infty} \sum_{l=1}^{d_m} |a_{m,l}|^2 < \infty \right\} \subseteq L^2. \quad (3.12)$$

对于 $f, g \in [\mathcal{H}_k]^s$ ，其系数分别为 $(a_{m,l})$ 和 $(b_{m,l})$ ，在 $[\mathcal{H}_k]^s$ 中定义内积为

$$\langle f, g \rangle_{[\mathcal{H}_k]^s} = \sum_{m=1}^{\infty} \sum_{l=1}^{d_m} \overline{a_{m,l}} b_{m,l} = \left\langle T^{-\frac{s}{2}} f, T^{-\frac{s}{2}} g \right\rangle_{L^2}. \quad (3.13)$$

显而易见， $[\mathcal{H}_k]^s$ 是一个以 $\{\mu_m^{s/2} e_{m,l} : m \geq 1, 1 \leq l \leq d_m\}$ 为标准正交基的可分希尔伯特空间。特别地，有 $[\mathcal{H}_k]^0 \subseteq L^2$ 以及 $[\mathcal{H}_k]^1 = \mathcal{H}_k$ 。此外，当 $s \geq t$ 时存在自然包含关系 $[\mathcal{H}_k]^s \subseteq [\mathcal{H}_k]^t$ ，且当 $s > t$ 时该包含映射是紧的。 L_k （以及 L_k^p ）在 $[\mathcal{H}_k]^s$ 上的限制也是一个具有相同谱的有界算子，因此为简便起见，本文仍将其记为 L_k （及 L_k^p ）。事实上， $L_k^p : [\mathcal{H}_k]^s \rightarrow [\mathcal{H}_k]^{s+2p}$ 给出了一个同构映射。注意到 L_k 看作是 $\mathcal{H}_k = [\mathcal{H}_k]^1$ 上的算子时，恰好是协方差算子 T_k 。

3.1.4 正则再生核希尔伯特空间

为了推导尽可能紧致的学习率，需要最大程度地刻画再生核希尔伯特空间中函数的正则性。由于 V_m 是 L^2 及其再生核希尔伯特空间 $\mathcal{H}_k$ 的有限维子空间，它

是关于 $\langle \cdot, \cdot \rangle_{L^2}$ 的再生核希尔伯特空间，其再生核 k_m 由下式唯一确定：

$$\langle k_m(x, \cdot), f \rangle_{L^2} = \int_{\mathcal{X}} \overline{k_m(x, x')} f(x') d\mu(x') = f(x), \quad \forall f \in V_m. \quad (3.14)$$

选择上述的一组标准正交基 $\{e_{m,l}\}_{l=1}^{d_m}$ ，本文可以显式得到：

$$k_m(x, x') = \sum_{l=1}^{d_m} \overline{e_{m,l}(x)} e_{m,l}(x'), \quad (3.15)$$

该表达式不随基的选择而改变。容易看出：

$$k(x, x') = \sum_{m=1}^{\infty} \mu_m k_m(x, x').$$

本章引入如下正则再生核希尔伯特空间条件：

假设 3.2 (正则再生核希尔伯特空间)： 存在常数 $M > 0$ 使得

$$\sup_{x \in \mathcal{X}} \sum_{m=1}^N k_m(x, x) = \sup_{x \in \mathcal{X}} \sum_{m=1}^N \sum_{l=1}^{d_m} |e_{m,l}(x)|^2 \leq M \sum_{m=1}^N d_m, \quad \forall N \geq 1. \quad (3.16)$$

在此情况下，称该再生核希尔伯特空间（相对于分布 μ ）是正则的。

显而易见，若特征函数是一致有界的，即 $\sup_{x \in \mathcal{X}} |e_{m,l}(x)|^2 \leq M$ ，则该再生核希尔伯特空间是正则的。此外，还存在更多满足正则性的情况，因此本文认为这是一个相当广义的假设。

例 3.1 (平移不变周期核)： 设 $\mathcal{X} = \mathbb{T}^d = [-\pi, \pi]^d$ 为 d 维环面， μ 为 $\mathbb{T}^d$ 上的均匀测度。考虑满足 $k(x, y) = h(x - y)$ 的平移不变核，其中 h 是定义在 $\mathbb{T}^d$ 上的函数。此时容易证明傅里叶基 $\{\phi_m = e^{i\langle m, x \rangle}, m \in \mathbb{Z}^d\}$ 构成了 T 的一组标准正交特征函数。因此，由于该基是一致有界的，该空间是正则的。此外，若相应的特征值满足 $\lambda_m \asymp (1 + \|m\|_2^2)^{-\alpha}$ ，则相应的再生核希尔伯特空间为 $\mathcal{H}_k \cong H^\alpha(\mathbb{T}^d)$ （即 $\mathbb{T}^d$ 上的周期 Sobolev 空间），且 $[\mathcal{H}_k]^s \cong H^{s\alpha}(\mathbb{T}^d)$ 。

例 3.2 (球面上的内积核)： 设 $\mathcal{X} = \mathbb{S}^d$ 为 d 维球面， μ 为 $\mathbb{S}^d$ 上的均匀测度。考虑满足 $k(x, y) = h(\langle x, y \rangle)$ 的内积核，其中 h 是 $[-1, 1]$ 上的函数。此时，Funk-Hecke 公式^[29]表明，球谐函数 $\{Y_{m,l} \mid m \geq 1, 1 \leq l \leq d_m\}$ 构成了 T 的一组标准正交特征函数，其中 $\{Y_{m,l} \mid 1 \leq l \leq d_m\}$ 是 m 阶齐次调和多项式，且 $d_m = \binom{m+d}{m} - \binom{m-2+d}{m-2} \asymp m^{d-1}$ 。利用球谐函数理论可以证明该再生核希尔伯特空间是正则的，参见第 3.5.2.1 小节。

例 3.3 (单位球上的内积核)： 现在，考虑 d 维单位球 $\mathcal{X} = \mathbb{B}^d = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ ，其测度 μ 与经典权重 $p(x) = (1 - \|x\|^2)^{-1/2}$ 成正比。仍考虑内积核 k 。此时，单位球上 Funk-Hecke 公式的类比（见 Dai et al.^[29] 的章节 11）表明，次数恰好为 m 的正交多项式空间 V_m^d 是 T 对应于同一特征值的特征子空间，且 $\dim V_m^d = \binom{m+d-1}{m} \asymp m^{d-1}$ 。与球面情况类似，可以证明该再生核希尔伯特空间是正则的，参见第 3.5.2.2 小节。

在以往的文献中^[34-36]，引入了如下 L^∞ -嵌入性质来刻画再生核希尔伯特空间的正则性。称 $\mathcal{H}$ 具有 $\alpha \in (0, 1]$ 阶的嵌入性质，如果 $[\mathcal{H}]^\alpha$ 可以连续嵌入到 $L^\infty(\mathcal{X}, \mu)$ 中，即算子范数

$$M_\alpha := \|[\mathcal{H}]^\alpha \hookrightarrow L^\infty(\mathcal{X}, \mu)\| < \infty. \quad (3.17)$$

随后，定义嵌入指数为 $\alpha_0 = \inf \{\alpha : M_\alpha < \infty\}$ 。显然 $\alpha \leq 1$ ，因为

$$\sup_{x \in \mathcal{X}} |f(x)| = \sup_{x \in \mathcal{X}} \langle k(x, \cdot), f \rangle_{\mathcal{H}_k} \leq \|k(x, \cdot)\|_{\mathcal{H}_k} \|f\|_{\mathcal{H}_k} \leq \kappa \|f\|_{\mathcal{H}_k},$$

且 Fischer et al.^[35] 的引理 10 表明 $\alpha_0 \geq 1/\beta$ 。为了获得紧致的集中性不等式，先前的工作^[36,89] 假设 $\alpha_0 = 1/\beta$ 。这里本文证明，正则再生核希尔伯特空间条件满足这一嵌入指数条件。

命题 3.1: 在假设 3.1 和假设 3.2 下，嵌入指数为 $\alpha_0 = 1/\beta$ 。

此外，注意到嵌入指数条件仅对具有幂律衰减的特征值有意义，而正则再生核希尔伯特空间条件则可以考虑更一般的衰减情况。事实上，正则再生核希尔伯特空间条件本质上考察的是特征函数而非特征值。

3.1.5 谱算法

为了引入谱算法，首先引入一些记号。对于 $x \in \mathcal{X}$ ，定义算子 $S_{k,x}^* : \mathcal{H}_k \rightarrow \mathbb{R}$ 由 $S_{k,x}^* f = \langle k_x, f \rangle_{\mathcal{H}_k} = f(x)$ 给出，它的伴随 $S_{k,x} : \mathbb{R} \rightarrow \mathcal{H}_k$ 由 $S_{k,x} y = y k_x$ 给出。由于 $k(x, x) \leq \kappa^2$ ，故有 $\|S_{k,x}\| \leq \kappa$ 和 $\|S_{k,x}^*\| \leq \kappa$ 。可以将经验协方差算子表示为这些单点算子的平均，对比 (3.4) 和 (3.6)，有

$$L_k f = \int_{\mathcal{X}} S_{k,x} f(x) d\mu(x) = \mathbb{E}_{x \sim \mu} [S_{k,x} f(x)],$$

以及

$$T_k = \int_{\mathcal{X}} S_{k,x} S_{k,x}^* d\mu(x) = \mathbb{E}_{x \sim \mu} [S_{k,x} S_{k,x}^*]. \quad (3.18)$$

对于给定的训练数据，定义经验协方差算子

$$\hat{T}_k = \frac{1}{n} \sum_{i=1}^n S_{k,x_i} S_{k,x_i}^* : \mathcal{H}_k \rightarrow \mathcal{H}_k, \quad (3.19)$$

它可以看作是将 (3.18) 中的测度 μ 替换为经验测度 $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ 后得到的算子。显然 $\hat{T}_k$ 是一个正定自伴算子，维度至多为 n 。还定义样本基函数

$$\hat{g}_Z = \frac{1}{n} \sum_{i=1}^n S_{k,x_i} y_i = \frac{1}{n} \sum_{i=1}^n y_i k(x_i, \cdot) \in \mathcal{H}_k. \quad (3.20)$$

那么, 谱算法是通过滤谱函数对 $\hat{g}_Z$ 应用 $\hat{T}_k$ 的“正则化逆”而得到的^[39]。本文首先引入滤谱函数的如下定义。

定义 3.3 (滤谱函数): 设 $\{\varphi_\lambda : [0, \kappa^2] \rightarrow \mathbb{R}_{\geq 0} \mid \lambda \in (0, 1)\}$ 为一组以正则化参数 λ 为索引的函数族, 并定义剩余函数 (remainder function)

$$\psi_\lambda(z) := 1 - z\varphi_\lambda(z). \quad (3.21)$$

如果满足以下条件, 称 $\{\varphi_\lambda \mid \lambda \in (0, 1)\}$ (或简记为 $\varphi_\lambda(z)$) 为一个滤谱函数:

- (i) 对于任何固定的 λ , $\psi_\lambda(z) \geq 0$ 关于 $z \in [0, \kappa^2]$ 单调递减。对于任何固定的 z , $\psi_\lambda(z)$ 随着 λ 的减小而减小。
- (ii) 存在常数 E 使得

$$\sup_{z \in [0, \kappa^2]} (z + \lambda)\varphi_\lambda(z) \leq E, \quad \forall \lambda \in (0, 1). \quad (3.22)$$

- (iii) 该滤谱函数的“修饰阶” (qualification) 为 $\tau_{\max} \in [1, \infty]$, 使得对于所有 $0 \leq \tau \leq \tau_{\max}$ (且 $\tau < \infty$), 有

$$\sup_{z \in [0, \kappa^2]} z^\tau \psi_\lambda(z) \leq F_\tau \lambda^\tau, \quad \forall \lambda \in (0, 1), \quad (3.23)$$

其中 F_τ 是仅依赖于 τ 的常数。

- (iv) 此外, 若 $\tau_{\max}$ 是有限的, 则存在常数 $\underline{F}$ 使得

$$\psi_\lambda(z) \geq \underline{F} \lambda^{\tau_{\max}}, \quad \forall z \in [0, \kappa^2], \lambda \in (0, 1).$$

现给定一个滤谱函数 $\varphi_\lambda(z)$, 谱算法定义为

$$\hat{f}_\lambda = \varphi_\lambda(\hat{T}_k) \hat{g}_Z \in \mathcal{H}_k. \quad (3.24)$$

可以将上述算子在 $\mathcal{H}_k$ 的子空间

$$\mathcal{H}_n = \text{span} \left\{ k_{x_i} : 1 \leq i \leq n \right\}$$

上表达, 从而得到如下的矩阵形式:

$$\hat{f}_\lambda(x) = k(x, X) \varphi_\lambda(K) \mathbf{y},$$

其中 $\mathbf{y} = (y_1, \dots, y_n)^\top$ 。本文将在本小节末尾给出谱算法的具体示例。

注 3.5: 注意到, 性质 (i) 和 (iv) 在现有文献^[39-41]的滤谱函数定义中并非本质要求, 但本文引入它们是为了避免证明中一些不必要的技巧性问题。特别地, 性质 (i) 描述了滤谱方案的一般模式, 性质 (iv) 则精确决定了该滤谱函数的修饰阶。还注意到, 正则化参数 λ 将依赖于样本量 n , 且当 n 趋于无穷大时 $\lambda(n)$ 趋于零, 因此 $\lambda \in (0, 1)$ 的上界并非本质要求, 如有必要可以取更小的值。

然而, 为了推导精确的泛化误差曲线, 上述滤谱函数的定义对于本文的目的而

言尚不充分。本文技术上的关键创新在于开发了一种基于解析算子演算（analytic functional calculus）的特殊证明思路。为此，本文引入关于解析滤谱函数的如下假设。据作者所知，现有工作尚未系统考虑滤谱函数的这类性质。

假设 3.3 (解析滤谱函数): 设

$$D_\lambda = \{z \in \mathbb{C} : \operatorname{Re} z \in [-\lambda/2, \kappa^2], |\operatorname{Im} z| \leq \operatorname{Re} z + \lambda/2\} \\ \cup \{z \in \mathbb{C} : |z - \kappa^2| \leq \kappa^2 + \lambda/2, \operatorname{Re} z \geq \kappa^2\}.$$

滤谱函数 $\varphi_\lambda(z)$ 可以延拓为包含 D_λ 的某个域上的解析函数，且对于所有 $\lambda \in (0, 1)$ ，以下条件成立：

(C1) $|(z + \lambda)\varphi_\lambda(z)| \leq \tilde{E}, \forall z \in D_\lambda;$

(C2) $|(z + \lambda)\psi_\lambda(z)| \leq \tilde{F}\lambda, \forall z \in D_\lambda;$

其中 $\tilde{E}, \tilde{F}$ 是正的常数。

注 3.6: 区域 D_λ 至关重要，它与证明中的解析算子演算论证有关，其示意图见图 3.2。条件 (C1) 和 (C2) 分别可以看作是 (3.22) 和 (3.23) 在复数域上的延拓，因此可以预见它们对于滤谱函数也是成立的。事实上，本文将在下文中展示许多常用的滤谱函数都满足假设 3.3。证明内容被推迟至第 3.5.3 小节。

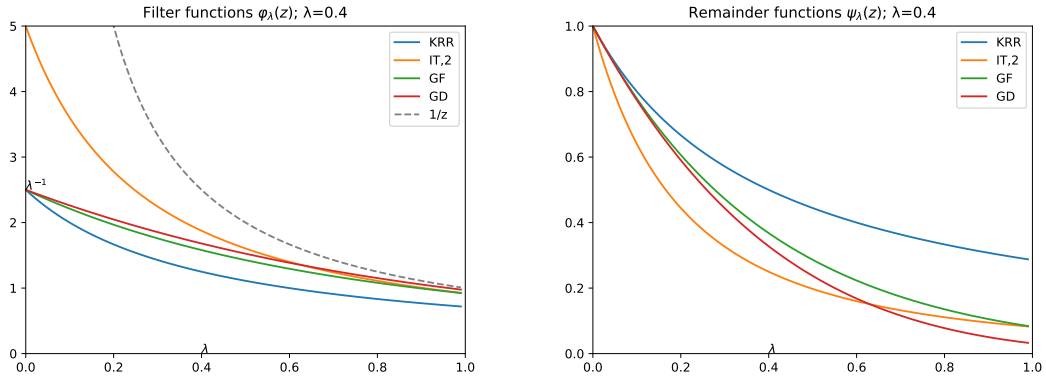


图 3.1 滤谱函数 φ_λ 和 ψ_λ 的示意图。

尽管上述假设 3.3 看起来是一个很强的假设，但如下所示的许多常用滤谱函数都满足该假设。

例 3.4 (核岭回归): 核岭回归 (KRR) 是由如下的带有 ℓ_2 正则化的经验风险最小化问题定义的：

$$\hat{f}_\lambda^{\text{Rid}} = \arg \min_{f \in \mathcal{H}_k} \left[\frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}_k}^2 \right],$$

其中 $\lambda > 0$ 为正则化参数。这是强凸的优化问题，存在唯一解，并且该解可以通过

闭式表达式给出。核岭回归的滤谱函数为

$$\varphi_\lambda^{\text{Rid}}(z) = \frac{1}{z + \lambda}, \quad \psi_\lambda^{\text{Rid}}(z) = \frac{\lambda}{z + \lambda}. \quad (3.25)$$

φ_λ 和 ψ_λ 在 $\{z \in \mathbb{C} : \operatorname{Re} z > -\lambda\} \supset D_\lambda$ 上均是解析的。该滤谱函数的修饰阶仅为 $\tau_{\max} = 1$ 。

例 3.5 (迭代岭回归): 为了克服核岭回归有限的修饰阶问题, Rosasco et al.^[41] 引入了如下迭代岭 (或迭代 Tikhonov) 方法。设 $p \geq 1$ 为定值。定义

$$\varphi_\lambda^{\text{IT},p}(z) = \frac{1}{z} \left[1 - \frac{\lambda^p}{(z + \lambda)^p} \right], \quad \psi_\lambda^{\text{IT},p}(z) = \frac{\lambda^p}{(z + \lambda)^p}. \quad (3.26)$$

易证 $z = 0$ 是 $\varphi_\lambda^{\text{IT},p}(z)$ 的可去奇点, 且 $\varphi_\lambda^{\text{IT},p}(z)$ 和 $\psi_\lambda^{\text{IT},p}(z)$ 在 $\{z \in \mathbb{C} : \operatorname{Re} z > -\lambda\} \supset D_\lambda$ 上均是解析的。该滤谱函数的优点在于其修饰阶为 $\tau_{\max} = p$ 。

为了理解“迭代”这一名称, 考虑 $p \in \mathbb{N}^*$ 的特殊情况。此时,

$$\varphi_\lambda^{\text{IT},p}(z) = \sum_{r=1}^p \lambda^{r-1} (z + \lambda)^{-r},$$

且该方法是通过岭回归方法进行 p 次迭代得到的:

$$(\hat{T}_k + \lambda) h_i = \hat{g}_Z + \lambda h_{i-1}, \quad i = 1, \dots, p.$$

例 3.6 (梯度流): 梯度流方法^[38]是另一种常用的正则化方法, 它对应于 (2.5) 中的神经切线核回归。考虑经验损失

$$\hat{\mathcal{L}}_n(f) = \frac{1}{2n} \sum_{i=1}^n (f(x_i) - y_i)^2 = \frac{1}{2n} \sum_{i=1}^n (S_{k,x_i}^* f - y_i)^2, \quad f \in \mathcal{H}_k.$$

那么, 可以考虑如下的 $\mathcal{H}_k$ 上的梯度流方程:

$$\dot{f}_t = -\nabla_f^{\mathcal{H}_k} \hat{\mathcal{L}}_n(f_t), \quad f_0 = 0, \quad (3.27)$$

其中 $\nabla_f^{\mathcal{H}_k}$ 表示在 $\mathcal{H}_k$ 中的梯度。具体地, 可以计算得到

$$\nabla_f^{\mathcal{H}_k} \hat{\mathcal{L}}_n(f) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i) k_{x_i} = \hat{T}_k f - \hat{g}_Z. \quad (3.28)$$

由此, 可以给出该梯度流方程的闭式解, 并给出滤谱函数

$$\varphi_\lambda^{\text{GF}}(z) = \frac{1 - e^{-tz}}{z}, \quad \psi_\lambda^{\text{GF}}(z) = e^{-tz}, \quad t = \lambda^{-1}. \quad (3.29)$$

同样容易证明 $z = 0$ 是 $\varphi_\lambda^{\text{GF}}(z)$ 的可去奇点, 因此 $\varphi_\lambda^{\text{GF}}(z)$ 和 $\psi_\lambda^{\text{GF}}(z)$ 在整个复平面上均是解析的。此外, 利用初等不等式可以证明, 梯度流方法的修饰阶为 $\tau_{\max} = \infty$, 且其常数为 $F_\tau = (\tau/e)^\tau$ 。

例 3.7 (梯度下降): 梯度下降方法是梯度流的离散版本。设 $\eta > 0$ 为固定的步长。

那么，对经验损失进行 t 步梯度下降迭代，即可产生滤谱函数

$$\begin{aligned}\varphi_\lambda^{\text{GD}}(z) &= \eta \sum_{k=0}^{t-1} (1 - \eta z)^k = \frac{1 - (1 - \eta z)^t}{z}, \quad \lambda = (\eta t)^{-1}, \\ \psi_\lambda^{\text{GD}}(z) &= (1 - \eta z)^t.\end{aligned}\tag{3.30}$$

此外，当 η 足够小时（例如 $\eta < 1/(2\kappa^2)$ ），对于 $z \in D_\lambda$ 有 $\text{Re}(1 - \eta z) > 0$ ，因此即使 t 不是整数，也可以取 $(1 - \eta z)^t$ 的单值分支。这样可以扩展滤谱函数的定义，使得 λ 可以取任意值，且 $t = (\eta\lambda)^{-1}$ 。同样容易证明 $z = 0$ 是 $\varphi_\lambda^{\text{GD}}(z)$ 的可去奇点。因此， $\varphi_\lambda^{\text{GD}}(z)$ 和 $\psi_\lambda^{\text{GD}}(z)$ 在 D_λ 上是解析的。与梯度流方法类似，梯度下降方法的修饰阶也为 $\tau_{\max} = \infty$ ，且 $F_\tau = (\tau/e)^\tau$ 。

3.2 谱算法的泛化误差与序列模型的等价性

3.2.1 序列模型中的收缩估计量

一种直观的理解谱算法的方法是考虑非参数回归所对应的高斯序列模型（Gaussian sequence model）^[40,90]，而将谱算法视为一种收缩估计量（shrinkage estimator）。回顾谱算法的估计量 $\hat{f}_\lambda = \varphi_\lambda(\hat{T}_k)\hat{g}_Z$ 。由 $\hat{T}_k$ 的表达式 (3.19)，可以将其视为对协方差算子 T 的经验估计；类似地， $\hat{g}_Z$ 也可以视为对 $g = S_k f^*$ 的经验估计。将这两个量替换成它们所对应的理想值后，得到如下的“理想”估计量：

$$\tilde{f}_\lambda^* = \varphi_\lambda(T)S_k f^*.$$

现在可以通过 (3.8) 所给出的特征函数基 $(e_j)_{j \geq 1}$ 展开该估计量。假设在这组基下，回归函数 f^* 具有展开

$$f^* = \sum_{j=1}^{\infty} \theta_j^* e_j,$$

其中 $\theta_j^* = \langle e_j, f^* \rangle_{L^2}$ 为展开系数。此时，这个估计量可以写成

$$\tilde{\theta}_j^* = \langle e_j, \tilde{f}_\lambda^* \rangle_{L^2} = [\varphi_\lambda(\lambda_j)\lambda_j] \theta_j^*, \quad j \geq 1.$$

这种形式的估计量被称为收缩估计量，因为它通过乘以因子 $\varphi_\lambda(\lambda_j)\lambda_j$ 来“收缩”原始系数 θ_j^* 。

虽然上述的理想化情况实际上忽视了噪声和有限样本效应，但是这样的序列形式展开为理解谱算法提供了有价值的视角。事实上，序列模型，特别是高斯序列模型（Gaussian sequence model），在非参数统计中被广泛用于分析各种估计量的性质^[90]。在高斯序列模型的框架下，观测到 n 个样本的非参数回归问题对应于如

下的高斯序列模型:

$$z_j = \theta_j^* + \sigma n^{-1/2} \xi_j, \quad j \geq 1, \quad (3.31)$$

其中无穷序列 $(z_j)_{j \geq 1}$ 为观测值, θ_j^* 是待估计的真实参数, $\sigma > 0$ 为噪声水平, 且 $\xi_j \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ 。本章将在第 4.3 节中给出两者之间的近似等价的更多解释。此时, 谱算法的估计量 $\hat{f}_\lambda$ 对应于如下的收缩估计量:

$$\hat{\theta}_\lambda = (\hat{\theta}_{\lambda,j})_{j \geq 1}, \quad \hat{\theta}_{\lambda,j} = [\varphi_\lambda(\lambda_j) \lambda_j] z_j, \quad j \geq 1. \quad (3.32)$$

对于序列模型的估计量 $\hat{\theta}_\lambda$, 其泛化误差定义为

$$\mathcal{R}_{\text{seq}}(\hat{\theta}_\lambda; \theta^*) = \mathbb{E} \left[\|\hat{\theta}_\lambda - \theta^*\|_2^2 \right] = \sum_{j=1}^{\infty} \mathbb{E} \left(\hat{\theta}_{\lambda,j} - \theta_j^* \right)^2. \quad (3.33)$$

事实上, 利用 (3.32) 和偏差-方差分解, 很容易计算这个收缩估计量的泛化误差, 它可以分解为偏差项和方差项之和: 偏差项为

$$\text{Bias}_{\text{seq}}^2(\varphi_\lambda; \theta^*) = \mathcal{R}_{\varphi}^2(\lambda; f^*) := \sum_{j=1}^{\infty} [\psi_\lambda(\lambda_j)]^2 (\theta_j^*)^2 = \|\psi_\lambda(T) f^*\|_{L^2}^2, \quad (3.34)$$

方差项为

$$\text{Var}_{\text{seq}}(\varphi_\lambda) = \frac{\sigma^2}{n} \sum_{j=1}^{\infty} [\lambda_j \varphi_\lambda(\lambda_j)]^2 = \frac{\sigma^2}{n} \mathcal{N}_{2,\varphi}(\lambda), \quad (3.35)$$

其中引入 $p \geq 1$ 阶的 φ -有效维度:

$$\mathcal{N}_{p,\varphi}(\lambda) := \sum_{j=1}^{\infty} [\lambda_j \varphi_\lambda(\lambda_j)]^p = \sum_{m=1}^{\infty} d_m [\mu_m \varphi_\lambda(\mu_m)]^p. \quad (3.36)$$

这个定义推广了经典的有效维度 (effective dimension) 概念^[14]。特别地, $p = 1$ 阶的 φ^{Rid} -有效维度即为前人工作中考虑的普通有效维度 (参见证明中的 (3.49))。将收缩估计量的泛化误差的显式表达式总结如下:

$$\mathcal{R}_{\text{seq}}(\hat{\theta}_\lambda; \theta^*) = \sum_{j=1}^{\infty} [\psi_\lambda(\lambda_j)]^2 (\theta_j^*)^2 + \frac{\sigma^2}{n} \sum_{j=1}^{\infty} [\lambda_j \varphi_\lambda(\lambda_j)]^2. \quad (3.37)$$

3.2.2 额外假设

在陈述主定理之前, 本章引入两个假设。第一个假设关于噪声。该假设非常标准, 若噪声独立于输入 x 且方差有界, 则该假设自然满足。

假设 3.4 (噪声): 假设

$$\mathbb{E}_{(x,y) \sim \rho} \left[(y - f^*(x))^2 \mid x \right] = \sigma^2 > 0, \quad \mu\text{-几乎处处 (a.e.) } x \in \mathcal{X}. \quad (3.38)$$

第二个假设关于回归函数 f^* 。回想插值空间的定义, 以及 $\{e_{m,l} : m \geq 1, 1 \leq$

$l \leq d_m$ 构成了 L^2 中的一组正交集。首先, 假设回归函数在 L^2 范数意义下具有如下展开:

$$f^* = \sum_{m=1}^{\infty} \sum_{l=1}^{d_m} f_{m,l} e_{m,l}. \quad (3.39)$$

这里注意到 $\tilde{f}_m^2 := \sum_{l=1}^{d_m} |f_{m,l}|^2 = \|P_{V_m} f^*\|_{L^2}^2$ 关于 V_m 的正交基 $\{e_{m,l}\}_{l=1}^{d_m}$ 的选择是不变的。随后, 假设回归函数满足如下源条件 (source condition)。

假设 3.5 (源条件): 存在 $s > 0$ 使得对于任何 $t < s$ 均有 $f^* \in [H_k]^t$, 且 $f^* \neq 0$; 若 $s < 2\tau_{\max}$, 则满足

$$\sum_{m: \mu_m < \lambda} \sum_{l=1}^{d_m} |f_{m,l}|^2 = \sum_{m: \mu_m < \lambda} \tilde{f}_m^2 = \Omega(\lambda^s). \quad (3.40)$$

该假设假定回归函数可以近似地由平滑指数为 s 的幂律衰减来刻画, 但不要要求 f^* 的系数严格按幂律衰减, 从而允许考虑更广泛的回归函数。注意到, 由于本章需要建立精确的泛化误差曲线, 因此以尾项和的形式呈现的下界也是必要的。下文给出了满足假设 3.5 的回归函数示例, 其证明推迟至第 3.5.4 小节。

例 3.8 (精确幂律 f^*): 设假设 3.1 成立, 且对于某些 $\gamma \geq 1$ 有 $\sum_{k=1}^m d_k \asymp m^\gamma$ 。假设 f^* 满足: 对于某些 $p > 1$, 有 $\tilde{f}_m \asymp m^{-(\gamma p + 1)/2}$ 。则 f^* 对于 $s = \frac{p}{\beta}$ 满足假设 3.5。特别地, 略微滥用符号, 若将 $(f_{m,l})$ 重新排列为

$$(f_j)_{j \geq 1} = (f_{1,1}, \dots, f_{1,d_1}, f_{2,1}, \dots, f_{m,1}, \dots, f_{m,d_m}, \dots).$$

那么, 该示例包含了 Cui et al.^[49], Li et al.^[88] 中考虑的 $|f_j| \asymp \lambda_j^{\frac{s}{2}} j^{-\frac{1}{2}} \asymp j^{-\frac{s\beta+1}{2}}$ 的情形。

还可以考虑如下系数存在“间隙”的情况。

例 3.9: 在本例中, 假设假设 3.1 成立, 并考虑前例中引入的 $(f_j)_{j \geq 1}$ 。假设对于某些 $q \geq 1, p > 1$, 有

$$\begin{cases} |f_{j(l)}| \asymp l^{-(p+1)/2}, & l = 1, 2, \dots, \\ f_j = 0, & \text{其他情形,} \end{cases}$$

其中 $j(l) \asymp l^q$ 。此时, f^* 对于 $s = p/(q\beta)$ 满足假设 3.5。

3.2.3 主定理

为了评估谱算法 $\hat{f}_\lambda$ 的性能, 考虑如下定义的泛化误差 (或超额风险)^[13]:

$$\|\hat{f}_\lambda - f^*\|_{L^2}^2 = \mathbb{E}_{x \sim \mu} (\hat{f}_\lambda(x) - f^*(x))^2.$$

此外, 考虑其关于 X 的条件期望更为方便, 定义

$$\mathcal{R}_{\text{Ker}}(\hat{f}_\lambda, f^*) := \mathbb{E} \left[\|\hat{f}_\lambda - f^*\|_{L^2}^2 \mid X \right]. \quad (3.41)$$

这仍是一个依赖于 X 的随机变量。若噪声 $\epsilon = y - f^*(x)$ 独立于 x , 则该条件期望仅是对噪声取期望。本章的主要结果将以高概率形式推导该量的精确估计。

现在, 可以陈述谱算法泛化误差的主定理。

定理 3.1: 在假设 3.1、3.2、3.3、3.4 和 3.5 下, 对于 $\lambda = \lambda(n) \rightarrow 0$, 有:

- 若存在 $\theta < \beta$ 使得 $\lambda = \Omega(n^{-\theta})$, 则

$$\mathcal{R}_{\text{Ker}}(\hat{f}_\lambda, f^*) = (1 + o_{\mathbb{P}}(1)) \mathcal{R}_{\text{Seq}}(\hat{\theta}_\lambda, \theta^*), \quad (3.42)$$

其中后者是序列模型中收缩估计量 $\hat{\theta}_\lambda$ 的泛化误差。

- 若 $\lambda = O(n^{-\beta})$, 则

$$\mathcal{R}_{\text{Ker}}(\hat{f}_\lambda, f^*) = \Omega_{\mathbb{P}}((\ln n)^{-4} \sigma^2). \quad (3.43)$$

3.3 结果的讨论与应用

本节从以下几个方面讨论本章的主要结果:

3.3.1 极小极大速率

定理 3.1 自然地导出了前人工作中推导出的谱算法的极小极大最优速率 (参见 Zhang et al.^[36], Lin et al.^[44] 及其参考文献)。进一步假设 $f^* \in [\mathcal{H}_k]^s$ 且 $s \leq 2\tau$ (如标准文献所示)。那么对于偏差项, 引理 3.3 表明:

$$\mathcal{R}_\varphi^2(\lambda; f^*) = \|\psi_\lambda(T)f^*\|_{L^2}^2 = O(\lambda^s).$$

对于方差项, 命题 3.9 表明:

$$\frac{\sigma^2}{n} \mathcal{N}_{2,\varphi}(\lambda) \asymp \frac{\sigma^2}{n} \lambda^{-1/\beta}.$$

因此, 选择 $\lambda \asymp n^{-\theta}$, 其中 $\theta = \frac{1}{s\beta+1}$ (如现有文献所示), 可得到最优速率的上界:

$$\mathbb{E} \left[\|\hat{f}_\lambda - f^*\|_{L^2}^2 \mid X \right] = O_{\mathbb{P}} \left(n^{-\frac{s\beta}{s\beta+1}} \right).$$

对于极小极大速率的下界, Caponnetto et al.^[14], Zhang et al.^[36], Lin et al.^[44] 已证明如下结果:

$$\liminf_{n \rightarrow \infty} \inf_{\hat{f}} \sup_{\substack{f^* \in [\mathcal{H}_k]^s \\ \|f^*\|_{[\mathcal{H}_k]^s} \leq R}} \mathbb{P} \left\{ \|\hat{f} - f^*\|_{L^2}^2 \geq cn^{-\frac{s\beta}{s\beta+1}} \right\} > 0. \quad (3.44)$$

3.3.2 精确的泛化误差曲线

本章的结果提供了谱算法泛化误差的完整图景，展示了正则化参数的选择、回归函数的源条件、噪声水平以及滤谱函数的选择所带来的影响。就正则化而言，随着正则化强度 λ 的减小，偏差减小而方差增加，表明谱算法也存在“偏差-方差权衡”（bias-variance trade-off），且学习曲线呈传统的 U 型。这也说明，为了达到最优速率，必须合理选择 λ 。

本章结果的主要优势在于，当 λ 处于合理范围内时，它为特定的谱算法提供了泛化误差的精确 $1 + o_{\mathbb{P}}(1)$ 形式。相比之下，之前关于极小极大最优速率的工作（例如 Caponnetto et al.^[14], Zhang et al.^[36], Lin et al.^[44]）通常仅能提供上界。虽然其结果也能部分反映偏差-方差权衡，但仅限于“上界形式”。更准确地说，他们证明了：

$$\text{Error}(\lambda) = \text{Bias}(\lambda) + \text{Var}(\lambda), \quad \text{且} \quad \text{Bias}(\lambda) \lesssim B(\lambda), \quad \text{Var}(\lambda) \lesssim V(\lambda),$$

然后通过平衡 $B(\lambda)$ 和 $V(\lambda)$ 来选择最优的 λ 。他们进一步验证，对于这样的 λ 选择，上界 $B(\lambda) + V(\lambda)$ 可以匹配极小极大下界。然而，由于 $B(\lambda)$ 和 $V(\lambda)$ 仅仅是上界，无法保证它们能反映精确的偏差-方差权衡。此外，近期关于谱算法的标准方法是应用所谓的“逼近-估计分解”（参见例如 Lin et al.^[44] 中定理 4.2 证明的开头）来分析泛化误差，这会直接丢失精确的偏差-方差权衡。

相比之下，本章的结果为任何满足 $\lambda \propto n^{-\theta}, \theta \in (0, \beta)$ 的 λ 建立了 $\text{Bias}(\lambda)$ 和 $\text{Var}(\lambda)$ 的 $1 + o_{\mathbb{P}}(1)$ 形式的精确逼近，在不丢失常数因子的前提下提供了精确的偏差-方差权衡。对于一般谱算法，据作者所知，本文首次给出了此类 $1 + o_{\mathbb{P}}(1)$ 形式的精确泛化误差曲线。

精确泛化误差的 $1 + o_{\mathbb{P}}(1)$ 形式意味着此结果可以给出泛化误差中的常数因子。它可以精确反映回归函数的量级和噪声如何影响泛化误差。此外，从先验（oracle）的角度来看，最小化 (3.42) 中的两项之和将产生 $\lambda = \lambda(n)$ 的最佳选择和最佳泛化误差，这已超越了单纯的渐近速率研究。

3.3.3 插值情形

本文将弱正则化（即 $\lambda = \lambda(n) = O(n^{-\beta})$ ）的情形称为“插值情形”（interpolating regime）。定理 3.1 表明，在该情形下，泛化误差的阶为 $\Omega_{\mathbb{P}}((\ln n)^{-4})$ ，这几乎是一个常数，说明估计器完全不具备泛化能力。

核方法在插值情形下的性能在以往文献中也有讨论。在受限设置下，多项工作^[45-47]展示了核最小范数插值器（即谱算法在 $\lambda \rightarrow 0^+$ 时的极限）的不一致性。与本文最相关的是，Li et al.^[89] 证明了对于满足嵌入指数条件的 Hölder 连续的核

函数相关的一般再生核希尔伯特空间，核回归在插值情形下的泛化误差对于任何 $\varepsilon > 0$ 均为 $\Omega_{\mathbb{P}}(n^{-\varepsilon})$ 。相比之下，本章去除了 Hölder 连续性的条件，并通过利用正则再生核希尔伯特空间条件改进分析，为插值情形下的谱算法提供了更好的下界 $\Omega_{\mathbb{P}}((\ln n)^{-4})$ 。本文认为，这一改进后的下界进一步证实了核方法在插值情形下不具备泛化能力，从而突显了正则化的必要性。

3.3.4 饱和效应

饱和效应是指对于某种修饰阶有限的谱算法，当回归函数的平滑度超过其修饰阶时，该算法无法达到极小极大最优收敛速率。由于传统文献仅提供泛化误差的上界，他们无法证明饱和效应。近期工作 Li et al.^[91] 严格证明了修饰阶 $\tau_{\max} = 1$ 的核岭回归的饱和效应，但对于具有更高（但有限）修饰阶的其他谱算法（如迭代岭正则化，见例子 3.5）尚无结果。

得益于精确的泛化误差曲线，本章的结果可以证明具有更高修饰阶的谱算法的饱和效应。考虑一个修饰阶有限（ $\tau_{\max} < \infty$ ）的谱算法，例如迭代岭正则化。那么，对于满足 $s \geq 2\tau_{\max}$ 且 $f^* \neq 0$ 的 $f^* \in [\mathcal{H}_k]^s$ ，显而易见：

$$\mathcal{R}_{\varphi}^2(\lambda; f^*) = \sum_{m=1}^{\infty} \psi_{\lambda}(\mu_m)^2 \tilde{f}_m^2 \geq \sum_{m=1}^{\infty} \underline{F} \lambda^{2\tau} \tilde{f}_m^2 = \underline{F} \|f^*\|_{L^2}^2 \lambda^{2\tau}.$$

因此，结合引理 3.3 中的上界，得出结论 $\mathcal{R}_{\varphi}^2(\lambda; f^*) \asymp \lambda^{2\tau}$ 。随后，当 λ 不太小时，主定理给出：

$$\begin{aligned} \mathbb{E} \left[\|\hat{f}_{\lambda} - f^*\|_{L^2}^2 \mid X \right] &= (1 + o_{\mathbb{P}}(1)) \left(\mathcal{R}_{\varphi}^2(\lambda; f^*) + \frac{\sigma^2}{n} \mathcal{N}_{2,\varphi}(\lambda) \right) \\ &= \Omega_{\mathbb{P}} \left(\lambda^{2\tau} + \frac{\sigma^2}{n} \lambda^{-1/\beta} \right) = \Omega_{\mathbb{P}} \left(n^{-\frac{2\tau\beta}{2\tau\beta+1}} \right). \end{aligned}$$

为了覆盖 λ 可能过小的情形，可以考虑 $\bar{\lambda} = \max(\lambda, n^{-\theta})$ （其中 $\theta < \beta$ ），并应用方差项的单调性（引理 3.5），如 Li et al.^[91] 所述。将其总结为以下推论。

推论 3.1 (饱和效应): 假设 3.1、3.2、3.3、3.4 成立。令 φ_{λ} 为修饰阶 $\tau = \tau_{\max} < \infty$ 的滤谱函数。那么，对于任何满足 $s \geq 2\tau$ 的非零 $f^* \in [\mathcal{H}_k]^s$ ，对于任何 $\lambda = \lambda(n) \rightarrow 0$ 的选择，有：

$$\mathbb{E} \left[\|\hat{f}_{\lambda} - f^*\|_{L^2}^2 \mid X \right] = \Omega_{\mathbb{P}} \left(n^{-\frac{2\tau\beta}{2\tau\beta+1}} \right). \quad (3.45)$$

此外，当下界在 $\lambda \asymp n^{-\theta}$ （其中 $\theta = \frac{\beta}{2\tau\beta+1}$ ）时达到。

3.3.5 与 Le Cam 等价性的比较

将非参数回归问题与高斯序列模型进行比较的想法也是传统文献中的一个重要主题，但此前主要以 Le Cam 等价性 (Le Cam equivalence) 的形式出现^[92-93]。这些工作旨在证明，在适当的正则性条件下，非参数回归问题与高斯序列模型在 Le Cam 距离意义下是等价的。Le Cam 距离衡量了两种统计实验之间的差异，定义为

$$\Delta(\mathcal{E}_{1,n}, \mathcal{E}_{2,n}; \Theta) = \sup_{\{i,j\}=\{1,2\}} \sup_{\hat{\theta}_{i,n}} \inf_{\hat{\theta}_{j,n}} \sup_{\theta^* \in \Theta} |\mathcal{R}_1(\hat{\theta}_{i,n}; \theta^*) - \mathcal{R}_2(\hat{\theta}_{j,n}; \theta^*)|,$$

其中 $\mathcal{E}_{i,n}$ 表示第 i 个统计实验， $\hat{\theta}_{i,n}$ 是在该实验下的估计量， $\mathcal{R}_i(\hat{\theta}_{i,n}; \theta^*)$ 是其风险。若 $\Delta(\mathcal{E}_{1,n}, \mathcal{E}_{2,n}; \Theta) \rightarrow 0$ ，则称这两个实验在参数空间 Θ 上是 Le Cam 等价的。然而，这样的 Le Cam 等价性结果虽然能说明上述风险差异趋于零，但往往无法说明该误差项趋于零的速率快于最优速率。此外，虽然在等价性的证明过程中可以给出非参数回归问题对应的高斯序列模型的特定估计量，但是对于谱算法，这些估计量的对应形式往往并不是前面讨论的收缩估计量。相比之下，本章的结果直接针对谱算法，提供了其泛化误差的精确估计，是比 Le Cam 等价性框架更加具体而精细的分析。

3.3.6 解析算子演算论证

作为本章的技术贡献之一，本章在谱算法的背景下开发了一种基于解析算子演算 (analytic functional calculus) 的“解析算子演算技巧”，这使本文能够推导出精确的泛化误差曲线。虽然算子演算在算子摄动理论中有着悠久的历史^[86,94]，但据作者所知，本文首次将其系统用于谱算法分析。首先，本文想说明这里的难点以及为什么现有技术不适用。

关于最优速率的传统文献仅关注上界，其基于“逼近-估计分解”的方法（例如 Zhang et al.^[36] 中的式 (88)）不适用于下界分析。此外，通常提供下界比提供上界更困难，因为前者要求误差项是无穷小量。对于核岭回归的简单情况，严格的工作^[88] 确定了渐近收敛速率，但其证明方法必须依赖于核岭回归的预解式恒等式 (resolvent identity)，即：

$$\varphi_\lambda^{\text{Rid}}(A) - \varphi_\lambda^{\text{Rid}}(B) = (A + \lambda)^{-1} - (B + \lambda)^{-1} = (A + \lambda)^{-1}(B - A)(B + \lambda)^{-1},$$

以及 $\psi_\lambda^{\text{Rid}}(A) = \lambda(A + \lambda)^{-1}$ 。这一恒等式对于将随机项（见 (3.48)）集中到 (3.42) 右侧出现的非随机对应项至关重要，期间会遇到诸如 $\varphi_\lambda(T) - \varphi_\lambda(\hat{T}_k)$ 和 $\psi_\lambda(T) - \psi_\lambda(\hat{T}_k)$ 之类的量。然而，对于一般的谱算法，不可能推导出类似的恒等式。此外，还必须考虑到 λ 的影响。

本章的“解析算子演算论证”利用解析算子演算克服了这些困难：

$$\begin{aligned}\varphi_\lambda(A) - \varphi_\lambda(B) &= -\frac{1}{2\pi i} \oint_{\Gamma} \varphi_\lambda(z) [R_A(z) - R_B(z)] dz \\ &= \frac{1}{2\pi i} \oint_{\Gamma} R_B(z)(A - B)R_A(z)\varphi_\lambda(z) dz,\end{aligned}$$

其中 $R_A(z) = (A - z)^{-1}$ 是 A 的预解式， Γ 是围道。此时，积分中的项与核岭回归的项相似（但注意此时 z 是复数）。令人惊讶的是，通过根据 λ 精心选择围道 Γ ，这一关键公式使本文能够应用关于预解式的集中结果并推导出非常精确的估计，从而得到精确的泛化误差曲线。本文相信，这种新颖的技术具有独立的理论价值，并可应用于其他问题。

3.4 主定理的证明

3.4.1 证明思路

现在给出 定理 3.1 的证明思路。第一步是传统的偏差-方差分解，这也是文献^[88]中的标准做法。首先定义一些基于样本点 X 的条件期望所得到的量：

$$\tilde{g}_X := \mathbb{E}(g_Z | X) = \frac{1}{n} \sum_{i=1}^n S_{k, x_i} f^*(x_i) \in \mathcal{H}_k, \quad (3.46)$$

$$\tilde{f}_\lambda := \mathbb{E}(\hat{f}_\lambda | X) = \varphi_\lambda(\hat{T}) \tilde{g}_X \in \mathcal{H}_k. \quad (3.47)$$

命题 3.2： 在假设 3.4 下，有

$$\begin{aligned}\mathbb{E} \left[\|\hat{f}_\lambda - f^*\|_{L^2}^2 \mid X \right] &= \|\tilde{f}_\lambda - f^*\|_{L^2}^2 + \frac{\sigma^2}{n^2} \sum_{i=1}^n \|\varphi_\lambda(\hat{T})k(x_i, \cdot)\|_{L^2}^2 \\ &=: \mathbf{Bias}^2(\lambda) + \mathbf{Var}(\lambda),\end{aligned} \quad (3.48)$$

其中注意到 $\mathbf{Bias}^2(\lambda)$ 和 $\mathbf{Var}(\lambda)$ 仍然是依赖于 X 的随机变量。

随后，将在定理 3.2 和定理 3.3 中分别证明，对于 $\lambda = \Omega(n^{-\theta})$ ， $\theta < \beta$ ，有：

$$\begin{aligned}\mathbf{Var}(\lambda) &= (1 + o_{\mathbb{P}}(1)) \frac{\sigma^2}{n} \mathcal{N}_{2, \varphi}(\lambda), \\ \mathbf{Bias}^2(\lambda) &= \mathcal{R}_\varphi^2(\lambda; f^*) + o_{\mathbb{P}}(1) \left(\mathcal{R}_\varphi^2(\lambda; f^*) + \frac{1}{n} \mathcal{N}_{2, \varphi}(\lambda) \right).\end{aligned}$$

此外，在推论 3.3 中，利用 $\mathbf{Var}(\lambda)$ 的单调性，还可以给出当 $\lambda = O(n^{-\beta})$ 时 $\mathbf{Var}(\lambda)$ 的一个下界。最后，将所有结论综合起来即可完成定理 3.1 的证明。

章节结构 在下文中，首先在命题 3.2 中给出偏差-方差分解的简单证明。在第 3.4.2 小节中，将推导后续使用的一些基本量的估计。在第 3.4.3 小节中，将使用集中不等式获得一些中间但关键量的高概率界。第 3.4.4 小节回顾解析算子演

算的基本事实，并定义证明中至关重要的围道。最后，在第 3.4.5 小节和第 3.4.6 小节中证明这两项的估计。

额外记号 在后续的证明中，将省略 L_k, T_k 和 $\hat{T}_k$ 的下标 k ，直接写作 L, T 和 $\hat{T}$ 。由于前面关于预备知识的讨论，也可以不显式区分积分算子 L 和协方差算子 T 。为了与前人文献^[95]保持一致，本章统一使用 T 来表示积分算子。本章将引入以下记号：

$$T_\lambda = T + \lambda I, \quad \hat{T}_\lambda = \hat{T} + \lambda I.$$

命题 3.2 的证明 设 $\epsilon_i = y_i - f^*(x_i)$ 为噪声。将 (3.20) 代入 (3.24)，可得

$$\begin{aligned} \hat{f}_\lambda &= \varphi_\lambda(\hat{T})\hat{g}_Z = \varphi_\lambda(\hat{T})\left(\frac{1}{n}\sum_{i=1}^n S_{k,x_i}(f^*(x_i) + \epsilon_i)\right) \\ &= \varphi_\lambda(\hat{T})\frac{1}{n}\sum_{i=1}^n S_{k,x_i}f^*(x_i) + \frac{1}{n}\sum_{i=1}^n \epsilon_i \varphi_\lambda(\hat{T})k_{x_i} \\ &= \tilde{f}_\lambda + \frac{1}{n}\sum_{i=1}^n \epsilon_i \varphi_\lambda(\hat{T})k_{x_i}, \end{aligned}$$

因此

$$\hat{f}_\lambda - f^* = \tilde{f}_\lambda - f^* + \frac{1}{n}\sum_{i=1}^n \epsilon_i \varphi_\lambda(\hat{T})k_{x_i}.$$

对 X 取条件期望，并利用 $\{\epsilon_i\}_{i=1}^n$ 在给定 X 条件下是独立同分布的零均值随机变量这一事实，以及假设 3.4，得到

$$\begin{aligned} \mathbb{E}\left[\|\hat{f}_\lambda - f^*\|_{L^2}^2 \mid X\right] &= \mathbb{E}\left[\langle \hat{f}_\lambda - f^*, \hat{f}_\lambda - f^* \rangle_{L^2} \mid X\right] \\ &= \|\tilde{f}_\lambda - f^*\|_{L^2}^2 + 2\mathbb{E}\left[\operatorname{Re}\left\langle \tilde{f}_\lambda - f^*, \frac{1}{n}\sum_{i=1}^n \epsilon_i \varphi_\lambda(\hat{T})k_{x_i} \right\rangle_{L^2} \mid X\right] \\ &\quad + \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i \varphi_\lambda(\hat{T})k_{x_i}\right\|_{L^2}^2 \mid X\right] \\ &= \|\tilde{f}_\lambda - f^*\|_{L^2}^2 + \frac{1}{n^2}\mathbb{E}\left[\left\langle \sum_{i=1}^n \epsilon_i \varphi_\lambda(\hat{T})k_{x_i}, \sum_{i=1}^n \epsilon_i \varphi_\lambda(\hat{T})k_{x_i} \right\rangle_{L^2} \mid X\right] \\ &= \|\tilde{f}_\lambda - f^*\|_{L^2}^2 + \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}(\epsilon_i^2 \mid X) \|\varphi_\lambda(\hat{T})k(x_i, \cdot)\|_{L^2}^2 \\ &= \|\tilde{f}_\lambda - f^*\|_{L^2}^2 + \frac{\sigma^2}{n^2}\sum_{i=1}^n \|\varphi_\lambda(\hat{T})k(x_i, \cdot)\|_{L^2}^2. \end{aligned}$$

■

3.4.2 基本量控制

记再生核希尔伯特空间 $\mathcal{H}_k$ 的 ($p \geq 1$ 阶) 有效维度为:

$$\mathcal{N}_p(\lambda) := \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\lambda + \lambda_i} \right)^p. \quad (3.49)$$

该量对应于之前在 (3.36) 中定义的 φ^{Rid} -有效维度。当 $p = 1$ 时, 它就是文献^[14]中常用的普通有效维度。

下面首先利用正则再生核希尔伯特空间条件给出正则化基函数的控制。

引理 3.1: 在假设 3.2 下, 对于任何 $x \in \mathcal{X}$,

$$\|\varphi_\lambda^{1/2}(T)k_x\|_{\mathcal{H}_k}^2 \leq M\mathcal{N}_{1,\varphi}(\lambda), \quad \|\varphi_\lambda(T)k_x\|_{L^2}^2 \leq M\mathcal{N}_{2,\varphi}(\lambda). \quad (3.50)$$

特别地, 对于 $\varphi_\lambda = \varphi_\lambda^{\text{Rid}}$, 有

$$\|T_\lambda^{-1/2}k_x\|_{\mathcal{H}_k}^2 \leq M\mathcal{N}_1(\lambda), \quad \|T_\lambda^{-1}k_x\|_{L^2}^2 \leq M\mathcal{N}_2(\lambda). \quad (3.51)$$

证明 利用特征值分解 (3.7) 以及 Mercer 分解 (3.8), 有

$$\begin{aligned} \|\varphi_\lambda^{1/2}(T)k_x\|_{\mathcal{H}_k}^2 &= \left\| \varphi_\lambda^{1/2}(T) \sum_{m=1}^{\infty} \mu_m \sum_{l=1}^{d_m} \overline{e_{m,l}(x)} e_{m,l} \right\|_{\mathcal{H}_k}^2 \\ &= \left\| \sum_{m=1}^{\infty} \varphi_\lambda^{1/2}(\mu_m) \mu_m \sum_{l=1}^{d_m} \overline{e_{m,l}(x)} e_{m,l} \right\|_{\mathcal{H}_k}^2 \\ &= \sum_{m=1}^{\infty} \mu_m \varphi_\lambda(\mu_m) \sum_{l=1}^{d_m} |e_{m,l}(x)|^2. \end{aligned}$$

注意到由于 μ_m 是递减的且 $t\varphi_\lambda(t)$ 随 t 减小而减小, 故 $a_m = \mu_m \varphi_\lambda(\mu_m)$ 是递减的。利用命题 3.10 和假设 3.2 可得

$$\|\varphi_\lambda^{1/2}(T)k_x\|_{\mathcal{H}_k}^2 \leq M \sum_{m=1}^{\infty} \mu_m \varphi_\lambda(\mu_m) d_m = M\mathcal{N}_{1,\varphi}(\lambda).$$

同理,

$$\|\varphi_\lambda(T)k_x\|_{L^2}^2 = \sum_{m=1}^{\infty} (\mu_m \varphi_\lambda(\mu_m))^2 \sum_{l=1}^{d_m} |e_{m,l}(x)|^2 \leq M\mathcal{N}_{2,\varphi}(\lambda).$$

■

在幂律衰减假设 3.1 下, 有有效维度的如下渐近性质, 证明参见附录中的命题 3.9。

引理 3.2: 设假设 3.1 成立, φ_λ 为满足 (3.22) 和 (3.23) 的滤谱函数。则对于任何

$p \geq 1$ 和 $\lambda > 0$, 有

$$\mathcal{N}_{p,\varphi}(\lambda) \asymp \lambda^{-1/\beta}. \quad (3.52)$$

特别地, 对于 $\varphi_\lambda = \varphi_\lambda^{\text{Rid}}$, 有 $\mathcal{N}_p(\lambda) \asymp \lambda^{-1/\beta}$.

以下引理控制了剩余项, 该引理将用于偏差项的证明。

引理 3.3: 假设 $f^* \in [\mathcal{H}_k]^t$. 令 φ_λ 为修饰阶为 τ 的滤谱函数, 且 $f_\lambda^* = \varphi_\lambda(T)Tf^*$. 则对于 $\gamma \in [0, t]$, 有

$$\|f^* - f_\lambda^*\|_{[\mathcal{H}_k]^\gamma} = \|\psi_\lambda(T)f^*\|_{[\mathcal{H}_k]^\gamma} \leq F_\theta \kappa^{(t-\gamma-2\tau)^+} \|f^*\|_{[\mathcal{H}_k]^t} \lambda^\theta, \quad (3.53)$$

其中 $\theta = \min\left(\frac{t-\gamma}{2}, \tau\right)$. 此外, 对于 $\gamma \in [0, 2+t]$, 有

$$\|f_\lambda^*\|_{[\mathcal{H}_k]^\gamma} \leq E \kappa^{(t-\gamma)^+} \|f^*\|_{[\mathcal{H}_k]^t} \lambda^{-\frac{(\gamma-t)^+}{2}}. \quad (3.54)$$

证明 由于 $f^* \in [\mathcal{H}_k]^t$, 可以找到 $h \in L^2$ 使得 $f^* = T^{t/2}h$ 且 $\|h\|_{L^2} = \|f^*\|_{[\mathcal{H}_k]^t}$. 那么,

$$\begin{aligned} \|\psi_\lambda(T)f^*\|_{[\mathcal{H}_k]^\gamma} &= \|T^{-\gamma/2}\psi_\lambda(T)T^{t/2}h\|_{L^2} \leq \|T^{\frac{t-\gamma}{2}}\psi_\lambda(T)\|_{\mathcal{B}(L^2)} \|h\|_{L^2} \\ &\leq \left\| T^{\frac{(t-\gamma-2\tau)^+}{2}} \right\| \|T^\theta\psi_\lambda(T)\|_{\mathcal{B}(L^2)} \|f^*\|_{[\mathcal{H}_k]^t} \\ &\leq F_\theta \kappa^{(t-\gamma-2\tau)^+} \|f^*\|_{[\mathcal{H}_k]^t} \lambda^\theta, \end{aligned}$$

其中 $\theta = \min\left(\frac{t-\gamma}{2}, \tau\right)$, 并且在最后一个不等式中使用了 (3.23). 对于第二个不等式, 有

$$\begin{aligned} \|f_\lambda^*\|_{[\mathcal{H}_k]^\gamma} &= \left\| T^{-\frac{\gamma}{2}} T \varphi_\lambda(T) T^{\frac{t}{2}} h \right\|_{L^2} \\ &\leq \left\| T^{\frac{2+t-\gamma}{2}} \varphi_\lambda(T) \right\|_{\mathcal{B}(L^2)} \|h\|_{L^2} \\ &\leq \left\| T^{\frac{(t-\gamma)^+}{2}} \right\| \left\| T^{1-\frac{(\gamma-t)^+}{2}} \varphi_\lambda(T) \right\|_{\mathcal{B}(L^2)} \|h\|_{L^2} \\ &\leq E \kappa^{(t-\gamma)^+} \|f^*\|_{[\mathcal{H}_k]^t} \lambda^{-\frac{(\gamma-t)^+}{2}}, \end{aligned}$$

其中最后一个不等式的第二项使用了 (3.90). ■

命题 3.3: 设 $f^* \in [\mathcal{H}_k]^0$. 假设 φ_λ 是在定义 3.3 中定义的修饰阶为 $\tau_{\max}$ 的滤谱函数. 则有

$$R_\varphi^2(\lambda; f^*) \geq \frac{1}{4} \sum_{m: \mu_m < \frac{\lambda}{2E}} \tilde{f}_m^2.$$

此外, 若 $\tau_{\max} < \infty$, 还有 $R_\varphi^2(\lambda; f^*) \geq \|f^*\|_{L^2}^2 E \lambda^{2\tau_{\max}}$.

因此, 在假设 3.5 下, 有

$$R_\varphi^2(\lambda; f^*) = \Omega(\lambda^{\min(s, 2\tau_{\max})}). \quad (3.55)$$

证明 回顾 (3.34) 中 $R_\varphi^2(\lambda; f^*)$ 的定义。对于第一个估计, 利用引理 3.10 中的最后一个控制, 当 $\mu_m \leq \frac{\lambda}{2E}$ 时, 有 $\psi_\lambda(\mu_m) \geq 1/2$, 故

$$R_\varphi^2(\lambda; f^*) = \sum_{m=1}^{\infty} \psi_\lambda(\mu_m)^2 \bar{f}_m^2 \geq \sum_{m: \mu_m < \frac{\lambda}{2E}} \psi_\lambda(\mu_m)^2 \bar{f}_m^2 \geq \frac{1}{4} \sum_{m: \mu_m < \frac{\lambda}{2E}} \bar{f}_m^2.$$

对于第二个估计, 利用定义 3.3 中的性质 (iv), 有

$$R_\varphi^2(\lambda; f^*) = \sum_{m=1}^{\infty} \psi_\lambda(\mu_m)^2 \bar{f}_m^2 \geq \sum_{m=1}^{\infty} \underline{F} \lambda^{2\tau_{\max}} \bar{f}_m^2 = \underline{F} \lambda^{2\tau_{\max}} \|f^*\|_{L^2}^2.$$

■

3.4.3 集中性质

利用正则再生核希尔伯特空间条件, 以下不等式是对现有文献 [89] Proposition 5.8 中集中结果的改进。主要的改进在于, 右侧出现的量 $\mathcal{N}_1(\lambda) \asymp \lambda^{-1/\beta}$ 严格小于其界限中出现的量 $M_\alpha \lambda^{-\alpha}$ ($\alpha > 1/\beta$), 而后者在 $\alpha \rightarrow 1/\beta$ 时是发散的。

命题 3.4: 在假设 3.2 下, 对于所有 $\delta \in (0, 1)$, 至少以概率 $1 - \delta$ 有:

$$\left\| T_\lambda^{-\frac{1}{2}} (T - \hat{T}) T_\lambda^{-\frac{1}{2}} \right\|_{\mathcal{B}(\mathcal{H}_k)} \leq \frac{2}{3} u + \sqrt{u}, \quad (3.56)$$

其中

$$u = u(n, \lambda) = \frac{2M\mathcal{N}_1(\lambda)}{n} \ln \frac{4(\|T\| + \lambda)\mathcal{N}_1(\lambda)}{\delta \|T\|}. \quad (3.57)$$

证明 将通过引理 3.13 进行证明。定义

$$A(x) = T_\lambda^{-\frac{1}{2}} (\hat{T} - T) T_\lambda^{-\frac{1}{2}}$$

以及 $A_i = A(x_i)$ 。那么容易看出 $\mathbb{E}(A_i) = 0$ 且

$$\frac{1}{n} \sum_{i=1}^n A_i = T_\lambda^{-\frac{1}{2}} (\hat{T} - T) T_\lambda^{-\frac{1}{2}},$$

这正是所关注的量。此外, 由于

$$T_\lambda^{-\frac{1}{2}} \hat{T} T_\lambda^{-\frac{1}{2}} = T_\lambda^{-\frac{1}{2}} S_{k,x} S_{k,x}^* T_\lambda^{-\frac{1}{2}} = T_\lambda^{-\frac{1}{2}} S_{k,x} \left[T_\lambda^{-\frac{1}{2}} S_{k,x} \right]^*,$$

根据 (3.51) 有

$$\left\| T_\lambda^{-\frac{1}{2}} \hat{T} T_\lambda^{-\frac{1}{2}} \right\|_{\mathcal{B}(\mathcal{H}_k)} = \left\| T_\lambda^{-\frac{1}{2}} S_{k,x} \right\|_{\mathcal{B}(\mathbb{R}, \mathcal{H}_k)}^2 = \left\| T_\lambda^{-\frac{1}{2}} k_x \right\|_{\mathcal{H}_k}^2 \leq M \mathcal{N}_1(\lambda). \quad (3.58)$$

通过取期望, 同样有 $\left\| T_\lambda^{-\frac{1}{2}} T T_\lambda^{-\frac{1}{2}} \right\|_{\mathcal{B}(\mathcal{H}_k)} \leq M \mathcal{N}_1(\lambda)$. 因此, 得到

$$\|A\|_{\mathcal{B}(\mathcal{H}_k)} \leq \left\| T_\lambda^{-\frac{1}{2}} T T_\lambda^{-\frac{1}{2}} \right\|_{\mathcal{B}(\mathcal{H}_k)} + \left\| T_\lambda^{-\frac{1}{2}} \hat{T} T_\lambda^{-\frac{1}{2}} \right\|_{\mathcal{B}(\mathcal{H}_k)} \leq 2M \mathcal{N}_1(\lambda) =: L.$$

对于条件的第二部分, 利用 $\mathbb{E}(B - \mathbb{E}(B))^2 \leq \mathbb{E}(B^2)$ 以及对于正自伴算子 B 有 $B^2 \leq \|B\| B$ 的事实 (其中 $\leq$ 表示由正算子诱导的偏序关系), 有

$$\mathbb{E}(A^2) \leq \mathbb{E} \left(T_\lambda^{-\frac{1}{2}} \hat{T} T_\lambda^{-\frac{1}{2}} \right)^2 \leq L \mathbb{E} \left(T_\lambda^{-\frac{1}{2}} \hat{T} T_\lambda^{-\frac{1}{2}} \right) = L T T_\lambda^{-1} =: V,$$

其中第二个 $\leq$ 来自 (3.58). 因此,

$$\|V\| = L \|T T_\lambda^{-1}\| = L \frac{\lambda_1}{\lambda + \lambda_1}, \quad \text{Tr } V = L \text{Tr} [T T_\lambda^{-1}] = L \mathcal{N}_1(\lambda),$$

$$B = \ln \frac{4 \text{Tr } V}{\delta \|V\|} = \ln \frac{4(\lambda_1 + \lambda) \mathcal{N}_1(\lambda)}{\delta \lambda_1}.$$

最后, 注意到引理中的量为:

$$\frac{2LB}{3n} = \frac{4M \mathcal{N}_1(\lambda) B}{3n} = \frac{2}{3} u, \quad \frac{2 \|V\| B}{n} \leq \frac{2M \mathcal{N}_1(\lambda) B}{n} = u.$$

■

下一个引理是命题 3.4 的推论。

引理 3.4: 假设假设 3.2 成立。固定 $\delta \in (0, 1)$ 。记

$$v = v(n, \lambda) = \frac{M \mathcal{N}_1(\lambda)}{n} \ln \frac{\mathcal{N}_1(\lambda)}{\delta}. \quad (3.59)$$

假设 $\lambda = \lambda(n) \rightarrow 0$ 满足 $v(n, \lambda) = o(1)$ 。那么当 n 足够大时, 至少以概率 $1 - \delta$ 有:

$$\|T_\lambda^{-1/2} (T - \hat{T}) T_\lambda^{-1/2}\| \leq C \sqrt{v}, \quad (3.60)$$

$$\begin{aligned} \|T_{X\lambda}^{-1/2} T_\lambda^{1/2}\|^2 &= \|T_\lambda^{1/2} T_{X\lambda}^{-1/2}\|^2 = \|T_\lambda^{1/2} T_{X\lambda}^{-1} T_\lambda^{1/2}\| \leq 3, \\ \|T_\lambda^{-1/2} T_{X\lambda}^{1/2}\|^2 &= \|T_{X\lambda}^{1/2} T_\lambda^{-1/2}\|^2 = \|T_\lambda^{-1/2} T_{X\lambda}^1 T_\lambda^{-1/2}\| \leq 2, \end{aligned} \quad (3.61)$$

其中 C 是一个绝对常数。

结合引理 3.2, 得到以下推论。

推论 3.2: 假设假设 3.1 和假设 3.2 成立。那么, 只要对于某些 $\theta < \beta$ 有 $\lambda(n) = \Omega(n^{-\theta})$, 对于固定的 $\delta \in (0, 1)$, 有 $v(n, \lambda) = o(1)$, 因此引理 3.4 中的结论成立。

引理 3.4 的证明 (3.60) 是命题 3.4 的直接推论, 其中 $v = o(1)$ 且

$$u = \frac{2M \mathcal{N}_1(\lambda)}{n} \ln \frac{4(\lambda_1 + \lambda) \mathcal{N}_1(\lambda)}{\delta \lambda_1} \leq C_0 v$$

对于某个绝对常数 C_0 成立。对于第二部分, 当 n 足够大使得 $u \leq 1/4$ 时,

$$\|T_\lambda^{-1/2} (T - \hat{T}) T_\lambda^{-1/2}\| \leq \frac{2}{3} u + \sqrt{u} \leq \frac{2}{3}.$$

注意到 $(T_{X\lambda}^{-1/2}T_\lambda^{1/2})^* = T_\lambda^{1/2}T_{X\lambda}^{-1/2}$ 以及 $\|A\|^2 = \|A^*\|^2 = \|A^*A\|$, 有

$$\begin{aligned}\|T_{X\lambda}^{-1/2}T_\lambda^{1/2}\|^2 &= \|T_\lambda^{1/2}T_{X\lambda}^{-1/2}\|^2 = \|T_\lambda^{1/2}(\hat{T} + \lambda)^{-1}T_\lambda^{1/2}\| \\ &= \|[T_\lambda^{-1/2}(\hat{T} + \lambda)T_\lambda^{-1/2}]^{-1}\| \\ &= \|[I - T_\lambda^{-1/2}(T - \hat{T})T_\lambda^{-1/2}]^{-1}\| \\ &\leq [1 - \|T_\lambda^{-1/2}(T - \hat{T})T_\lambda^{-1/2}\|]^{-1} \leq 3,\end{aligned}$$

其中最后一个不等式利用了 $\|(I - A)^{-1}\| \leq (1 - \|A\|)^{-1}$.

对于另一部分, 有

$$\begin{aligned}\|T_\lambda^{-1/2}T_{X\lambda}^{1/2}\|^2 &= \|T_{X\lambda}^{1/2}T_\lambda^{-1/2}\|^2 = \|T_\lambda^{-1/2}T_{X\lambda}T_\lambda^{-1/2}\| \\ &= \|I + T_\lambda^{-1/2}(\hat{T} - T)T_\lambda^{-1/2}\| \\ &\leq 1 + \|T_\lambda^{-1/2}(\hat{T} - T)T_\lambda^{-1/2}\| \leq 2.\end{aligned}$$

■

3.4.4 解析算子演算

基于解析算子的论证是本文的主要创新点之一。首先回顾一些关于解析算子演算的基本事实。数学细节可参考例如 Simon^[86]。

定义 3.4: 设 A 为 Banach 空间 X 上的线性算子。其“预解集” $\rho(A)$ 由下式给出:

$$\rho(A) := \{\lambda \in \mathbb{C} \mid A - \lambda \text{ 是可逆的}\},$$

记 $R_A(\lambda) := (A - \lambda)^{-1}$ 。 A 的谱定义为

$$\sigma(A) := \mathbb{C} \setminus \rho(A).$$

解析算子演算中的一个简单但关键的成分是如下“预解式恒等式”:

$$R_A(\lambda) - R_B(\lambda) = R_A(\lambda)(B - A)R_B(\lambda) = R_B(\lambda)(B - A)R_A(\lambda). \quad (3.62)$$

预解式能够模拟 Cauchy 积分公式的形式来定义 $f(A)$ 的值, 其中 A 是一个算子, f 是解析函数。这通常被称为解析算子演算, 参见例如 Simon^[86]Theorem 2.3.1。

命题 3.5 (解析算子演算): 设 A 为希尔伯特空间 H 上的算子, f 为定义在 $D_f \subset \mathbb{C}$ 上的解析函数。令 Γ 为包含在 D_f 中并围住 $\sigma(A)$ 的一条围道。则有

$$f(A) = \frac{1}{2\pi i} \oint_{\Gamma} f(z)(z - A)^{-1} dz = -\frac{1}{2\pi i} \oint_{\Gamma} f(z)R_A(z) dz, \quad (3.63)$$

且该结果与 Γ 的选择无关。

注 3.7: 对于自伴紧算子 A , 有谱分解

$$A = \sum_{i=1}^{\infty} \lambda_i \langle e_i, \cdot \rangle e_i,$$

通常定义 $f(A)$ 为

$$f(A) = \sum_{i=1}^{\infty} f(\lambda_i) \langle e_i, \cdot \rangle e_i. \quad (3.64)$$

事实上, 该定义与命题 3.5 中的定义是一致的。注意到 (3.64) 对于连续函数 f 同样有效, 且通过谱定理 (例如参考 Simon^[86] 的第 5 章节), 也可以将其扩展到自伴 (不一定是紧) 算子。

现在, 令 Γ 为包含在 D_f 中并同时围住 $\sigma(A)$ 和 $\sigma(B)$ 的一条围道。利用 (3.62), 得到

$$\begin{aligned} f(A) - f(B) &= -\frac{1}{2\pi i} \oint_{\Gamma} f(z) [R_A(z) - R_B(z)] dz \\ &= \frac{1}{2\pi i} \oint_{\Gamma} R_B(z)(A - B)R_A(z)f(z) dz. \end{aligned} \quad (3.65)$$

将使用如下谱映射定理来控制证明中的一些算子范数, 参见 Simon^[86] 的定理 5.1.11。

命题 3.6 (谱映射定理): 设 A 为有界自伴算子, f 为 $\sigma(A)$ 上的连续函数。则有

$$\sigma(f(A)) = \{f(\lambda) \mid \lambda \in \sigma(A)\}. \quad (3.66)$$

因此, $\|f(A)\| = \sup_{\lambda \in \sigma(A)} |f(\lambda)| \leq \|f\|_{\infty}$ 。

最后, 本章定义围道 Γ_{λ} :

$$\begin{aligned} \Gamma_{\lambda} &= \Gamma_{\lambda,1} \cup \Gamma_{\lambda,2} \cup \Gamma_{\lambda,3} \\ \Gamma_{\lambda,1} &= \{x \pm (x + \eta)i \in \mathbb{C} \mid x \in [-\eta, 0]\} \\ \Gamma_{\lambda,2} &= \{x \pm (x + \eta)i \in \mathbb{C} \mid x \in (0, \kappa^2)\} \\ \Gamma_{\lambda,3} &= \{z \in \mathbb{C} \mid |z - \kappa^2| = \kappa^2 + \eta, \operatorname{Re}(z) \geq \kappa^2\}, \end{aligned} \quad (3.67)$$

其中 $\eta = \lambda/2$, 参见图 3.2。由于 T 和 $\hat{T}$ 是满足 $\|T\|, \|\hat{T}\| \leq \kappa^2$ 的正自伴算子, 有 $\sigma(T), \sigma(\hat{T}) \subseteq [0, \kappa^2]$ 。因此, Γ_{λ} 确实是满足命题 3.5 要求的一条围道。

命题 3.7: 当 (3.61) 成立时, 存在一个绝对常数 C , 使得对于任何 $z \in \Gamma_{\lambda}$, 有:

$$\begin{aligned} \left\| T_{\lambda}^{\frac{1}{2}}(T - z)^{-1}T_{\lambda}^{\frac{1}{2}} \right\| &\leq C \\ \left\| T_{\lambda}^{\frac{1}{2}}(\hat{T} - z)^{-1}T_{\lambda}^{\frac{1}{2}} \right\| &\leq 3C. \end{aligned} \quad (3.68)$$

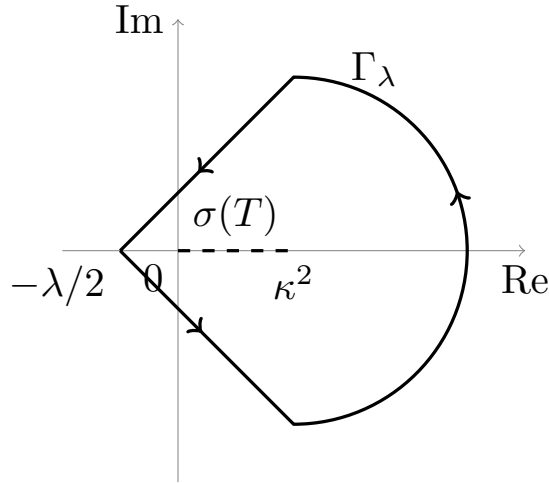


图 3.2 (3.67) 中定义的围道 Γ_λ 示意图。 Γ_λ 所围成的区域正是假设 3.3 中的 D_λ 。虚线区间 $[0, \kappa^2]$ 包含 T 和 $\hat{T}$ 的谱。

证明 利用谱映射定理, 对于谱满足 $\sigma(A) \subseteq [0, \kappa^2]$ 的自伴算子 A , 有

$$\left\| A_\lambda^{\frac{1}{2}} (A - z)^{-1} A_\lambda^{\frac{1}{2}} \right\| = \sup_{t \in \sigma(A)} \left| \frac{t + \lambda}{t - z} \right|.$$

现在, 当 $z = x + (x + \lambda/2)i \in \Gamma_{\lambda,1} \cup \Gamma_{\lambda,2}$ (其中 $x \in [-\lambda/2, \kappa^2]$) 时, 得到

$$\sup_{t \in \sigma(A)} \left| \frac{t + \lambda}{t - z} \right|^2 \leq \sup_{t \geq 0} \left| \frac{t + \lambda}{t - z} \right|^2 = \begin{cases} \frac{4\lambda^2}{\lambda^2 + 4\lambda x + 8x^2}, & -\frac{1}{2}\lambda \leq x \leq -\frac{1}{2(2+\sqrt{2})}\lambda, \\ \frac{5\lambda^2 + 12\lambda x + 8x^2}{(\lambda + 2x)^2}, & x \geq \frac{1}{2(2+\sqrt{2})}\lambda, \end{cases}$$

经过繁琐计算可知, 右侧在 $x = -\lambda/4$ 处达到最大值 8, 故有

$$\sup_{t \in \sigma(A)} \left| \frac{t + \lambda}{t - z} \right|^2 \leq 8, \quad z \in \Gamma_{\lambda,1} \cup \Gamma_{\lambda,2}.$$

当 $z \in \Gamma_{\lambda,3}$ 时, 对于 $t \in \sigma(A) \subseteq [0, \kappa^2]$ 有 $|t - z| \geq \kappa^2$, 故

$$\sup_{t \in \sigma(A)} \left| \frac{t + \lambda}{t - z} \right| \leq \sup_{t \in \sigma(A)} \left| \frac{t + \lambda}{\kappa^2} \right| \leq \frac{\lambda + \kappa^2}{\kappa^2} \leq 2.$$

综上所述, 存在绝对常数 C 使得

$$\left\| A_\lambda^{\frac{1}{2}} (A - z)^{-1} A_\lambda^{\frac{1}{2}} \right\| \leq C.$$

因此, 令 $A = T$ 即可得到第一个不等式。对于第二个不等式, 注意到

$$\begin{aligned} \left\| T_\lambda^{\frac{1}{2}} (\hat{T} - z)^{-1} T_\lambda^{\frac{1}{2}} \right\| &= \left\| T_\lambda^{\frac{1}{2}} T_{X\lambda}^{-\frac{1}{2}} \cdot T_{X\lambda}^{\frac{1}{2}} (\hat{T} - z)^{-1} T_{X\lambda}^{\frac{1}{2}} \cdot T_{X\lambda}^{-\frac{1}{2}} T_\lambda^{\frac{1}{2}} \right\| \\ &\leq \left\| T_\lambda^{\frac{1}{2}} T_{X\lambda}^{-\frac{1}{2}} \right\| \cdot \left\| T_{X\lambda}^{\frac{1}{2}} (\hat{T} - z)^{-1} T_{X\lambda}^{\frac{1}{2}} \right\| \cdot \left\| T_{X\lambda}^{-\frac{1}{2}} T_\lambda^{\frac{1}{2}} \right\| \\ &\leq 3C, \end{aligned}$$

其中利用了 (3.61) 以及针对 $A = \hat{T}$ 的范数界。 ■

3.4.5 方差项

以下定理极大地改进了 Li et al.^[89] 的定理 A.10 以及 Zhang et al.^[96] 中的结果。除了考虑更通用的滤谱算法这一主要区别外，它还：(1) 去除了 Li et al.^[89] 中对核函数 Hölder 连续性的要求；(2) 与 Zhang et al.^[96] 相比，给出了不损失常数因子的精确 $1 + o_{\mathbb{P}}(1)$ 形式；(3) 允许更广泛的 λ 范围，从而在推论 3.3 中导出对数下界。

定理 3.2: 在假设 3.1、3.2 和 3.3 下，假设 $\lambda = \lambda(n) \rightarrow 0$ 满足

$$\frac{\lambda^{-1/\beta}}{n} (\ln \lambda^{-1})^3 = o(1), \quad (3.69)$$

则有

$$\mathbf{Var}(\lambda) = [1 + o_{\mathbb{P}}(1)] \frac{\sigma^2}{n} \mathcal{N}_{2,\varphi}(\lambda). \quad (3.70)$$

特别地，满足 (3.69) 的充分条件是 $\lambda = \Omega(n^{-\theta})$ (对于某些 $\theta < \beta$) 或 $\lambda = \Omega(n^{-\beta} \ln^p n)$ (对于任何 $p > 3\beta$)。

证明 回顾:

$$\mathbf{Var}(\lambda) = \frac{\sigma^2}{n} \frac{1}{n} \sum_{i=1}^n \left\| \varphi_{\lambda}(\hat{T}) k_{x_i} \right\|_{L^2}^2.$$

引理 3.2 给出:

$$\mathcal{N}_1(\lambda) \asymp \mathcal{N}_2(\lambda) \asymp \mathcal{N}_{2,\varphi}(\lambda) \asymp \lambda^{-1/\beta}. \quad (3.71)$$

因此，只要 n 足够大，由于 $\lambda = \Omega(n^{-\theta})$ (其中 $\theta < \beta$)，引理 3.4 中的条件 (3.59) 即可满足。随后，应用引理 3.4、引理 3.7 和引理 3.6，当 n 足够大时，至少以概率 $1 - \delta$ 有:

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^n \left\| \varphi_{\lambda}(\hat{T}) k_{x_i} \right\|_{L^2}^2 - \int_{\mathcal{X}} \left\| \varphi_{\lambda}(T) k_x \right\|_{L^2}^2 d\mu(x) \right| \\ & \leq \frac{1}{n} \sum_{i=1}^n \left| \left\| \varphi_{\lambda}(\hat{T}) k_{x_i} \right\|_{L^2}^2 - \left\| \varphi_{\lambda}(T) k_{x_i} \right\|_{L^2}^2 \right| + \left| \frac{1}{n} \sum_{i=1}^n \left\| \varphi_{\lambda}(T) k_{x_i} \right\|_{L^2}^2 - \int_{\mathcal{X}} \left\| \varphi_{\lambda}(T) k_x \right\|_{L^2}^2 d\mu(x) \right| \\ & \leq CM \left(\sqrt{v \mathcal{N}_1(\lambda) \ln \lambda^{-1}} + \sqrt{\mathcal{N}_{2,\varphi}(\lambda)} \right) \sqrt{v \mathcal{N}_1(\lambda) \ln \lambda^{-1}} + M \mathcal{N}_{2,\varphi}(\lambda) \sqrt{\frac{2}{n} \ln \frac{2}{\delta}} \\ & = o(\mathcal{N}_{2,\varphi}(\lambda)), \end{aligned}$$

其中对于最后一个估计，回顾 v 由 (3.59) 给出，故根据 (3.69) 可得:

$$\left[\sqrt{v \mathcal{N}_1(\lambda) \ln \lambda^{-1}} \right]^2 = \frac{M \mathcal{N}_1^2(\lambda)}{n} \ln \frac{\mathcal{N}_1(\lambda)}{\delta} (\ln \lambda^{-1})^2$$

$$\leq C \ln \frac{1}{\delta} \cdot \frac{\lambda^{-1/\beta}}{n} (\ln \lambda^{-1})^3 \cdot \lambda^{-1/\beta} = o(\lambda^{-1/\beta}) = o(\mathcal{N}_{2,\varphi}(\lambda)).$$

最后, 利用 Mercer 展开, 发现:

$$\begin{aligned} \|\varphi_\lambda(T)k_x\|_{L^2}^2 &= \left\| \varphi_\lambda(T) \sum_{m=1}^{\infty} \mu_m \sum_{l=1}^{d_m} \overline{e_{m,l}(x)} e_{m,l} \right\|_{L^2}^2 \\ &= \left\| \sum_{m=1}^{\infty} \varphi_\lambda(\mu_m) \mu_m \sum_{l=1}^{d_m} \overline{e_{m,l}(x)} e_{m,l} \right\|_{L^2}^2 \\ &= \sum_{m=1}^{\infty} (\varphi_\lambda(\mu_m) \mu_m)^2 \sum_{l=1}^{d_m} |e_{m,l}(x)|^2, \end{aligned}$$

因此确定性项写为:

$$\begin{aligned} \int_{\mathcal{X}} \|\varphi_\lambda(T)k_x\|_{L^2}^2 d\mu(x) &= \int_{\mathcal{X}} \left[\sum_{m=1}^{\infty} (\varphi_\lambda(\mu_m) \mu_m)^2 \sum_{l=1}^{d_m} |e_{m,l}(x)|^2 \right] d\mu(x) \\ &= \sum_{m=1}^{\infty} (\varphi_\lambda(\mu_m) \mu_m)^2 = \mathcal{N}_{2,\varphi}(\lambda). \end{aligned}$$

■

引理 3.5: 方差项 $\mathbf{Var}(\lambda)$ 随 λ 减小而增大, 即对于任何 $\lambda_1 \leq \lambda_2$, 有 $\mathbf{Var}(\lambda_1) \geq \mathbf{Var}(\lambda_2)$ 。

证明 定义核矩阵 $K = \frac{1}{n}(k(x_i, x_j))_{n \times n}$ 。容易证明 $\hat{T}$ 在集合 $\{k_{x_i}\}_{i=1}^n$ 上的表示矩阵由 K 给出 (参见例如 Li et al.^[89]Section A.1)。因此, 记列向量 $k(X, \cdot) = (k_{x_1}, \dots, k_{x_n})^\top$, 有

$$\varphi_\lambda(\hat{T})k(X, \cdot) = \varphi_\lambda(K)k(X, \cdot),$$

其中左侧 $\varphi_\lambda(\hat{T})$ 的作用是逐元素的。那么,

$$\begin{aligned} \mathbf{Var}(\lambda) &= \frac{\sigma^2}{n^2} \sum_{i=1}^n \|\varphi_\lambda(\hat{T})k_{x_i}\|_{L^2}^2 = \frac{\sigma^2}{n^2} \sum_{i=1}^n \int_{\mathcal{X}} |(\varphi_\lambda(\hat{T})k_{x_i})(x)|^2 d\mu(x) \\ &= \frac{\sigma^2}{n^2} \int_{\mathcal{X}} \sum_{i=1}^n |(\varphi_\lambda(\hat{T})k_{x_i})(x)|^2 d\mu(x) = \frac{\sigma^2}{n^2} \int_{\mathcal{X}} \|(\varphi_\lambda(\hat{T})k(X, \cdot))(x)\|_{\mathbb{R}^n}^2 d\mu(x) \\ &= \frac{\sigma^2}{n^2} \int_{\mathcal{X}} \|(\varphi_\lambda(K)k(X, \cdot))(x)\|_{\mathbb{R}^n}^2 d\mu(x) \\ &= \frac{\sigma^2}{n^2} \int_{\mathcal{X}} \|\varphi_\lambda(K)k(X, x)\|_{\mathbb{R}^n}^2 d\mu(x) \\ &= \frac{\sigma^2}{n^2} \int_{\mathcal{X}} [k(X, x)]^H \varphi_\lambda^2(K)k(X, x) d\mu(x), \end{aligned}$$

其中 $[k(X, x)]^H$ 是 $k(X, x)$ 的共轭转置。此外, 滤谱函数的性质 (i) 意味着 $\varphi_\lambda(z)$ 随

λ 减小而增大。因此, 得到 $\varphi_{\lambda_1}^2(K) \geq \varphi_{\lambda_2}^2(K)$, 结论成立。 ■

推论 3.3: 当 $\lambda = \lambda(n) = O(n^{-\beta})$ 时, 有

$$\mathbf{Var}(\lambda) = \Omega_{\mathbb{P}}((\ln n)^{-4} \sigma^2). \quad (3.72)$$

证明 选择 $\tilde{\lambda} = n^{-\beta}(\ln n)^{4\beta}$, 则当 n 足够大时有 $\lambda \leq \tilde{\lambda}$ 。利用引理 3.5, 得到 $\mathbf{Var}(\lambda) \geq \mathbf{Var}(\tilde{\lambda})$ 。由于 $\tilde{\lambda}$ 的选择满足条件 (3.69), 故应用定理 3.2 可得:

$$\mathbf{Var}(\tilde{\lambda}) = [1 + o_{\mathbb{P}}(1)] \frac{\sigma^2}{n} \mathcal{N}_{2,\varphi}(\tilde{\lambda}) = \Omega_{\mathbb{P}}\left(\frac{\sigma^2}{n} \tilde{\lambda}^{-1/\beta}\right) = \Omega_{\mathbb{P}}(\sigma^2(\ln n)^{-4}).$$

引理 3.6: 至少以概率 $1 - \delta$, 有

$$\left| \frac{1}{n} \sum_{i=1}^n \|\varphi_{\lambda}(T)k_{x_i}\|_{L^2}^2 - \int_{\mathcal{X}} \|\varphi_{\lambda}(T)k_x\|_{L^2}^2 d\mu(x) \right| \leq M \mathcal{N}_{2,\varphi}(\lambda) \sqrt{\frac{2}{n} \ln \frac{2}{\delta}}. \quad (3.73)$$

证明 令 $\xi(x) = \|\varphi_{\lambda}(T)k_x\|_{L^2}^2$ 且 $\xi_i = \xi(x_i)$ 。则它们是独立同分布随机变量, 根据 (3.50) 有:

$$|\xi|^2 = \|\varphi_{\lambda}(T)k_x\|_{L^2}^2 \leq M \mathcal{N}_{2,\varphi}(\lambda).$$

利用引理 3.11 即可得到所需结果。 ■

引理 3.7: 在假设 3.2 和假设 3.3 下, 当 (3.60) 和 (3.61) 成立时, 有:

$$\begin{aligned} \sup_{x \in \mathcal{X}} \left| \|\varphi_{\lambda}(\hat{T})k_x\|_{L^2}^2 - \|\varphi_{\lambda}(T)k_x\|_{L^2}^2 \right| &\leq CM \left(\sqrt{v \mathcal{N}_1(\lambda) \ln \lambda^{-1}} + \sqrt{\mathcal{N}_{2,\varphi}(\lambda)} \right) \\ &\quad \cdot \sqrt{v \mathcal{N}_1(\lambda) \ln \lambda^{-1}}. \end{aligned} \quad (3.74)$$

证明 从下式开始:

$$D = \left| \|\varphi_{\lambda}(\hat{T})k_x\|_{L^2} - \|\varphi_{\lambda}(T)k_x\|_{L^2} \right| \leq \left\| T^{\frac{1}{2}} [\varphi_{\lambda}(T) - \varphi_{\lambda}(\hat{T})] k_x \right\|_{\mathcal{H}_k}.$$

利用算子演算, 得到:

$$\begin{aligned} &T^{\frac{1}{2}} [\varphi_{\lambda}(T) - \varphi_{\lambda}(\hat{T})] k_x \\ &= T^{\frac{1}{2}} \left[\frac{1}{2\pi i} \oint_{\Gamma_{\lambda}} R_{\hat{T}}(z)(T - \hat{T})R_T(z)\varphi_{\lambda}(z) dz \right] k_x \\ &= \frac{1}{2\pi i} \oint_{\Gamma_{\lambda}} T^{\frac{1}{2}}(\hat{T} - z)^{-1}(T - \hat{T})(T - z)^{-1}k_x\varphi_{\lambda}(z) dz \\ &= \frac{1}{2\pi i} \oint_{\Gamma_{\lambda}} T^{\frac{1}{2}}T_{\lambda}^{-\frac{1}{2}} \cdot T_{\lambda}^{\frac{1}{2}}(\hat{T} - z)^{-1}T_{\lambda}^{\frac{1}{2}} \cdot T_{\lambda}^{-\frac{1}{2}}(T - \hat{T})T_{\lambda}^{-\frac{1}{2}} \cdot T_{\lambda}^{\frac{1}{2}}(T - z)^{-1}T_{\lambda}^{\frac{1}{2}} \cdot T_{\lambda}^{-\frac{1}{2}}k_x\varphi_{\lambda}(z) dz. \end{aligned}$$

因此, 取范数可得:

$$\begin{aligned}
 D &\leq C \left\| T^{\frac{1}{2}} T_{\lambda}^{-\frac{1}{2}} \right\| \cdot \left\| T_{\lambda}^{\frac{1}{2}} (\hat{T} - z)^{-1} T_{\lambda}^{\frac{1}{2}} \right\| \cdot \left\| T_{\lambda}^{-\frac{1}{2}} (T - \hat{T}) T_{\lambda}^{-\frac{1}{2}} \right\| \cdot \left\| T_{\lambda}^{\frac{1}{2}} (T - z)^{-1} T_{\lambda}^{\frac{1}{2}} \right\| \\
 &\quad \cdot \left\| T_{\lambda}^{-\frac{1}{2}} k_x \right\|_{\mathcal{H}_k} \oint_{\Gamma_{\lambda}} |\varphi_{\lambda}(z) dz| \\
 &\leq C \cdot 1 \cdot C \cdot C \sqrt{\nu} \cdot C \cdot \sqrt{M \mathcal{N}_1(\lambda)} \oint_{\Gamma_{\lambda}} |\varphi_{\lambda}(z) dz| \\
 &\leq C \sqrt{M \nu \mathcal{N}_1(\lambda)} \oint_{\Gamma_{\lambda}} |\varphi_{\lambda}(z) dz|,
 \end{aligned}$$

其中在第二个估计中, 对各部分分别应用了: (1) 算子演算, (2,4) 命题 3.7, (3) 估计 (3.60), 以及 (5) 估计 (3.51)。根据假设 3.3, 有:

$$\oint_{\Gamma_{\lambda}} |\varphi_{\lambda}(z) dz| \leq C \oint_{\Gamma_{\lambda}} \frac{1}{|z + \lambda|} |dz|.$$

现在关注后一个积分。对于 $z \in \Gamma_{\lambda,1}$, 有 $|z + \lambda| \geq \lambda(2\sqrt{2})$, 因此:

$$\int_{\Gamma_{\lambda,1}} \frac{1}{|z + \lambda|} |dz| \leq 2\sqrt{2}\lambda^{-1} |\Gamma_{\lambda,1}| \leq C,$$

其中注意到 $|\Gamma_{\lambda,1}| \leq C\lambda$ 。对于 $\Gamma_{\lambda,2}$, 有:

$$\begin{aligned}
 \int_{\Gamma_{\lambda,2}} \frac{1}{|z + \lambda|} |dz| &= 2 \int_0^{\kappa^2} \frac{1}{|x + (x + \lambda/2)i + \lambda|} \sqrt{2} dx \\
 &\leq C \int_0^{\kappa^2} \frac{1}{x + \lambda} dx \\
 &\leq C \ln \lambda^{-1}.
 \end{aligned}$$

对于 $z \in \Gamma_{\lambda,3}$, 有 $|z + \lambda| \geq \kappa^2$, 因此:

$$\int_{\Gamma_{\lambda,3}} \frac{1}{|z + \lambda|} |dz| \leq \frac{1}{\kappa^2} |\Gamma_{\lambda,3}| \leq C.$$

因此, 得到:

$$\oint_{\Gamma_{\lambda}} \frac{1}{|z + \lambda|} |dz| \leq C \ln \lambda^{-1}, \quad (3.75)$$

故:

$$D = \left| \|\varphi_{\lambda}(\hat{T})k_x\|_{L^2} - \|\varphi_{\lambda}(T)k_x\|_{L^2} \right| \leq C \sqrt{M \nu \mathcal{N}_1(\lambda)} \ln \lambda^{-1}.$$

随后, 结合 (3.50) 中的第二个估计, 有:

$$\begin{aligned}
 \|\varphi_{\lambda}(\hat{T})k_x\|_{L^2} + \|\varphi_{\lambda}(T)k_x\|_{L^2} &\leq 2 \|\varphi_{\lambda}(T)k_x\|_{L^2} + D \\
 &\leq C \sqrt{M} \left(\sqrt{\mathcal{N}_{2,\varphi}(\lambda)} + \sqrt{\nu \mathcal{N}_1(\lambda)} \ln \lambda^{-1} \right).
 \end{aligned}$$

最后：

$$\begin{aligned}
 \left| \|\varphi_\lambda(\hat{T})k_x\|_{L^2}^2 - \|\varphi_\lambda(T)k_x\|_{L^2}^2 \right| &= \left| \|\varphi_\lambda(\hat{T})k_x\|_{L^2} - \|\varphi_\lambda(T)k_x\|_{L^2} \right| \\
 &\quad \cdot (\|\varphi_\lambda(\hat{T})k_x\|_{L^2} + \|\varphi_\lambda(T)k_x\|_{L^2}) \\
 &\leq CM \left(\sqrt{\mathcal{N}_{2,\varphi}(\lambda)} + \sqrt{v\mathcal{N}_1(\lambda) \ln \lambda^{-1}} \right) \\
 &\quad \cdot \sqrt{v\mathcal{N}_1(\lambda) \ln \lambda^{-1}}.
 \end{aligned}$$

■

3.4.6 偏差项

定理 3.3: 令 $\lambda = \lambda(n)$ 满足 $\lambda = \Omega(n^{-\theta})$ (对于某些 $\theta < \beta$)。在假设 3.1、3.2、3.3 和 3.5 下, 有:

$$\mathbf{Bias}^2(\lambda) = \mathcal{R}_\varphi^2(\lambda; f^*) + o_{\mathbb{P}} \left(\mathcal{R}_\varphi^2(\lambda; f^*) + \frac{1}{n} \mathcal{N}_{2,\varphi}(\lambda) \right). \quad (3.76)$$

更精确地, 令 $\tilde{s} = \min(s, 2\tau_{\max})$, 若 $\tilde{s} \leq 2$, 或者 $\tilde{s} > 2$ 且满足 $n^{-1}\lambda^{-(\beta^{-1}+\tilde{s}-2)}(\ln n)^3 = o(1)$, 则有:

$$\mathbf{Bias}^2(\lambda) = (1 + o_{\mathbb{P}}(1))\mathcal{R}_\varphi^2(\lambda; f^*). \quad (3.77)$$

定理 3.3 的证明 首先, 可以应用推论 3.2 使得 (3.60) 和 (3.61) 成立。回顾:

$$g^* = Tf^*, \quad f_\lambda^* = \varphi_\lambda(T)Tf^* = \varphi_\lambda(T)g^*.$$

正如第 3.4.1 小节中所述, 偏差项定义为:

$$\mathbf{Bias}^2(\lambda) = \|\tilde{f}_\lambda - f^*\|_{L^2}^2 = \|f_\lambda^* - f^* + \tilde{f}_\lambda - f_\lambda^*\|_{L^2}^2.$$

因此有:

$$\|f_\lambda^* - f^*\|_{L^2} - \|\tilde{f}_\lambda - f_\lambda^*\|_{L^2} \leq \mathbf{Bias}(\lambda) \leq \|f_\lambda^* - f^*\|_{L^2} + \|\tilde{f}_\lambda - f_\lambda^*\|_{L^2},$$

其中 $\|f_\lambda^* - f^*\|_{L^2} = \mathcal{R}_\varphi(\lambda; f^*)$ 是 (3.34) 中定义的主项。对于误差项, 进行如下分解:

$$\begin{aligned}
 \tilde{f}_\lambda - f_\lambda^* &= \varphi_\lambda(\hat{T})\tilde{g}_X - (\psi_\lambda(\hat{T}) + \varphi_\lambda(\hat{T})\hat{T})f_\lambda^* \\
 &= \varphi_\lambda(\hat{T})(\tilde{g}_X - \hat{T}f_\lambda^*) - \psi_\lambda(\hat{T})T\varphi_\lambda(T)f^* \\
 &= \varphi_\lambda(\hat{T})(\tilde{g}_X - \hat{T}f_\lambda^*) - \varphi_\lambda(\hat{T})\psi_\lambda(T)g^* + \varphi_\lambda(\hat{T})\psi_\lambda(T)g^* - \psi_\lambda(\hat{T})T\varphi_\lambda(T)f^* \\
 &= \varphi_\lambda(\hat{T})[\tilde{g}_X - \hat{T}f_\lambda^* - \psi_\lambda(T)g^*] + [\varphi_\lambda(\hat{T})\psi_\lambda(T)Tf^* - \psi_\lambda(\hat{T})T\varphi_\lambda(T)f^*] \\
 &= \varphi_\lambda(\hat{T})(\tilde{g}_X - \hat{T}f_\lambda^* - g^* + Tf_\lambda^*) + (\varphi_\lambda(\hat{T})T\psi_\lambda(T) - \psi_\lambda(\hat{T})T\varphi_\lambda(T))f^*.
 \end{aligned} \quad (3.78)$$

对于 (3.78) 中的第一项:

$$\begin{aligned}
 & \left\| \varphi_\lambda(\hat{T})(\tilde{g}_X - \hat{T}f_\lambda^* - g^* + Tf_\lambda^*) \right\|_{L^2} \\
 &= \left\| T^{\frac{1}{2}} \varphi_\lambda(\hat{T})(\tilde{g}_X - \hat{T}f_\lambda^* - g^* + Tf_\lambda^*) \right\|_{\mathcal{H}_k} \\
 &\leq \left\| T^{\frac{1}{2}} T_\lambda^{-\frac{1}{2}} \right\| \cdot \left\| T_\lambda^{\frac{1}{2}} \varphi_\lambda(\hat{T}) T_\lambda^{\frac{1}{2}} \right\| \cdot \left\| T_\lambda^{-\frac{1}{2}} [(\tilde{g}_X - \hat{T}f_\lambda^*) - (g^* - Tf_\lambda^*)] \right\|_{\mathcal{H}_k} \\
 &\stackrel{(i)}{\leq} 1 \cdot E \left\| T_\lambda^{\frac{1}{2}} T_{X\lambda}^{-1} T_\lambda^{\frac{1}{2}} \right\| \cdot \left\| T_\lambda^{-\frac{1}{2}} [(\tilde{g}_X - \hat{T}f_\lambda^*) - (g^* - Tf_\lambda^*)] \right\|_{\mathcal{H}_k} \\
 &\leq C \left\| T_\lambda^{-\frac{1}{2}} [(\tilde{g}_X - \hat{T}f_\lambda^*) - (g^* - Tf_\lambda^*)] \right\|_{\mathcal{H}_k},
 \end{aligned}$$

其中 (i) 的第二个控制来自 (3.22), 最后一个控制可由 (3.61) 导出. 利用命题 3.3, 同样有:

$$\mathcal{R}_\varphi(\lambda; f^*) = \Omega(\lambda^{\tilde{s}/2}), \quad (3.79)$$

这里记 $\tilde{s} = \min(s, 2\tau_{\max})$. 因此, 根据引理 3.8 (其中 t 足够接近 s), 有:

$$\left\| T_\lambda^{\frac{1}{2}} (\tilde{g}_X - \hat{T}f_\lambda^* - g^* + Tf_\lambda^*) \right\|_{L^2} = o_{\mathbb{P}}(\mathcal{R}_\varphi(\lambda; f^*)). \quad (3.80)$$

对于 (3.78) 中的第二项, 由于对于某些 $\theta < \beta$ 有 $\lambda = \Omega(n^{-\theta})$, 正如方差项证明中 (3.71) 所讨论的, 有:

$$v = O(n^{-1} \mathcal{N}_1(\lambda) \ln \lambda^{-1}) = O(n^{-1} \lambda^{-\frac{1}{\beta}} \ln \lambda^{-1}) = o(1),$$

故满足引理 3.4 中的条件. 结合引理 3.4 和引理 3.9, 对于任何满足 $t < s$ 且 $t \leq 2$ 的固定 t :

$$\left\| (\varphi_\lambda(\hat{T})T\psi_\lambda(T) - \psi_\lambda(\hat{T})T\varphi_\lambda(T))f^* \right\|_{L^2} = O_{\mathbb{P}}\left(\mathcal{R}_\varphi(\lambda; f^*) + \|f^*\|_{[\mathcal{H}_k]^t} \lambda^{\frac{t}{2}}\right) \sqrt{v} \ln \lambda^{-1}.$$

此外, 由于 $\sqrt{v} \ln \lambda^{-1} = o(1)$, 故:

$$\mathcal{R}_\varphi(\lambda; f^*) \sqrt{v} \ln \lambda^{-1} = o(\mathcal{R}_\varphi(\lambda; f^*)).$$

对于最后一项, 注意到:

$$\left(\lambda^{\frac{t}{2}} \sqrt{v} \ln \lambda^{-1} \right)^2 = O\left(\lambda^t n^{-1} \lambda^{-\frac{1}{\beta}} (\ln \lambda^{-1})^3 \right).$$

考虑以下情况:

• **情况 1:** 利用 (3.79), 如果对于某些 $t < s$ 且 $t \leq 2$, 有:

$$n^{-1} \lambda^{-(\beta^{-1} + \tilde{s} - t)} (\ln \lambda^{-1})^3 = O\left(n^{-1} \lambda^{-(\beta^{-1} + \tilde{s} - t)} (\ln n)^3 \right) = o(1), \quad (3.81)$$

则有 $\lambda^{\frac{t}{2}} \sqrt{v} \ln \lambda^{-1} = o(\mathcal{R}_\varphi(\lambda; f^*))$.

- **情况 2:** 利用 (3.71), 如果 $\lambda^t (\ln \lambda^{-1})^3 = O(\lambda^t (\ln n)^3) = o(1)$, 则有 $\lambda^{\frac{t}{2}} \sqrt{v} \ln \lambda^{-1} = o\left(\frac{1}{n} \mathcal{N}_{2,\varphi}(\lambda)\right)^{1/2}$.

现在, 如果 $\tilde{s} \leq 2$, 由于 $\lambda = \Omega(n^{-\theta})$ (其中 $\theta < \beta$), 通过选择足够接近 s 的 t (即 $t > \tilde{s} - \left(\frac{1}{\theta} - \frac{1}{\beta}\right)$), 总能满足 (3.81), 故情况 1 总是成立。

另一方面, 如果 $\tilde{s} > 2$, 固定 $t = 2$ 并固定某个 $0 < \theta_0 < (\beta^{-1} + \tilde{s} - 2)^{-1}$. 那么, 当 $\lambda \geq n^{-\theta_0}$ 时, 适用情况 1; 当 $\lambda \leq n^{-\theta_0}$ 时, 适用情况 2. 综上所述, 总有:

$$\lambda^{\frac{t}{2}} \sqrt{v} \ln \lambda^{-1} = o\left(\mathcal{R}_{\varphi}^2(\lambda; f^*) + \frac{1}{n} \mathcal{N}_{2,\varphi}(\lambda)\right)^{1/2}.$$

因此, 证明了:

$$\|[\varphi_{\lambda}(\hat{T})T\psi_{\lambda}(T) - \psi_{\lambda}(\hat{T})T\varphi_{\lambda}(T)]f^*\|_{L^2} = o_{\mathbb{P}}\left(\mathcal{R}_{\varphi}^2(\lambda; f^*) + \frac{1}{n} \mathcal{N}_{2,\varphi}(\lambda)\right)^{1/2}.$$

结合 (3.80), 证明了误差项 $\|\tilde{f}_{\lambda} - f_{\lambda}^*\|_{L^2}$ 同样具有该阶, 从而 (3.76) 成立. ■

以下引理控制了偏差项中的近似误差, 其思路类似于 Li et al.^[88] 中的 Lemma A.5 和 Lemma A.10 的结合, 但此处考虑通用的滤谱算法. 此外, 也应用了 Zhang et al.^[36] 中的技术来处理错配情形. 证明推迟至附录。

引理 3.8: 在假设 3.1 和 3.2 成立下, 令 $f^* \in [\mathcal{H}_k]^t$, φ_{λ} 为修饰阶为 τ 的滤谱函数. 假设对于某些 $\theta < \beta$ 有 $\lambda = \Omega(n^{-\theta})$. 则存在 (取决于 θ 的) 某个 $\varepsilon > 0$, 使得有:

$$\left\|T_{\lambda}^{-\frac{1}{2}}[(\tilde{g}_X - \hat{T}f_{\lambda}^*) - (g^* - Tf_{\lambda}^*)]\right\|_{\mathcal{H}_k} = O_{\mathbb{P}}\left(n^{-\varepsilon} \lambda^{\tilde{t}/2}\right). \quad (3.82)$$

这里 $g^* = Tf^*$, $f_{\lambda}^* = T\varphi_{\lambda}(T)f^*$, 且 $\tilde{t} = \min(t, 2\tau)$.

接下来的引理处理了 (3.78) 中的交互项, 此处应用了解析算子演算论据。

引理 3.9: 在假设 3.2、3.3 下, 假设 $f^* \in [\mathcal{H}_k]^t$, 当 (3.60) 和 (3.61) 成立时, 有:

$$\|(\varphi_{\lambda}(\hat{T})T\psi_{\lambda}(T) - \psi_{\lambda}(\hat{T})T\varphi_{\lambda}(T))f^*\|_{L^2} \leq C\left(\mathcal{R}_{\varphi}(\lambda; f^*) + \|f^*\|_{[\mathcal{H}_k]^{\tilde{t}}} \lambda^{\tilde{t}/2}\right) \sqrt{v} \ln \lambda^{-1},$$

其中 $\tilde{t} = \min(t, 2)$.

证明 首先进行如下分解:

$$\begin{aligned} & \|(\varphi_{\lambda}(\hat{T})T\psi_{\lambda}(T) - \psi_{\lambda}(\hat{T})T\varphi_{\lambda}(T))f^*\|_{L^2} \\ &= \left\|T^{\frac{1}{2}}(\varphi_{\lambda}(\hat{T})T\psi_{\lambda}(T) - \psi_{\lambda}(\hat{T})T\varphi_{\lambda}(T))f^*\right\|_{\mathcal{H}_k} \\ &\leq \left\|T^{\frac{1}{2}}(\varphi_{\lambda}(\hat{T})T\psi_{\lambda}(T) - \psi_{\lambda}(T)T\varphi_{\lambda}(T))f^*\right\|_{\mathcal{H}_k} \\ &\quad + \left\|T^{\frac{1}{2}}(\psi_{\lambda}(\hat{T})T\varphi_{\lambda}(T) - \psi_{\lambda}(T)T\varphi_{\lambda}(T))f^*\right\|_{\mathcal{H}_k}. \end{aligned} \quad (3.83)$$

对于 (3.83) 中的第二项, 使用类似于引理 3.7 中的论证. 根据在 (3.67) 定义的

Γ_λ 上的命题 3.5, 有:

$$\begin{aligned}
 & T^{\frac{1}{2}}(\psi_\lambda(\hat{T})T\varphi_\lambda(T) - \psi_\lambda(T)T\varphi_\lambda(T))f^* \\
 &= T^{\frac{1}{2}} \left[\frac{1}{2\pi i} \oint_{\Gamma_\lambda} R_{\hat{T}}(z)(T - \hat{T})R_T(z)\psi_\lambda(z) dz \right] T\varphi_\lambda(T)f^* \\
 &= \frac{1}{2\pi i} \oint_{\Gamma_\lambda} T^{\frac{1}{2}}(\hat{T} - z)^{-1}(T - \hat{T})(T - z)^{-1}\psi_\lambda(z)T\varphi_\lambda(T)f^* dz \\
 &= \frac{1}{2\pi i} \int_{\Gamma_\lambda} T^{\frac{1}{2}}T_\lambda^{-\frac{1}{2}} \cdot T_\lambda^{\frac{1}{2}}(T_X - z)^{-1}T_\lambda^{\frac{1}{2}} \cdot T_\lambda^{-\frac{1}{2}}(T - \hat{T})T_\lambda^{-\frac{1}{2}} \\
 &\quad \cdot T_\lambda^{\frac{1}{2}}(T - z)^{-1}T_\lambda^{\frac{1}{2}} \cdot T_\lambda^{-\frac{1}{2}}T\varphi_\lambda(T)f^*\psi_\lambda(z) dz.
 \end{aligned}$$

因此:

$$\begin{aligned}
 & \left\| T^{\frac{1}{2}}(\psi_\lambda(\hat{T})T\varphi_\lambda(T) - \psi_\lambda(T)T\varphi_\lambda(T))f^* \right\|_{\mathcal{H}_k} \\
 & \leq C \int_{\Gamma_\lambda} \left\| T^{\frac{1}{2}}T_\lambda^{-\frac{1}{2}} \right\| \cdot \left\| T_\lambda^{\frac{1}{2}}(T_X - z)^{-1}T_\lambda^{\frac{1}{2}} \right\| \cdot \left\| T_\lambda^{-\frac{1}{2}}(T - \hat{T})T_\lambda^{-\frac{1}{2}} \right\| \\
 & \quad \cdot \left\| T_\lambda^{\frac{1}{2}}(T - z)^{-1}T_\lambda^{\frac{1}{2}} \right\| \cdot \left\| T_\lambda^{-\frac{1}{2}}T\varphi_\lambda(T)f^* \right\|_{\mathcal{H}_k} |\psi_\lambda(z) dz| \\
 & \stackrel{(a)}{\leq} C \cdot 1 \cdot C \cdot \sqrt{\nu} \cdot C \cdot \left\| T_\lambda^{-\frac{1}{2}}T\varphi_\lambda(T)f^* \right\|_{\mathcal{H}_k} \cdot \lambda \int_{\Gamma_\lambda} \left| \frac{1}{z + \lambda} \right| dz \\
 & \stackrel{(b)}{\leq} C\sqrt{\nu} \left\| T_\lambda^{-\frac{1}{2}}T\varphi_\lambda(T)f^* \right\|_{\mathcal{H}_k} \lambda \ln \lambda^{-1},
 \end{aligned}$$

其中在 (a) 中, 分别对各项应用了: (1) 算子演算, (2,4) 命题 3.7, (3) 估计 (3.60), 以及 (6) 假设 3.3 中的条件 (C2); 在 (b) 中对最后一项应用了 (3.75)。

令 $\tilde{t} = \min(t, 2)$ 。由于 $f^* \in [\mathcal{H}_k]^t$, 故也有 $f^* \in [\mathcal{H}_k]^{\tilde{t}}$, 故可写 $f^* = T^{\tilde{t}/2}h$ (对于满足 $\|h\|_{L^2} = \|f^*\|_{[\mathcal{H}_k]^{\tilde{t}}}$ 的某个 $h \in L^2$)。由此可得:

$$\begin{aligned}
 \left\| T_\lambda^{-\frac{1}{2}}T\varphi_\lambda(T)f^* \right\|_{\mathcal{H}_k} &= \left\| T^{-\frac{1}{2}}T_\lambda^{-\frac{1}{2}}T\varphi_\lambda(T)T^{\frac{\tilde{t}}{2}}h \right\|_{L^2} \\
 &\leq \left\| T^{\frac{1}{2}}T_\lambda^{-\frac{1}{2}} \right\| \cdot \left\| T^{\frac{\tilde{t}}{2}}\varphi_\lambda(T) \right\| \cdot \|h\|_{L^2} \\
 &\leq C \|f^*\|_{[\mathcal{H}_k]^{\tilde{t}}} \lambda^{\frac{\tilde{t}}{2}-1},
 \end{aligned}$$

其中最后一个不等式来自引理 3.10。因此, 有:

$$\left\| T^{\frac{1}{2}}(\psi_\lambda(\hat{T})T\varphi_\lambda(T) - \psi_\lambda(T)T\varphi_\lambda(T))f^* \right\|_{\mathcal{H}_k} \leq C \|f^*\|_{[\mathcal{H}_k]^{\tilde{t}}} \sqrt{\nu} \lambda^{\frac{\tilde{t}}{2}} \ln \lambda^{-1}. \quad (3.84)$$

对于 (3.83) 中的第一项，同样采用解析算子演算论据：

$$\begin{aligned}
 & T^{\frac{1}{2}}(\varphi_\lambda(\hat{T})T\psi_\lambda(T) - \psi_\lambda(T)T\varphi_\lambda(T))f^* \\
 &= T^{\frac{1}{2}}(\varphi_\lambda(\hat{T}) - \varphi_\lambda(T))T\psi_\lambda(T)f^* \\
 &= \frac{1}{2\pi i} \oint_{\Gamma_\lambda} T^{\frac{1}{2}}(\hat{T} - z)^{-1}(\hat{T} - T)(T - z)^{-1}\varphi_\lambda(z)T\psi_\lambda(T)f^* dz \\
 &= \frac{1}{2\pi i} \oint_{\Gamma_\lambda} T^{\frac{1}{2}}T_\lambda^{-\frac{1}{2}} \cdot T_\lambda^{\frac{1}{2}}(T_X - z)^{-1}T_\lambda^{\frac{1}{2}} \cdot T_\lambda^{-\frac{1}{2}}(T - \hat{T})T_\lambda^{-\frac{1}{2}} \\
 &\quad \cdot T_\lambda^{\frac{1}{2}}(T - z)^{-1}T_\lambda^{\frac{1}{2}} \cdot T_\lambda^{-\frac{1}{2}}T^{\frac{1}{2}} \cdot T^{\frac{1}{2}}\psi_\lambda(T)f^*\varphi_\lambda(z) dz.
 \end{aligned}$$

因此有：

$$\begin{aligned}
 & \|T^{\frac{1}{2}}(\varphi_\lambda(\hat{T})T\psi_\lambda(T) - \psi_\lambda(T)T\varphi_\lambda(T))f^*\|_{\mathcal{H}_k} \\
 &\leq \oint_{\Gamma_\lambda} \left\| T^{\frac{1}{2}}T_\lambda^{-\frac{1}{2}} \right\| \cdot \left\| T_\lambda^{\frac{1}{2}}(\hat{T} - z)^{-1}T_\lambda^{\frac{1}{2}} \right\| \cdot \left\| T_\lambda^{-\frac{1}{2}}(T - \hat{T})T_\lambda^{-\frac{1}{2}} \right\| \\
 &\quad \cdot \left\| T_\lambda^{\frac{1}{2}}(T - z)^{-1}T_\lambda^{\frac{1}{2}} \right\| \left\| T_\lambda^{-\frac{1}{2}}T^{\frac{1}{2}} \right\| \cdot \left\| T^{\frac{1}{2}}\psi_\lambda(T)f^* \right\|_{\mathcal{H}_k} |\varphi_\lambda(z) dz| \\
 &\leq C\sqrt{\nu}\mathcal{R}_\varphi(\lambda; f^*) \oint_{\Gamma_\lambda} |\varphi_\lambda(z) dz| \\
 &\leq C\sqrt{\nu}\mathcal{R}_\varphi(\lambda; f^*) \ln \lambda^{-1},
 \end{aligned}$$

其中最后一个控制由于假设 3.3 中的条件 (C2) 成立。随后，将之前的控制和 (3.84) 代入 (3.83) 即可得到所需结果。 ■

3.5 辅助结论

这一初等命题证明了假设 3.1 中 (3.9) 与 (3.10) 的等价性。

命题 3.8： 设 $(a_j)_{j \geq 1}$ 为一个单调递减趋于零的正数列。则有

$$a_j = \Theta(j^{-\beta}), \iff \#\{j : a_j \geq \lambda\} = \Theta(\lambda^{-1/\beta}) \text{ 当 } \lambda \rightarrow 0 \text{ 时}.$$

证明 首先注意到，由于 a_j 是递减的，故有 $\max\{j : a_j \geq \lambda\} = \#\{j : a_j \geq \lambda\}$ 。

($\implies$): 假设 $cj^{-\beta} \leq a_j \leq Cj^{-\beta}$ 。则有

$$\max\{j : a_j \geq \lambda\} \geq \max\{j : cj^{-\beta} \geq \lambda\} = \Omega(\lambda^{-1/\beta}).$$

另一方面，

$$\max\{j : a_j \geq \lambda\} = \min\{j : a_{j+1} < \lambda\} \leq \min\{j : C(j+1)^{-\beta} < \lambda\} = O(\lambda^{-1/\beta}).$$

($\impliedby$): 令 $N(\lambda) = \max\{j : a_j \geq \lambda\}$ 且假设 $c\lambda^{-1/\beta} \leq N(\lambda) \leq C\lambda^{-1/\beta}$ 。注意到 $N(\lambda) \geq$

j 意味着 $a_j \geq \lambda$, 故有

$$a_j \geq \sup \{ \lambda : N(\lambda) \geq j \} \geq \sup \{ \lambda : c \lambda^{-1/\beta} \geq j \} = \Omega(j^{-\beta}).$$

另一方面, $N(\lambda) < j$ 意味着 $a_j < \lambda$, 故有

$$a_j \leq \inf \{ \lambda : N(\lambda) < j \} \leq \inf \{ \lambda : C \lambda^{-1/\beta} < j \} = O(j^{-\beta}).$$

■

命题 3.9: 设 $(\lambda_j)_{j \geq 1}$ 为计及重数的特征值递减序列。定义

$$\Phi(\varepsilon) = \# \{ j : \lambda_j \geq \varepsilon \}. \quad (3.85)$$

假设 φ_λ 是满足 (3.22) 和 (3.23) 的滤谱函数。则对于任意 $p \geq 1$ 和 $\lambda > 0$, 有

$$2^{-p} \Phi(2F_1 \lambda) \leq \mathcal{N}_{p,\varphi}(\lambda) \leq p E^p \lambda^{-1} \int_0^\lambda \Phi(x) dx. \quad (3.86)$$

特别地, 若满足假设 3.1, 则

$$\mathcal{N}_{p,\varphi}(\lambda) = \Theta(\lambda^{-1/\beta}), \quad \text{当 } \lambda \rightarrow 0 \text{ 时}. \quad (3.87)$$

证明 首先处理上界。滤谱函数的性质 (3.22) 给出

$$z \varphi_\lambda(z) \leq E \frac{z}{\lambda + z} \leq E \min(1, \lambda^{-1} z).$$

因此,

$$\mathcal{N}_{p,\varphi}(\lambda) = \sum_{j=1}^{\infty} [\lambda_j \varphi_\lambda(\lambda_j)]^p \leq E^p \sum_{j=1}^{\infty} \min(1, \lambda^{-p} \lambda_j^p) = E^p \lambda^{-p} \sum_{j=1}^{\infty} \min(\lambda, \lambda_j)^p.$$

注意到 $p \int_0^a x^{p-1} dx = a^p$, 有

$$\begin{aligned} \sum_{j=1}^{\infty} \min(\lambda, \lambda_j)^p &= \sum_{j=1}^{\infty} p \int_0^{\min(\lambda, \lambda_j)} x^{p-1} dx \\ &= p \int_0^\infty \left(\sum_{j=1}^{\infty} \mathbf{1}_{\{\min(\lambda, \lambda_j) \geq x\}} \right) x^{p-1} dx \\ &= p \int_0^\lambda \Phi(x) x^{p-1} dx. \end{aligned}$$

由此可见

$$\mathcal{N}_{p,\varphi}(\lambda) \leq p E^p \lambda^{-p} \int_0^\lambda \Phi(x) x^{p-1} dx \leq p E^p \lambda^{-1} \int_0^\lambda \Phi(x) dx, \quad (3.88)$$

其中最后一个不等式源于当 $x \leq \lambda$ 时 $x/\lambda \leq 1$ 。

对于下界, 首先由 (3.23) 可得 $\psi_\lambda(z) \leq F_1 \lambda z^{-1}$ 。结合 $z \varphi_\lambda(z) = 1 - \psi_\lambda(z)$, 得到

$$z \varphi_\lambda(z) = 1 - \psi_\lambda(z) \geq 1 - F_1 \lambda z^{-1} \geq \frac{1}{2}, \quad \forall z \geq 2F_1 \lambda.$$

因此, 记 $\tilde{\lambda} = 2F_1\lambda$, 有

$$\begin{aligned}\mathcal{N}_{p,\varphi}(\lambda) &= \sum_{j=1}^{\infty} [\lambda_j \varphi_{\lambda}(\lambda_j)]^p \\ &\geq \sum_{j \leq \Phi(\tilde{\lambda})} [\lambda_j \varphi_{\lambda}(\lambda_j)]^p \\ &\geq \sum_{j \leq \Phi(\tilde{\lambda})} 2^{-p} = 2^{-p} \Phi(2F_1\lambda).\end{aligned}$$

最后, 若满足假设 3.1, 则命题 3.8 意味着 $\Phi(\lambda) \asymp \lambda^{-1/\beta}$, 其中 $1/\beta < 1$, 故

$$\lambda^{-1} \int_0^{\lambda} \Phi(x) dx \leq C \lambda^{-1} \int_0^{\lambda} x^{-1/\beta} dx \leq C \lambda^{-1} \cdot \lambda^{1-1/\beta} = C \lambda^{-1/\beta},$$

这表明 (3.86) 两侧具有相同的阶 $\lambda^{-1/\beta}$. ■

命题 3.10: 设 $(a_m)_{m \geq 1}$ 为正数递减序列, $(b_m)_{m \geq 1}, (c_m)_{m \geq 1}$ 为两个正数列且满足

$$\sum_{k=1}^m b_k \leq \sum_{k=1}^m c_k, \quad \forall m \geq 1.$$

则对于任意 $N \geq 1$, 有

$$\sum_{m=1}^N a_m b_m \leq \sum_{m=1}^N a_m c_m.$$

证明 使用 Abel 求和公式, 有

$$\begin{aligned}\sum_{m=1}^N a_m b_m &= \sum_{m=1}^{N-1} (a_m - a_{m+1}) \sum_{k=1}^m b_k + a_N \sum_{k=1}^N b_k \\ &\leq \sum_{m=1}^{N-1} (a_m - a_{m+1}) \sum_{k=1}^m c_k + a_N \sum_{k=1}^N c_k = \sum_{m=1}^N a_m c_m.\end{aligned}$$
■

3.5.1 滤谱函数的性质

以下是关于 $\varphi_{\lambda}^{\text{KR}}$ 的一个众所周知的初等性质:

命题 3.11: 对于 $\lambda > 0$ 且 $\alpha \in [0, 1]$, 有

$$\frac{z^{\alpha}}{z + \lambda} \leq \lambda^{\alpha-1}. \quad (3.89)$$

引理 3.10: 设 φ_{λ} 为定义 3.3 中定义的滤谱函数. 则对于 $s \in [0, 1]$, 有

$$\sup_{z \in [0, \kappa^2]} \varphi_{\lambda}(z) z^s \leq E \lambda^{s-1}. \quad (3.90)$$

此外, 假设 (3.23) 对于 τ 成立, 则常数 F_r 满足

$$F_r \leq F_0^{1-\frac{r}{\tau}} F_\tau^{\frac{r}{\tau}}, \quad \forall r \in [0, \tau]. \quad (3.91)$$

更进一步, 当 $z \leq \frac{\lambda}{2E}$ 时, 有 $\psi_\lambda(z) \geq 1/2$ 。

证明 第一个不等式是 (3.22) 和 (3.89) 的结果。第二个不等式源于

$$z^r \psi_\lambda(z) = \psi_\lambda(z)^{1-\frac{r}{\tau}} \left(z^\tau \psi_\lambda(z) \right)^{\frac{r}{\tau}} \leq F_0^{1-\frac{r}{\tau}} (F_\tau \lambda^\tau)^{\frac{r}{\tau}} = F_0^{1-\frac{r}{\tau}} F_\tau^{\frac{r}{\tau}} \lambda^r.$$

对于最后一个不等式, 注意到 $\varphi_\lambda(z) \leq E\lambda^{-1}$, 故当 $z \leq \frac{\lambda}{2E}$ 时,

$$\psi_\lambda(z) = 1 - z\varphi_\lambda(z) \geq 1 - \frac{\lambda}{2E} E\lambda^{-1} = \frac{1}{2}.$$

■

3.5.2 正则 RKHS

命题 3.1 的证明 Fischer et al.^[35] 中的定理 9 表明, 嵌入映射的范数满足

$$\| [H]^\alpha \hookrightarrow L^\infty(\mathcal{X}, \mu) \| = \| k_\mu^\alpha \|_{L^\infty} := \operatorname{ess\,sup}_{x \in \mathcal{X}, \mu} \sum_{i=1}^{\infty} \lambda_i^\alpha e_i(x)^2 = \operatorname{ess\,sup}_{x \in \mathcal{X}, \mu} \sum_{m=1}^{\infty} \mu_m^\alpha \sum_{l=1}^{d_m} |e_{m,l}(x)|^2. \quad (3.92)$$

随后, 结合假设 3.2, 利用命题 3.10 取极限可得

$$\begin{aligned} \sum_{m=1}^{\infty} \mu_m^\alpha \sum_{l=1}^{d_m} |e_{m,l}(x)|^2 &\leq M \sum_{m=1}^{\infty} d_m \mu_m^\alpha \\ &= M \sum_{i=1}^{\infty} \lambda_i^\alpha \leq M \sum_{i=1}^{\infty} i^{-\alpha\beta}, \end{aligned}$$

因此, 只要 $\alpha > 1/\beta$, 嵌入范数就是有限的。

■

3.5.2.1 球面上的内积核

设 $\mathcal{X} = \mathbb{S}^d$ 为 d 维球面, μ 为 $\mathbb{S}^d$ 上的均匀测度。经典结果^[29]表明, 球面拉普拉斯算子 $\Delta_{\mathbb{S}^d}$ 的特征分解给出了如下正交直和分解:

$$L^2(\mathbb{S}^d) = \bigoplus_{m=0}^{\infty} \mathcal{H}_m^d(\mathbb{S}^d),$$

其中 $\mathcal{H}_m^d(\mathbb{S}^d)$ 由 $d+1$ 个变量的 n 次齐次调和多项式在 $\mathbb{S}^d$ 上的限制组成, 且 $\mathcal{H}_m^d(\mathbb{S}^d)$ 是 $\Delta_{\mathbb{S}^d}$ 对应于特征值 $-m(m+d-1)$ 的特征子空间。此外, $\mathcal{H}_m^d(\mathbb{S}^d)$ 的维度由下式给出:

$$a_m := \dim \mathcal{H}_m^d(\mathbb{S}^d) = \binom{m+d}{m} - \binom{m-2+d}{m-2} \asymp m^{d-1},$$

且满足 $\sum_{k \leq m} a_k = \binom{m+d}{m} + \binom{m-1+d}{m-1} \asymp m^d$ 。

此外, $\mathcal{H}_m^d(\mathbb{S}^d)$ 的再生核 $Z_m(x, y)$ 是良置且唯一的, 可以显式表示为

$$Z_m(x, y) = \sum_{l=1}^{a_m} Y_{m,l}(x) Y_{m,l}(y),$$

其中 $\{Y_{m,l}\}_{l=1}^{a_m}$ 是 $\mathcal{H}_m^d(\mathbb{S}^d)$ 的任意一组标准正交基。令 C_m^λ 表示 Gegenbauer 多项式, 其通常由如下幂级数定义:

$$\sum_{m=0}^{\infty} C_m^\lambda(t) \alpha^m = \frac{1}{(1 - 2t\alpha + \alpha^2)^\lambda}. \quad (3.93)$$

那么, 当 $d \geq 2$ 时, 有 (同 (2.25))

$$Z_m(x, y) = \frac{n + \lambda}{\lambda} C_m^\lambda(u), \quad u = \langle x, y \rangle, \quad \lambda = \frac{d-1}{2}, \quad (3.94)$$

此外, Dai et al.^[29]Corollary 1.27 表明

$$Z_m(x, x) = \frac{n + \lambda}{\lambda} C_m^\lambda(1) = \dim \mathcal{H}_m^d(\mathbb{S}^d). \quad (3.95)$$

更进一步, 有 Funk-Hecke 公式, 同 (2.26)。

通过对比 (3.6) 和 (2.26), 得出结论: $\mathcal{H}_m^d(\mathbb{S}^d)$ 是 T 对应于特征值 $\mu_m = \mu_m(h)$ 的特征子空间。因此, 得到 $Z_m = k_m$, 从而根据 (3.95) 有

$$k_m(x, x) = Z_m(x, x) = \dim \mathcal{H}_m^d(\mathbb{S}^d) = \dim V_m,$$

故假设 3.2 在 $M = 1$ 时成立。

3.5.2.2 单位球上的内积核

单位球的情况与球面类似。更多细节请参见 Dai et al.^[29]Section 11。考虑 d 维单位球 $\mathcal{X} = \mathbb{B}^d = \{x \in \mathbb{R}^{d+1} : \|x\| \leq 1\}$, 其中 μ 与经典权重 $W(x) = (1 - \|x\|^2)^{-1/2}$ 成正比。用 V_m^d 表示关于内积

$$\langle f, g \rangle_W = c_W \int_{\mathbb{B}^d} f(x) g(x) W(x) dx$$

的次数恰为 n 的正交多项式空间, 其中 c_W 是归一化常数。那么, 有

$$\dim V_m^d = \binom{m + d - 1}{m}.$$

此外, 有如下 Funk-Hecke 公式的类比^[29]Theorem 11.1.9:

命题 3.12 (Funk-Hecke 公式): 设 $\lambda = \frac{d-1}{2}$, h 为可积函数且满足 $\int_{-1}^1 |h(t)| (1 - t^2)^{\lambda-1/2} dt$ 有限。则对于每个 $P_m \in V_m^d$, 有

$$c_W \int_{\mathbb{B}^d} h(\langle x, y \rangle) P_m(y) W(y) dy = \mu_m(f) P_m(x), \quad \forall x \in \mathbb{B}^d, \quad (3.96)$$

其中 $\mu_m(h)$ 是由下式定义的常数:

$$\mu_m(h) = c_\lambda \int_{-1}^1 h(t) \frac{C_m^\lambda(t)}{C_m^\lambda(1)} (1-t^2)^{\lambda-\frac{1}{2}} dt,$$

且 c_λ 是使得 $\mu_0(1) = 1$ 的常数。

因此, 对于任何内积核 $k(x, y) = h(\langle x, y \rangle)$, V_m^d 构成了 T 对应于特征值 $\mu_m = \mu_m(h)$ 的一个特征子空间。

此外, Dai et al.^{[29]Corollary 11.1.8} 表明 V_m^d 的再生核 k_m 为

$$k_m(x, y) = \frac{n+\lambda}{2\lambda} [C_n^\lambda(\langle x, y \rangle + x_{d+1}y_{d+1}) + C_n^\lambda(\langle x, y \rangle - x_{d+1}y_{d+1})],$$

其中 $x_{d+1} = \sqrt{1 - \|x\|^2}$, $y_{d+1} = \sqrt{1 - \|y\|^2}$ 。因此,

$$k_m(x, x) = \frac{n+\lambda}{2\lambda} (C_n^\lambda(1) + C_n^\lambda(2\|x\|^2 - 1)) \leq \frac{n+\lambda}{\lambda} C_n^\lambda(1) = \frac{n+\lambda}{\lambda} \frac{\Gamma(2\lambda+n)}{\Gamma(2\lambda)n!}.$$

这里在 $C_n^\lambda(1)$ 中使用了 Dai et al.^{[29]Eq. (B.2.2)}。将 $\lambda = (d-1)/2$ 代入展开, 得到

$$\frac{n+\lambda}{\lambda} \frac{\Gamma(2\lambda+n)}{\Gamma(2\lambda)n!} / \dim V_n^d = \frac{2n+d-1}{n+d-1} \leq 2,$$

故假设 3.2 在 $M = 2$ 时也成立。

3.5.3 解析滤谱函数

为了进一步分析复平面上滤谱函数的性质, 首先回顾复分析中的一些结果。

命题 3.13 (最大模原理): 设 f 是开集 Ω 上的解析函数, $K \subset \Omega$ 是一个紧集。则满足

$$\sup_{z \in K} |f(z)| = \sup_{z \in \partial K} |f(z)|.$$

命题 3.14: 核岭回归 (例子 3.4)、迭代岭回归 (例子 3.5) 和梯度流 (例子 3.6) 的滤谱函数均满足假设 3.3。

证明 对于核岭回归的情况, 结论是显然的。

(I) 迭代岭回归与梯度流的条件 (C1): 定义 $H = \{z \in \mathbb{C} : \operatorname{Re}(z) \geq -\frac{\lambda}{2}\}$ 。注意梯度流和迭代岭回归的滤谱函数均具有 $\varphi_\lambda(z) = g(u)/(\lambda u)$ 的形式, 其中 $g(u)$ 是 H 上的解析函数, 且 $u = z/\lambda$ 。具体而言, 对于迭代岭回归, $g(u) = 1 - \frac{1}{(u+1)^p}$; 对于梯度流, $g(u) = 1 - e^{-u}$ 。注意到 $\frac{g(u)}{u}$ 可以延拓为 H 上的解析函数。因此,

$$|(z+\lambda)\varphi_\lambda(z)| = |(u+1)g(u)/u| =: G(u) > 0$$

也可以被视为 H 上的解析函数。在迭代岭回归的情况中, $\limsup_{|u| \rightarrow \infty} G(u) = \limsup_{|u| \rightarrow \infty} g(u) = 1$ 是有限的。因此, $\sup_{u \in H} G(u) < \infty$, 这意味着条件 (C1) 成立。梯度流的滤谱函数同样满足此条件, 因为 $\limsup_{|u| \rightarrow \infty} g(u) \leq \sup_{u \in H} g(u) \leq$

$1 + \sqrt{e} < \infty$ 。

(II) 迭代岭回归与梯度流的条件 (C2): 在迭代岭回归的情况中, 对于所有满足 $\operatorname{Re}(z) \geq -\frac{\lambda}{2}$ 的 z , 有

$$\left| \frac{z + \lambda}{\lambda} \psi_{\lambda}^{\text{IT}, p}(z) \right| = \left(\frac{|\lambda|}{|z + \lambda|} \right)^{p-1} \leq 2^{p-1}.$$

对于梯度流, 当 $\operatorname{Re}(z) \leq 0$ 且 $z \in D_{\lambda}$ 时,

$$\left| \frac{z + \lambda}{\lambda} \psi_{\lambda}^{\text{GF}}(z) \right| = \left| \frac{z + \lambda}{\lambda} \right| e^{\operatorname{Re}(-\frac{z}{\lambda})} \leq \frac{5}{2} \sqrt{e}.$$

若 $\operatorname{Re}(z) > 0$ 且 $z \in D_{\lambda}$, 则

$$\left| \frac{z + \lambda}{\lambda} \psi_{\lambda}^{\text{GF}}(z) \right| \leq 2 \left(1 + \frac{\operatorname{Re}(z)}{\lambda} \right) e^{\operatorname{Re}(-\frac{z}{\lambda})}.$$

注意到函数 $f(x) := (1+x)e^{-x}$ 在 $[0, +\infty)$ 上单调递减。因此, $\left| \frac{z + \lambda}{\lambda} \psi_{\lambda}^{\text{GF}}(z) \right| \leq 2$ 。 ■

命题 3.15: 梯度下降 (例子 3.7) 的滤谱函数满足假设 3.3。

证明 对于条件 (C1), 注意到 $(z + \lambda)\varphi_{\lambda}^{\text{GD}}(z)$ 可以延拓为 D_{λ} 上的解析函数。根据命题 3.13, 只需要证明 $\sup_{z \in \Gamma_{\lambda}} |(z + \lambda)\varphi_{\lambda}^{\text{GD}}(z)|$ 被一个与 λ 无关的常数所控制。实际上, 对于所有 $z \in \Gamma_{\lambda}$, 有 $\frac{|z + \lambda|}{|z|} \leq 1 + \frac{|\lambda|}{|z|} \leq 3$ 且 $|1 - \eta z|^{\frac{1}{\eta\lambda}} \leq \left| 1 + \frac{1}{2}\eta\lambda \right|^{\frac{1}{\eta\lambda}} \leq \sqrt{e}$ 。因此, 条件 (C1) 对 $\varphi_{\lambda}^{\text{GD}}$ 同样成立。

对于条件 (C2), 需要证明对于某个常数 $\varepsilon > 0$, 有

$$\sup_{z \in D_{\lambda}, \lambda \in (0, \varepsilon)} \left| \frac{z + \lambda}{\lambda} \psi_{\lambda}^{\text{GD}}(z) \right| = \sup_{z \in D_{\lambda}, \lambda \in (0, \varepsilon)} \left| \frac{z + \lambda}{\lambda} (1 - \eta z)^{\frac{1}{\eta\lambda}} \right| < \infty.$$

事实上, 当 $|1 - \eta z| \geq 1$ 时,

$$\left| \frac{z + \lambda}{\lambda} (1 - \eta z)^{\frac{1}{\eta\lambda}} \right| \leq \frac{3}{2} \left| 1 + \frac{1}{2}\eta\lambda \right|^{\frac{1}{\eta\lambda}} \leq \frac{3}{2} \sqrt{e}$$

对所有 $z \in D_{\lambda}$ 成立。当 $|1 - \eta z| < 1$ 时,

$$\left| \frac{z + \lambda}{\lambda} (1 - \eta z)^{\frac{1}{\eta\lambda}} \right| \leq 1 + \left| \frac{z}{\lambda} (1 - \eta z)^{\frac{1}{\eta\lambda}} \right| \leq 1 + \left| z (1 - \eta z)^{\frac{1}{\eta}} \right| \leq \frac{3}{2} + 2\kappa^2$$

对所有 $z \in D_{\lambda}$ 且 $\lambda \in (0, 1)$ 成立。 ■

3.5.4 回归函数的源条件

例子 3.8 的证明 在假设 3.1 和 $\sum_{k=1}^m d_k \asymp m^{\gamma}$ 的条件下, 有 $\mu_m \asymp m^{-\gamma\beta}$ 。取 $s = \frac{p}{\beta}$ 。那么对于所有 $t < s$, 有

$$\sum_{m=1}^{\infty} \frac{\tilde{f}_m^2}{\mu_m^t} = O \left(\sum_{m=1}^{\infty} m^{-(\gamma p + 1)} m^{t\gamma\beta} \right) = O \left(\sum_{m=1}^{\infty} m^{-1 - \gamma(p - \beta t)} \right) < \infty.$$

因此, 对所有 $t < s$ 均有 $f^* \in [\mathcal{H}_k]^t$ 。

另一方面,

$$\sum_{m: \mu_m \leq \lambda} \bar{f}_m^2 \geq c \sum_{m \geq C\lambda^{-(\gamma\beta)^{-1}}} m^{-\gamma p-1} = \Omega(\lambda^{\frac{p}{\beta}}) = \Omega(\lambda^s).$$

最后, 若 $f_j \asymp j^{-(p+1)/2}$, 有

$$\bar{f}_m^2 = \sum_{j=D_{m-1}}^{D_m} f_j^2 \asymp \int_{(m-1)^\gamma}^{m^\gamma} x^{-p-1} dx \asymp m^{-(\gamma p+1)},$$

其中 $D_m := \sum_{k=1}^m d_k \asymp m^\gamma$. ■

例子 3.9 的证明 令 $s = \frac{p}{q\beta}$. 那么在假设 3.1 下, 对于所有 $t < s$, 有

$$\sum_{j=1}^{\infty} \lambda_j^{-t} |f_j|^2 = \sum_{l=1}^{\infty} \lambda_{j(l)}^{-t} |f_{j(l)}|^2 \leq C \sum_{l=1}^{\infty} l^{tq\beta} l^{-(p+1)} \leq C \sum_{l=1}^{\infty} l^{-1-(p-tq\beta)},$$

这里回顾 $j(l) \asymp l^q$ 且 $|f_{j(l)}| \asymp l^{-(p+1)/2}$, 故对于所有 $t < s$ 均有 $f^* \in [\mathcal{H}_k]^t$.

另一方面,

$$\sum_{j: \lambda_j \leq \lambda} |f_j|^2 = \sum_{l: \lambda_{j(l)} \leq \lambda} |f_{j(l)}|^2 \geq c \sum_{l \geq C\lambda^{-1/(q\beta)}} l^{-(p+1)} = \Omega(\lambda^{\frac{p}{q\beta}}) = \Omega(\lambda^s).$$
■

3.5.5 引理 3.8 的证明

该证明遵循与 Li et al.^[88] 中 Lemma A.5 和 Lemma A.10 相同的思路, 但将结果建立在滤谱算法之上, 并通过正则再生核希尔伯特空间性质利用引理 3.1 改进了估计。证明分为两种情况: $t > 1/\beta$ 和 $t \leq 1/\beta$ 。对于第一种情况, 回归函数 f^* 是 L^∞ 有界的, 因此可以直接应用 Bernstein 不等式。对于第二种情况, 必须应用截断技术 (truncation technique)。

$t > 1/\beta$ 情况的证明。 根据 $[\mathcal{H}_k]^t$ 的包含关系, 可以将 t 替换为 $\tilde{t} = \min(t, 2\tau)$, 因此现在有 $t \leq 2\tau$ 。将使用引理 3.12 中的 Bernstein 不等式。记

$$\xi(x) = T_\lambda^{-\frac{1}{2}} (K_x f^*(x) - T_x f_\lambda^*). \quad (3.97)$$

那么, 有

$$T_\lambda^{-\frac{1}{2}} [(\tilde{g}_X - \hat{T} f_\lambda^*) - (g - T f_\lambda^*)] = \frac{1}{n} \sum_{i=1}^n \xi(x_i) - \mathbb{E}_{x \sim \mu} \xi(x).$$

此外, 有

$$\mathbb{E} \|\xi(x)\|_{\mathcal{H}}^m = \mathbb{E} \left\| T_\lambda^{-\frac{1}{2}} K_x (f^*(x) - f_\lambda^*(x)) \right\|_{\mathcal{H}_k}^m$$

$$\leq \mathbb{E} \left[\left\| T_\lambda^{-\frac{1}{2}} k_x \right\|_{\mathcal{H}}^m \cdot \mathbb{E}(|f^*(x) - f_\lambda^*(x)|^m \mid x) \right]. \quad (3.98)$$

(3.98) 中的第一项通过 (3.50) 和引理 3.2 有界:

$$\left\| T_\lambda^{-\frac{1}{2}} k_x \right\|_{\mathcal{H}}^2 \leq M \mathcal{N}_1(\lambda) \leq C \lambda^{-1/\beta}.$$

对于第二项, 由于 $t > 1/\beta$, 利用嵌入性质命题 3.1 以及引理 3.3, 对于 $\alpha \in (1/\beta, t)$, 有

$$\|f_\lambda^* - f^*\|_{L^\infty} \leq M_\alpha \|f_\lambda^* - f^*\|_{[\mathcal{H}_k]^\alpha} \leq M_\alpha F_{(t-\alpha)/2} \|f^*\|_{[\mathcal{H}_k]^t} \lambda^{\frac{t-\alpha}{2}} \leq C M_\alpha R \lambda^{\frac{t-\alpha}{2}}.$$

此外, 引理 3.3 还意味着

$$\mathbb{E} |f_\lambda^*(x) - f^*(x)|^2 = \|f_\lambda^*(x) - f^*(x)\|_{L^2}^2 \leq F_{t/2}^2 \|f^*\|_{[\mathcal{H}_k]^t}^2 \lambda^t \leq C R^2 \lambda^t.$$

将这些估计代入 (3.98), 得到

$$\begin{aligned} (3.98) &\leq (C \lambda^{-1/(2\beta)})^m \cdot \|f^* - f^*\|_{L^\infty}^{m-2} \cdot \mathbb{E} |f_\lambda^*(x) - f^*(x)|^2 \\ &\leq (C \lambda^{-1/(2\beta)})^m \cdot \left(C M_\alpha R \lambda^{\frac{t-\alpha}{2}} \right)^{m-2} \cdot C R^2 \lambda^t \\ &\leq \frac{1}{2} m! (C R^2 \lambda^{t-1/\beta}) \left(C M_\alpha R \lambda^{\frac{t-\alpha-1/\beta}{2}} \right)^{m-2}. \end{aligned}$$

因此, 应用引理 3.12, 其中

$$\sigma^2 = C R^2 \lambda^{t-1/\beta}, \quad L = C M_\alpha R \lambda^{\frac{t-\alpha-1/\beta}{2}},$$

可得

$$\left\| T_\lambda^{-\frac{1}{2}} [(\tilde{g}_X - \hat{T} f_\lambda^*) - (g^* - T f_\lambda^*)] \right\|_{\mathcal{H}_k} \leq C R \sqrt{\frac{\lambda^{-1/\beta}}{n}} \left(1 + M_\alpha \sqrt{\frac{\lambda^{-\alpha}}{n}} \right) \lambda^{t/2}.$$

由于对于某个 $\theta < \beta$ 有 $\lambda = \Omega(n^{-\theta})$, 选择足够接近 $1/\beta$ 的 $\alpha > 1/\beta$ 即可得到所需结果。

$t \leq 1/\beta$ 情况的证明。 在这种情况下, 必须应用截断技术。为了估计由截断引起的额外误差项, 需要使用如下关于再生核希尔伯特空间的 L^q 嵌入命题^{[36]Theorem 5}。

命题 3.16: 在假设 3.2 下, 对于任何 $0 < s \leq \alpha_0$ 和 $\alpha > \alpha_0$, 有如下嵌入关系:

$$[\mathcal{H}_k]^s \hookrightarrow L^q(\mathcal{X}, d\mu), \quad q = \frac{2\alpha}{\alpha - s}. \quad (3.99)$$

现在, 考虑 $\Omega_B = \{x \in \mathcal{X} : |f^*(x)| \leq BR\}$ 以及

$$\bar{\xi}(x) = \xi(x) \mathbf{1}_{\Omega_B}(x),$$

其中 ξ 的定义见 (3.97), B 的选择将在后文确定。那么,

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n \xi(x_i) - \mathbb{E}\xi(x) \right\|_{\mathcal{H}_k} &\leq \left\| \frac{1}{n} \sum_{i=1}^n \bar{\xi}(x_i) - \mathbb{E}\bar{\xi}(x) \right\|_{\mathcal{H}_k} + \left\| \frac{1}{n} \sum_{i=1}^n \xi(x_i) \mathbf{1}_{\Omega_B^c}(x_i) \right\|_{\mathcal{H}_k} \\ &\quad + \left\| \mathbb{E}\xi(x) \mathbf{1}_{\Omega_B^c}(x) \right\|_{\mathcal{H}_k}. \end{aligned} \quad (3.100)$$

对于 (3.100) 中的第一项, 可以重复第一种情况中的相同论证, 并利用额外的界限:

$$\begin{aligned} \left\| \mathbf{1}_{\{x \in \Omega_r\}} (f_\lambda^* - f^*) \right\|_{L^\infty} &\leq \|f_\lambda^*\|_{L^\infty} + \left\| \mathbf{1}_{\{x \in \Omega_r\}} f^* \right\|_{L^\infty} \\ &\leq M_\alpha \|f_\lambda^*\|_{[\mathcal{H}_k]^\alpha} + BR \\ &\leq M_\alpha R \lambda^{\frac{t-\alpha}{2}} + BR, \end{aligned}$$

其中在最后一个不等式中应用了引理 3.3。因此得到

$$\left\| \frac{1}{n} \sum_{i=1}^n \bar{\xi}(x_i) - \mathbb{E}\bar{\xi}(x) \right\|_{\mathcal{H}_k} \leq CR \sqrt{\frac{\lambda^{-1/\beta}}{n}} \left(1 + \frac{M_\alpha \lambda^{-\alpha/2} + B \lambda^{-t/2}}{\sqrt{n}} \right) \lambda^{t/2}. \quad (3.101)$$

对于 (3.100) 中的第二项, 命题 3.16 结合 Markov 不等式可得

$$\mathbb{P}_\mu(\Omega_B^c) = \mathbb{P}_{x \sim \mu} \{ |f^*(x)| > BR \} \leq (BR)^{-q} \|f^*\|_{L^q}^q, \quad (3.102)$$

其中 $q = \frac{2\alpha}{\alpha-s}$ 。那么,

$$\mathbb{P} \{ x_i \in \Omega_B, i = 1, \dots, n \} = \left(1 - \mathbb{P}_\mu(\Omega_B^c) \right)^n \geq (1 - (BR)^{-q} \|f^*\|_{L^q}^q)^n.$$

因此, 只要满足

$$B^{-q} = o\left(\frac{1}{n}\right), \quad \text{或者等价地, } B = \omega(n^{1/q}), \quad (3.103)$$

就有 $\mathbb{P} \{ x_i \in \Omega_B, i = 1, \dots, n \} \rightarrow 1$, 从而当 n 足够大时, (3.100) 中的第二项以高概率消失。

对于 (3.100) 中的第三项,

$$\begin{aligned} \left\| \mathbb{E}\xi(x) \mathbf{1}_{\Omega_B^c}(x) \right\|_{\mathcal{H}_k} &\leq \mathbb{E} \left\| \xi(x) \mathbf{1}_{\Omega_B^c}(x) \right\|_{\mathcal{H}_k} \\ &= \mathbb{E} \left[\left\| \mathbf{1}_{\Omega_B^c}(x) (f^*(x) - f_\lambda^*(x)) \right\| \varphi_\lambda^{1/2}(T) k_x \right]_{\mathcal{H}_k} \\ &\stackrel{(a)}{\leq} C \lambda^{-1/(2\beta)} \mathbb{E} \left[\left\| \mathbf{1}_{\Omega_B^c}(x) (f^*(x) - f_\lambda^*(x)) \right\| \right] \\ &\stackrel{(b)}{\leq} C \lambda^{-1/(2\beta)} \|f^* - f_\lambda^*\|_{L^2} \mathbb{P}(\Omega_B^c)^{\frac{1}{2}} \end{aligned}$$

$$\stackrel{(c)}{\leq} C \lambda^{-1/(2\beta)} \lambda^{t/2} B^{-q/2} \|f^*\|_{L^q}^{q/2} = C \|f^*\|_{L^q}^{q/2} \lambda^{t/2} \sqrt{\lambda^{-1/\beta} B^{-q}}, \quad (3.104)$$

其中在 (a) 中应用了 (3.50), 在 (b) 中应用了 Cauchy-Schwarz 不等式, 在 (c) 中针对 L^2 范数应用了引理 3.3, 针对概率应用了 (3.102)。

最后, 由于 $\theta < \beta$, 可以选择 $\alpha \in (\beta^{-1}, \theta^{-1})$ 。由此可得

$$1 - \theta/\beta > 1 - \alpha\theta > 0,$$

可以选择某个 l 使得

$$\frac{1}{q} = \frac{1}{2}(1 - \frac{t}{\alpha}) < l < \frac{1 - \theta t}{2}.$$

现在令 $B = n^l$ 。那么 (3.103) 立即成立。对于项 (3.101), 有

$$\sqrt{\frac{\lambda^{-1/\beta}}{n}} = O(n^{-(1-\theta/\beta)/2}), \quad \frac{\lambda^{-\alpha/2}}{\sqrt{n}} = O(n^{-(1-\alpha\theta)/2}), \quad \frac{B\lambda^{-t/2}}{\sqrt{n}} = O(n^{-(\frac{1-\theta t}{2}-l)}),$$

对于项 (3.104), 有

$$\sqrt{\lambda^{-1/\beta} B^{-q}} = O(n^{-(ql-\theta/\beta)/2}) = O(n^{-(1-\theta/\beta)/2}).$$

由于所有的指数均为负值, 由此完成了证明。

3.5.6 集中不等式

如下给出了标准的有界随机变量的 Hoeffding 不等式, 参见 Wainwright^[97] 的命题 2.5。

引理 3.11 (Hoeffding 不等式): 设 $\xi, \xi_1, \dots, \xi_n$ 为独立同分布 (i.i.d.) 随机变量, 满足 $|\xi| \leq B$ 几乎处处成立。则对于任意 $\delta \in (0, 1)$, 至少以概率 $1 - \delta$ 有

$$\left| \frac{1}{n} \sum_{i=1}^n \xi_i - \mathbb{E}\xi \right| \leq \sqrt{\frac{2B^2}{n} \ln \frac{2}{\delta}}. \quad (3.105)$$

下面的引理给出了更为一般的 Bernstein 类型不等式, 只需要对应的随机元的 (高阶) 矩条件^[14]。

引理 3.12: 设 H 为可分希尔伯特空间。设 $\xi, \xi_1, \dots, \xi_n$ 为取值于 H 的独立同分布随机元。假设

$$\mathbb{E} \|\xi - \mathbb{E}\xi\|_H^m \leq \frac{1}{2} m! \sigma^2 L^{m-2}, \quad \forall m = 2, 3, \dots, \quad (3.106)$$

成立。则对于固定的 $\delta \in (0, 1)$, 有

$$\mathbb{P} \left\{ \left\| \frac{1}{n} \sum_{i=1}^n \xi_i - \mathbb{E}\xi \right\|_H \leq 2 \left(\frac{L}{n} + \frac{\sigma}{\sqrt{n}} \right) \ln \frac{2}{\delta} \right\} \geq 1 - \delta. \quad (3.107)$$

特别地, 满足 (3.106) 的一个充分条件是

$$\|\xi\|_H \leq \frac{L}{2} \text{ 几乎处处成立, 且 } \mathbb{E} \|\xi\|_H^2 \leq \sigma^2.$$

以下关于随机自伴 Hilbert-Schmidt 算子的 Bernstein 不等式在文献中常被使用 (例如 Li et al.^[89] 的引理 B.5), 它是 Tropp^[98] 的定理 7.7.1 的一个稍作修改的版本。

引理 3.13: 设 H 为可分希尔伯特空间。设 $A_1, \dots, A_n$ 为独立同分布随机元, 其取值为 H 上的自伴 Hilbert-Schmidt 算子, 满足 $\mathbb{E}(A_1) = 0$, 且对于某个 $L > 0$ 有 $\|A_1\| \leq L$ 几乎处处成立, 且对于某个正迹类算子 V 有 $\mathbb{E}(A_1^2) \leq V$ 。则对于任意 $\delta \in (0, 1)$, 至少以概率 $1 - \delta$ 有

$$\left\| \frac{1}{n} \sum_{i=1}^n A_i \right\| \leq \frac{2LB}{3n} + \left(\frac{2\|V\|B}{n} \right)^{1/2}, \quad \text{其中 } B = \ln \frac{4 \operatorname{Tr} V}{\delta \|V\|}.$$

第4章 神经网络的自适应特征模型

前两章分别研究了神经切线核的谱性质以及相应核方法泛化误差的精确刻画。这些结果一方面揭示了神经切线核方法在若干统计设定下的有效性，另一方面也表明：当问题具有更复杂的结构，尤其是在高维情形下，固定核方法往往会受到维度灾难的限制，难以解释真实神经网络所展现出的优异性能。造成这一差异的关键原因在于，神经切线核方法本质上基于训练初始时刻的线性化近似，其特征表示在训练过程中保持静态；而真实神经网络则会在训练中不断调整内部表示，从而主动适应数据结构。

基于这一观察，本章提出一个一般性的自适应特征模型 (adaptive feature model) 框架，通过引入参数化的特征映射，并联合优化特征映射与线性系数，来刻画神经网络动态学习特征的能力。为了衡量训练过程中所学特征映射与目标函数之间的匹配程度，本文进一步利用第3章中核方法泛化误差的精确刻画，提出特征误差度量 (feature error measure)。本章将给出这一框架的基本定义与理论动机，并说明它如何为后续的对角参数化模型 (第5章) 和方向参数化模型 (第6章) 提供统一的分析语言。这些结果将展示自适应特征模型在理解神经网络特征学习动力学及其对泛化影响方面的潜力。

4.1 自适应特征模型

虽然自适应特征学习已在各种背景下得到了探索^[17,62,65,99]，但仍缺乏一个捕捉其核心原理的统一框架。在作者此前综述工作^[100]的基础上，本章提出一个一般的自适应特征模型 (adaptive feature model)，将神经网络的动态学习能力整合到一个用于非参数回归的结构化统计框架中。

回顾非参数回归问题 (见第1.4节)。传统的非参数回归方法依赖于固定的特征映射 $\Phi: \mathcal{X} \rightarrow H$ ，将输入 x 转换为希尔伯特空间 H (如 $\ell^2(\mathbb{N})$) 中的特征表示。预测器定义为 $f(x) = \langle \Phi(x), \beta \rangle_H$ ，其中 $\beta \in H$ 是可训练系数。然而，选择有效的 Φ 具有挑战性，往往导致与 f^* 的对齐不佳以及性能低下。

在固定特征映射之外，本文提出了一个参数化特征映射 $\Phi_\theta: \mathcal{X} \rightarrow H$ ，其中 θ 是另一个可训练参数， H 是固定的希尔伯特空间。预测器变为

$$f(x) = \langle \Phi_\theta(x), \beta \rangle_H, \quad (4.1)$$

并采用梯度下降方法联合优化 θ 和 β 。回顾经验损失（或称样本损失）为

$$\hat{\mathcal{L}}_n = \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i))^2, \quad (4.2)$$

其中额外的 $1/2$ 系数是为了简化梯度表达式。自适应特征模型通过梯度下降（流）同时训练 θ 和 β ：

$$\begin{cases} \dot{\theta}_t = -\nabla_{\theta} \hat{\mathcal{L}}_n, \\ \dot{\beta}_t = -\nabla_{\beta} \hat{\mathcal{L}}_n, \end{cases} \quad (4.3)$$

其中 β 通常初始化为零，而 θ 的初始化取决于 Φ_{θ} 的参数化方式。

自适应特征模型允许特征映射 Φ_{θ} 在训练期间演化，发现与 f^* 对齐更紧密的表示。此外，这一过程是通过梯度下降自动完成的，不需要特定于问题的估计方法。这种动力学模拟了神经网络的行为，即网络权重形成的特征表示是通过训练隐式学习的。

参数化 Φ_{θ} 的灵活性使该方案能够整合多种模型。如果 Φ_{θ} 是固定的，则退化为标准的核梯度下降法^[38]。 Φ_{θ} 的其他实例包括过参数化线性回归、对角自适应核方法和方向自适应特征方法，这些将在后文中介绍。通过 Φ_{θ} 的各种参数化，自适应特征模型以统一的方式适应不同类型的数据结构。例如，在过参数化线性回归中，调整特征映射 Φ_{θ} 有助于识别稀疏信号分量。同样，在对角参数化自适应特征模型（第5章）中，模型调整核的谱权重，改善了特征映射与潜在函数 f^* 之间的对齐。此外，在方向参数化自适应特征模型（第6章）中，对于单指标模型，模型会学习相关方向以捕捉有效的投影结构。

虽然自适应性提供了明显的直观优势，但也为理论分析带来了新挑战。与由于静态结构而能够得到最优系数闭式解的固定特征方法不同，通过梯度下降同时演化的自适应方案通常缺乏解析解。此外，这种联合优化导致了非线性动力学，因为同步更新以复杂的、依赖数据的方式将特征映射和系数耦合在一起。此外，如果梯度下降无限期运行，可能会过拟合噪声训练数据，因此对早停（early stopping）时间的精细分析往往是必要的。

4.2 特征误差度量

面对一族特征映射 Φ_{θ} ，一个关键问题是：如何评价所学特征映射的有效性？为了解决这个问题，本章将引入特征误差度量（feature error measure），这是一种理论度量，旨在量化所学特征映射 Φ_{θ} 与目标函数 f^* 的对齐程度。考虑如下形式

的特征映射：

$$\Phi(x) = \left(\lambda_j^{\frac{1}{2}} e_j(x) \right)_{j \in N} \in \ell^2(N), \quad (4.4)$$

其中 $N \subseteq \mathbb{N}$ 是索引集, $\ell^2(N)$ 是平方可和序列空间, $\{e_j(\cdot)\}_{j \in N}$ 是 $L^2(\mathcal{X}, d\mu)$ 中的标准正交系 (不一定是基), $\lambda_j \geq 0$ 是权重。令 L_Φ^2 为由 $\{e_j\}_{j \in N}$ 张成的 $L^2(\mathcal{X}, d\mu)$ 子空间, 并用 $P_{L_\Phi^2} f^* = \sum_{j \in N} f_j^* e_j$ 表示 f^* 在该子空间上的投影, 其中 $f_j^* = \langle f^*, e_j \rangle_{L^2}$ 是 f^* 在该标准正交系中的系数。本文给出特征误差度量的如下定义：

定义 4.1： 特征误差度量 (关于 $\delta, \epsilon^2 > 0$ 的函数) 定义为：

$$\mathcal{E}(\delta, \epsilon^2; \Phi, f^*) = \mathcal{E}_{\text{Proj}}(\Phi, f^*) + \mathcal{E}_{\text{Stat}}(\delta, \epsilon^2; \Phi, f^*), \quad (4.5)$$

其中投影误差 $\mathcal{E}_{\text{Proj}}(\Phi, f^*)$ 和统计误差 $\mathcal{E}_{\text{Stat}}(\delta, \epsilon^2; \Phi, f^*)$ 分别由下式给出：

$$\begin{aligned} \mathcal{E}_{\text{Proj}}(\Phi, f^*) &:= \|f^* - P_{L_\Phi^2} f^*\|_{L^2(\mathcal{X}, d\mu)}^2, \\ \mathcal{E}_{\text{Stat}}(\delta, \epsilon^2; \Phi, f^*) &:= \mathcal{E}_V + \mathcal{E}_B := |\{j \in N : \lambda_j \geq \delta\}| \cdot \epsilon^2 + \sum_{j \in N} (f_j^*)^2 \mathbf{1}\{\lambda_j < \delta\}. \end{aligned}$$

特征误差度量 $\mathcal{E}(\delta, \epsilon^2; \Phi, f^*)$ 量化了特征映射 Φ 在逼近目标函数 f^* 时的理论误差, 其中参数 $\epsilon^2 > 0$ 定义了有效噪声水平, $\delta > 0$ 作为截断水平。它可分解为三个部分：

- **投影误差** $\mathcal{E}_{\text{Proj}}(\Phi, f^*)$ ：衡量特征映射 Φ 表示 f^* 的能力上限；
- **方差项** $\mathcal{E}_V$ ：通过显著分量的数量 (即 $\lambda_j \geq \delta$) 反映模型复杂度；
- **偏差项** $\mathcal{E}_B$ ：捕捉来自小权重特征 (即 $\lambda_j < \delta$) 的逼近误差。

通过引入 ϵ^2 , 而不是直接用样本量 n 来刻画误差, 这一理论量使本文能够从更一般的角度理解和分析特征映射的性能。同时, 它使本文能够将特征映射的训练与特征映射的评估分开, 从而更清晰地理解特征映射的表现。对于 n 个样本, ϵ^2 通常按 $\epsilon^2 \asymp \frac{1}{n}$ 缩放。

另一方面, δ 设置了一个阈值: 权重 $\lambda_j \geq \delta$ 的特征对模型的复杂度有贡献, 而低于该阈值的特征则被有效忽略。在早停梯度下降的背景下, 通常有 $\delta \asymp t^{-1}$, 其中 t 是训练时间——早停对应于较大的 δ , 限制了活跃特征的数量以防止过拟合。此外, 可以通过优化 δ 来最小化误差, 从而定义最优调节误差和最优截断水平：

$$\mathcal{E}^*(\epsilon^2; \Phi, f^*) := \inf_{\delta \geq 0} \mathcal{E}(\delta, \epsilon^2; \Phi, f^*), \quad \delta^*(\epsilon^2; \Phi, f^*) := \arg \min_{\delta \geq 0} \mathcal{E}(\delta, \epsilon^2; \Phi, f^*). \quad (4.6)$$

然而, 本文在此强调, 最优调节误差仍取决于特征映射 Φ , 且对于不同的特征映射, 其差异可能非常显著。

注 4.1： 总是可以通过使用正交补和零权重扩展特征映射, 将投影误差重新表述为统计误差中的偏差项。然而, 这种方法会影响特征映射原始形式的可解释性, 因

此本文希望在特征误差度量中保留投影误差。

特征误差度量量化了与给定特征映射 Φ 相关的泛化误差。特别是考虑到序列模型并假设 $(e_j)_{j \geq 1}$ 是一个基，它捕捉了与该特征映射相关的线性估计器^[90]的泛化误差。具体而言，误差度量 $\mathcal{E}_{\text{Stat}}(\delta, \epsilon^2; \Phi, f^*)$ 恰好对应于估计器 $\hat{f}_j = \mathbf{1}_{\{\lambda_j \geq \delta\}} z_j$ 的泛化误差。此外，考虑序列模型中的梯度下降估计器 $\hat{f}_t^{\text{Seq}}$ ，它也是一个具有闭式解 $\hat{f}_j = (1 - e^{-\lambda_j t}) z_j$ 的线性估计器。其泛化误差为：

$$\mathbb{E} \left\| f^* - \hat{f}_t^{\text{Seq}} \right\|_{L^2}^2 = \mathcal{E}_V^{\text{GD}}(t) + \mathcal{E}_B^{\text{GD}}(t), \quad \mathcal{E}_V^{\text{GD}}(t) = \frac{1}{n} \sum_{j \geq 1} (1 - e^{-\lambda_j t})^2, \quad \mathcal{E}_B^{\text{GD}}(t) = \sum_{j \geq 1} e^{-2\lambda_j t} (f_j^*)^2. \quad (4.7)$$

通过设置 $t = \delta^{-1}$ ，项 $\mathcal{E}_V^{\text{GD}}(t)$ 和 $\mathcal{E}_B^{\text{GD}}(t)$ 分别与 $\mathcal{E}_V$ 和 $\mathcal{E}_B$ 紧密接近。这些联系表明，特征误差度量是衡量特征映射 Φ 学习目标函数 f^* 质量的一个具有代表性的 Oracle 代理。

4.3 序列模型简化

自适应特征模型的动力学方程 (4.3) 往往是复杂且难以分析的。为了在专注于自适应特征动力学的同时保留非参数回归的本质，本章进一步提出考虑高斯序列模型^[90]的简化。假设给定 $L^2(\mathcal{X}, d\mu)$ 中一个固定的标准正交基 $(\phi_\ell)_{\ell \in \mathcal{I}}$ ，其中 $\mathcal{I}$ 是索引集。文献^[90,92-93,100]已经观察到，在非参数回归问题 (1.3) 中观测 n 个独立同分布的样本实际上近似于在序列模型中观测到序列 $(z_\ell)_{\ell \in \mathcal{I}}$ ：

$$z_\ell = f_\ell^* + \xi_\ell, \quad f_\ell^* = \langle f^*, \phi_\ell \rangle_{L^2(\mathcal{X}, d\mu)}, \quad \xi_\ell \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2/n), \quad \ell \in \mathcal{I}. \quad (4.8)$$

这里， f_ℓ^* 代表目标函数 f^* 在标准正交基中的真实系数， ξ_ℓ 是噪声项（在不同索引之间相互独立）， σ^2 是方差参数。本文指出，噪声的方差 σ^2/n 随样本量缩放，反映了对 n 个样本取平均的效果。

设 f 为候选函数。由于 $(\phi_\ell)_{\ell \in \mathcal{I}}$ 是标准正交基，总体损失（超额风险）可以写为 $\|f - f^*\|_2^2 = \sum_{\ell \in \mathcal{I}} (f_\ell - f_\ell^*)^2$ ，其中 $f_\ell = \langle f, \phi_\ell \rangle_{L^2}$ 。因此，本文相应地在序列模型中定义序列模型损失（简称为“序列损失”）。严格地说，当索引集 $\mathcal{I}$ 为无限集时，带有独立噪声的全体观测序列通常不属于 $\ell^2(\mathcal{I})$ ，直接对所有坐标求和的随机损失可能不是有限值。因此，可以先在有限截断集合 $\mathcal{I}_M \subset \mathcal{I}$ 上定义良定的截断损失，再令 $M \rightarrow \infty$ 。本文为简化记号，仍使用如下形式表示这一截断极限或形式化等价的序列模型代理：

$$\bar{L}_n(f) = \frac{1}{2} \sum_{\ell \in \mathcal{I}} (f_\ell - z_\ell)^2, \quad f_\ell = \langle f, \phi_\ell \rangle_{L^2}, \quad (4.9)$$

其梯度和动力学均可理解为有限截断问题的坐标形式，再在所需的收敛条件下取

极限。这是非参数回归问题中的经验损失 $\hat{\mathcal{L}}_n$ 的序列模型近似。此时，可以考虑在序列损失 $\bar{\mathcal{L}}_n$ 下的自适应特征模型动力学：

$$\begin{cases} \frac{d}{dt}\bar{\theta}_t = -\nabla_{\theta}\bar{\mathcal{L}}_n, \\ \frac{d}{dt}\bar{\beta}_t = -\nabla_{\beta}\bar{\mathcal{L}}_n. \end{cases} \quad (4.10)$$

那么，可以期待，(4.10) 所描述的自适应特征模型动力学能够近似地捕捉原始非参数回归问题中自适应特征模型的行为。从而可以通过分析序列模型中的自适应特征模型来获得对原始问题的理解。

从有限样本到序列模型观测的转变不仅有所谓的“Le Cam 等价性”^[92-93]的佐证，这一点也得到了最近关于具有固定特征映射的梯度下降泛化误差研究^[95]的支持，见后面的第 4.3.1 小节。

此外，对于自适应特征模型的其他实例，也可以经验性地观察到经验损失和序列损失下预测器之间的接近性。如图 4.2 所示，只要训练时间保持在一定范围内，两个预测器之间的差距就会随着样本量的增加而消失。这两种动力学的相似性使得序列损失下的自适应特征模型可以被视为一种有效的近似。本文假设，这两种训练过程之间强大的“路径等价性”在很大一类自适应特征模型中普遍成立，这超出了本文目前工作的焦点，但将是一个有趣的未来研究方向。本文用一个不严格的猜想来总结这一观察：

猜想 4.1 (非参数回归与序列模型的路径等价性)： 考虑非参数回归问题中的自适应特征模型动力学 (4.3)，以及其在序列损失 (4.9) 下的对应动力学 (4.10)。设 $\hat{f}_t$ 和 $\bar{f}_t$ 分别为这两种动力学在时间 t 时刻的估计量。在适当的参数化条件下，对于某一类测试泛函 $G(\Phi, \beta) \in \mathcal{G}$ ，对于任何固定的训练时间 $T > 0$ ，有

$$\sup_{t \in [0, T]} \sup_{G \in \mathcal{G}} \left| \mathbb{E}G(\Phi_{\theta_t}, \beta_t) - \mathbb{E}G(\Phi_{\bar{\theta}_t}, \bar{\beta}_t) \right| \xrightarrow{n \rightarrow \infty} 0, \quad (4.11)$$

以下，本章将通过固定核回归和数值实验来说明这种路径等价性。

4.3.1 核方法下的例证

现在考虑特征映射 Φ 固定、不依赖于参数 θ 的情形。此时，自适应特征模型 (4.3) 退化为核回归中的梯度流（梯度下降）方法（见例子 3.6）。固定 $L^2(\mathcal{X}, d\mu)$ 中的标准正交基 $(e_j)_{j \geq 1}$ ，考虑

$$\Phi(x) = \left(\lambda_j^{\frac{1}{2}} e_j(x) \right)_{j \geq 1} : \mathcal{X} \rightarrow \ell^2, \quad (4.12)$$

其中 $\lambda_j \geq 0$ 是一组权重, 满足 $\sum_{j=1}^{\infty} \lambda_j < \infty$ 。利用 $f_t(x) = \langle \Phi(x), \beta_t \rangle_{\ell^2}$, 此时 (4.3) 所给出的参数训练过程为

$$\dot{\beta}_t = -\nabla_{\beta} \hat{\mathcal{L}}_n = -\frac{1}{n} \sum_{i=1}^n (f_t(x_i) - y_i) \Phi(x_i), \quad (4.13)$$

从而给出函数值的梯度流:

$$\begin{aligned} \frac{d}{dt} f_t(x) &= \langle \Phi(x), \dot{\beta}_t \rangle_{\ell^2} \\ &= -\frac{1}{n} \sum_{i=1}^n (f_t(x_i) - y_i) \langle \Phi(x), \Phi(x_i) \rangle_{\ell^2}. \end{aligned}$$

如果考虑对应的核函数 $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\ell^2}$, 那么上式可以写为

$$\frac{d}{dt} f_t(x) = -\frac{1}{n} \sum_{i=1}^n (f_t(x_i) - y_i) k(x, x_i). \quad (4.14)$$

比较由核函数 $k(x, x')$ 所诱导的核回归梯度流, 回顾例子 3.6 中的 (3.27):

$$\dot{f}_t = -\nabla_f^{\mathcal{H}_k} \hat{\mathcal{L}}_n(f_t) = -\frac{1}{n} \sum_{i=1}^n (f_t(x_i) - y_i) k_{x_i} \in \mathcal{H}_k,$$

可以发现这两个梯度流在函数值上是完全相同的。因此, 在固定特征映射的情况下, 自适应特征模型 (4.3) 与核回归中的梯度流是等价的。进一步地, 可以将核回归中的梯度流与序列模型中的梯度流联系起来。

为了进一步说明序列模型与非参数回归模型之间的联系, 首先将再生核希尔伯特空间中的梯度流 (4.13) 写成矩阵形式。定义 n 个样本点的观测向量 $\mathbf{y} = (y_1, \dots, y_n)^\top$, 以及特征矩阵 $\Phi(X) = (\Phi(x_1), \dots, \Phi(x_n))_{\infty \times n}$ 。那么, 展开 $f_t(x) = \Phi(x)^\top \beta_t$, 梯度流 (4.13) 可以写作:

$$\dot{\beta}_t = -\frac{1}{n} \Phi(X) \Phi(X)^\top \beta_t + \frac{1}{n} \Phi(X) \mathbf{y}. \quad (4.15)$$

由于 $\frac{1}{n} \Phi(X) \Phi(X)^\top$ 的第 j, l 个分量为

$$\left(\frac{1}{n} \Phi(X) \Phi(X)^\top \right)_{j,l} = \frac{\lambda_j^{\frac{1}{2}} \lambda_l^{\frac{1}{2}}}{n} \sum_{i=1}^n e_j(x_i) e_l(x_i),$$

并且 $(e_j)_{j \geq 1}$ 是 $L^2(\mathcal{X}, d\mu)$ 中的标准正交基, 大数定律表明当样本量 n 足够大时,

$$\frac{1}{n} \Phi(X) \Phi(X)^\top \approx \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots),$$

也就是说该梯度流可以在特征基下实现对角化。不仅如此, 假设 $f^*(x) = \langle \Phi(x), \beta^* \rangle$,

可以展开

$$\begin{aligned}\frac{1}{n}\Phi(X)\mathbf{y} &= \frac{1}{n}\sum_{i=1}^n y_i \Phi(x_i) = \frac{1}{n}\sum_{i=1}^n (f^*(x_i) + \varepsilon_i) \Phi(x_i) \\ &= \frac{1}{n}\Phi(X)\Phi(X)^\top \boldsymbol{\beta}^* + \frac{1}{n}\sum_{i=1}^n \varepsilon_i \Phi(x_i).\end{aligned}$$

其中的第一项近似为 $\Lambda \boldsymbol{\beta}^*$ ，而第二项的第 j 个分量为

$$\frac{\lambda_j^{\frac{1}{2}}}{n} \sum_{i=1}^n e_j(x_i) \varepsilon_i,$$

中心极限定理表明它可以近似为 $\lambda_j^{\frac{1}{2}} \xi_j$ ，其中 ξ_j 是一个均值为零、方差为 σ^2/n 的正态随机变量。于是

$$\frac{1}{n}\Phi(X)\mathbf{y} \approx \Lambda \boldsymbol{\beta}^* + \Lambda^{\frac{1}{2}} \boldsymbol{\xi} = \Lambda^{\frac{1}{2}} \mathbf{z}.$$

把上述近似代入 (4.15)，得到如下近似的梯度流，它可以写成独立分量的形式：

$$\dot{\beta}_j(t) \approx -\lambda_j \beta_j + \lambda_j^{\frac{1}{2}} z_j = -\lambda_j (\beta_j(t) - \lambda_j^{-\frac{1}{2}} z_j). \quad (4.16)$$

另一方面，考虑在序列损失 (4.9) 下的梯度流。对于 $f(x) = \langle \Phi(x), \boldsymbol{\beta} \rangle_{\ell^2}$ ，有

$$f_j := \langle f, e_j \rangle_{L^2} = \lambda_j^{\frac{1}{2}} \beta_j.$$

此时序列损失的梯度为

$$\nabla_{\beta_j} \bar{\mathcal{L}}_n = \frac{\partial \bar{\mathcal{L}}_n}{\partial f_j} \cdot \frac{\partial f_j}{\partial \beta_j} = (f_j - z_j) \cdot \lambda_j^{\frac{1}{2}} = \lambda_j^{\frac{1}{2}} \left(\lambda_j^{\frac{1}{2}} \beta_j - z_j \right).$$

于是，序列损失下的演化方程 (4.10) 变成

$$\frac{d}{dt} \bar{\beta}_j(t) = -\lambda_j (\bar{\beta}_j(t) - \lambda_j^{-\frac{1}{2}} z_j), \quad (4.17)$$

发现它与上面从核梯度流中得到的近似梯度流 (4.16) 完全相同。不仅如此，继续得到

$$\frac{d}{dt} \bar{f}_j(t) = \lambda_j^{\frac{1}{2}} \frac{d}{dt} \bar{\beta}_j(t) = -\lambda_j (\bar{f}_j(t) - z_j). \quad (4.18)$$

这是一个独立分量的线性微分方程，其解为

$$\bar{f}_j(t) = (1 - e^{-\lambda_j t}) z_j = [\lambda_j \varphi_\lambda^{\text{GF}}(\lambda_j)] z_j, \quad \lambda = t^{-1}, \quad (4.19)$$

其中 $\varphi_\lambda^{\text{GF}}$ 正是例子 3.6 中梯度流的滤谱函数。这也说明，在固定特征映射的情况下，序列损失下的梯度流给出的估计量，正是序列模型中对应于梯度流方法的收缩估计量 (3.32)。

最后，本文在前一章中得到的“序列模型等价性”，定理 3.1，事实上说明了，

在一般的假设下，核梯度流与序列模型中的梯度流在泛化误差上是等价的，并且这个等价性极其精确。由此，可以期待，在适当的假设下，不仅可以得到泛化误差的等价性，还可以得到更强的路径等价性。

4.3.2 数值验证

本节进行进一步的数值实验，以说明经验损失和序列损失下自适应特征模型的路径等价性。首先，图 4.1 展示了对角自适应特征模型（见第 5 章）和方向自适应特征模型（见第 6 章）中训练过程中特征误差度量（FEM）的演化。可以看到，在经验损失和序列损失下，特征误差度量随着训练的进行而减小，两者表现相似。

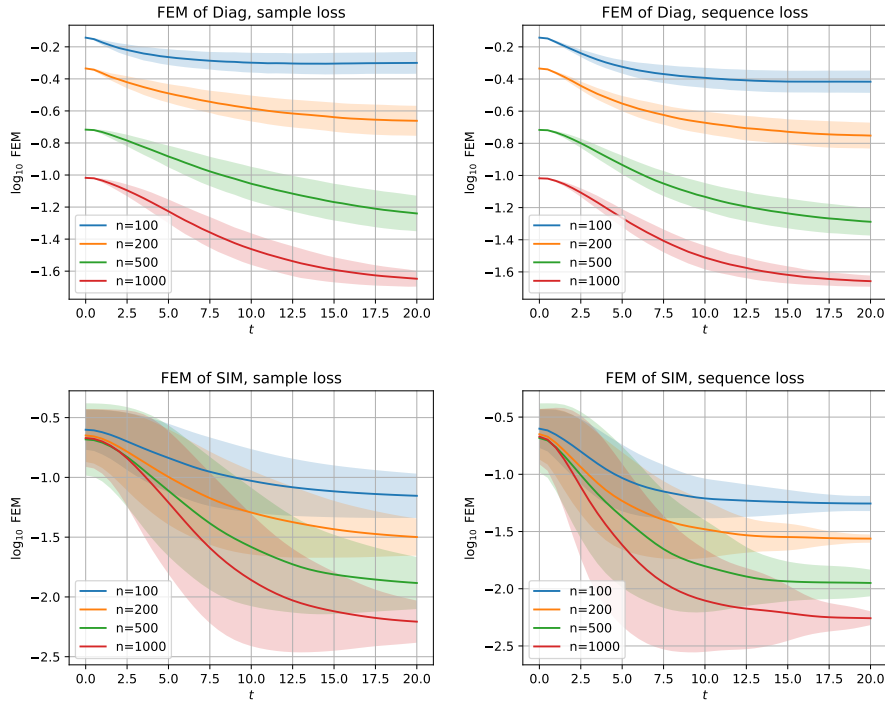


图 4.1 训练过程中特征误差度量 $\mathcal{E}^*$ 的下降。上行：对角自适应特征（Diag）；下行：单指标模型（SIM）的方向自适应特征。左列：经验损失；右列：序列损失。阴影区域表示由 200 次运行计算出的标准差。

在图 4.2 中，通过基于 L^2 范数的能量距离，度量了两个函数分布之间的距离，展示了在经验损失和序列损失下训练的预测器演化过程的相似性。图 4.3 展示了两种损失下特征误差度量（FEM）分布之间的能量距离。这些结果表明，随着样本量的增加，经验损失和序列损失下的自适应特征模型的训练路径变得越来越相似，这支持了本文关于路径等价性的猜想。

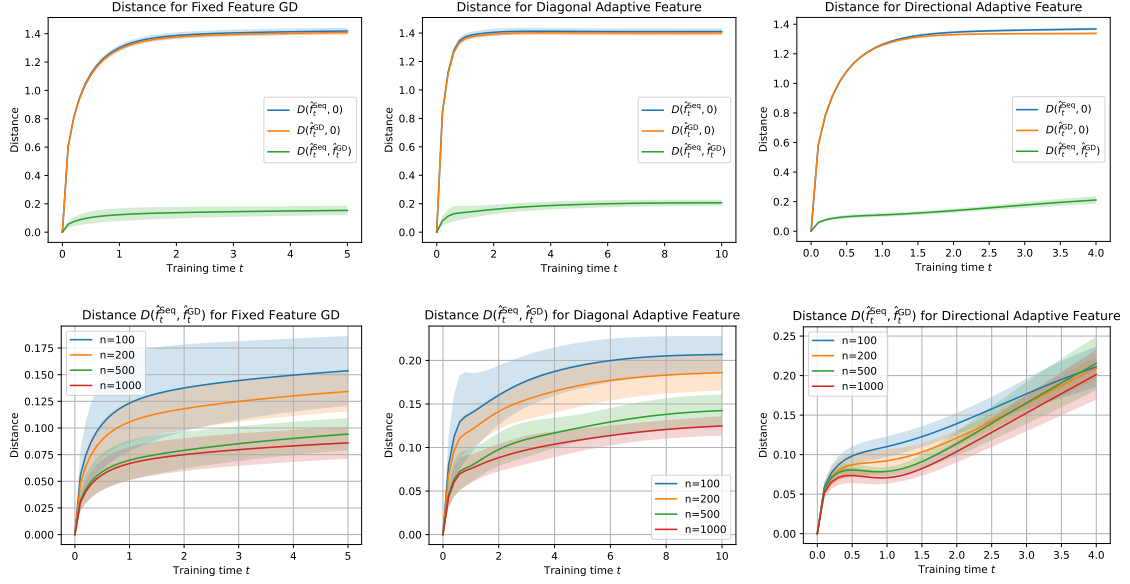


图 4.2 经验损失 $\mathcal{L}_n$ 和序列损失 $\bar{\mathcal{L}}_n$ 下训练曲线的相似性。绘制了由 200 次独立运行估计出的能量距离，阴影区域表示通过自助法（bootstrapping）估计得到的标准差。上行：在训练路径上， $D(\hat{f}_t^{\text{Seq}}, \hat{f}_t^{\text{GD}})$ 远小于 $D(\hat{f}_t^{\text{Seq}}, 0)$ 和 $D(\hat{f}_t^{\text{GD}}, 0)$ 。下行： $\hat{f}_t^{\text{GD}}$ 和 $\hat{f}_t^{\text{Seq}}$ 之间的差异随着 n 的增加而减小。三列中的方法分别为固定特征方法、对角自适应核方法和方向自适应特征方法。

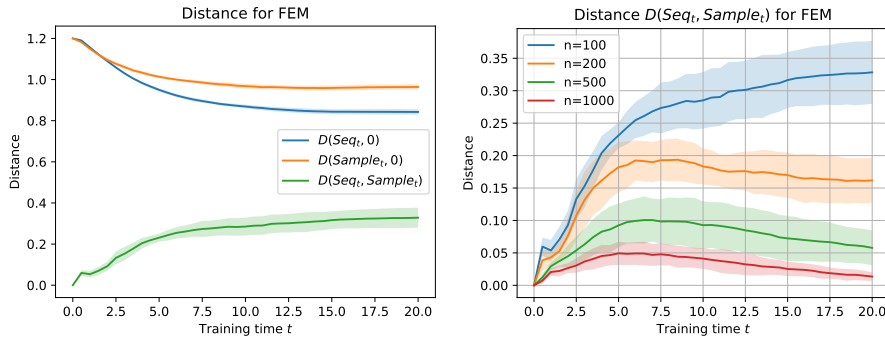


图 4.3 经验损失 $\mathcal{L}_n$ 和序列损失 $\bar{\mathcal{L}}_n$ 下特征误差度量 $\mathcal{E}^*$ 之间的能量距离。

第 5 章 对角参数化模型的自适应特征理论

本章研究对角参数化模型中的自适应特征学习机制。在这一模型中，特征函数保持固定，而对应的特征值权重可以在训练过程中动态调整；因此，它为理解“特征如何在梯度下降下被重新加权”提供了一个结构清晰、又足够非平凡的理论对象。第 5.1 节将回顾固定核回归的局限性，以说明为何需要自适应方法。随后，第 5.2 节将介绍对角参数化模型的设定。接着，第 5.3 节将陈述泛化能力的主要结果，而第 5.4 节将给出特征映射改进的结果。第 5.5 节将展示数值实验以验证理论发现。余下的章节将给出这些结果的证明细节。

5.1 固定核回归的局限性

仍然考虑第 3 章中的非参数回归问题设置。设 $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 为连续正定核， $\mathcal{H}_k$ 为其相关的再生核希尔伯特空间（RKHS）。核函数的 Mercer 分解为：

$$k(x, y) = \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(y), \quad (5.1)$$

其中 $\{\lambda_j\}_{j \geq 1}$ 是非负特征值的降序列， $\{e_j\}_{j \geq 1}$ 是对应的正交归一化特征函数基底。利用 Mercer 分解，可以给出核函数所对应的特征映射的显式形式：

$$\Phi(x) = (\lambda_j^{\frac{1}{2}} e_j(x))_{j \geq 1} : \mathcal{X} \rightarrow \ell^2, \quad (5.2)$$

它满足

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\ell^2}. \quad (5.3)$$

此外，核函数所对应的再生核希尔伯特空间可以表示为：

$$\mathcal{H}_k = \left\{ f : \mathcal{X} \rightarrow \mathbb{R} \mid f(x) = \sum_{j=1}^{\infty} \beta_j \lambda_j^{\frac{1}{2}} e_j(x), (\beta_j)_{j \geq 1} \in \ell^2 \right\}, \quad (5.4)$$

由此，利用特征映射，函数 $f \in \mathcal{H}_k$ 可以表示为

$$f(x) = \langle \Phi(x), \beta \rangle_{\ell^2}, \quad \beta \in \ell^2.$$

固定核的梯度下降估计量是自适应核回归框架 (4.3) 中的基线方法，对应于不进行特征学习的平凡情形。回顾第 4.3.1 小节，取 $\Phi_\theta \equiv \Phi$ ，参数训练由 (4.13) 给

出, 即

$$\dot{\beta}_t = -\nabla_{\beta} \mathcal{L} = \frac{1}{n} \sum_{i=1}^n (y_i - \langle \Phi(x_i), \beta_t \rangle_{\ell^2}) \Phi(x_i), \quad \text{其中 } \beta_0 = \mathbf{0}, \quad (5.5)$$

而所对应的估计量 $\hat{f}_t^{\text{GF}}$ 为

$$\hat{f}_t^{\text{GF}}(x) = \langle \Phi(x), \beta_t \rangle_{\ell^2},$$

它所满足的演化方程为 (4.14)。这种核梯度下降 (流) 估计量也对应于无限宽极限下的神经网络, 见第 2.1 节。

利用本文已在第 3 章中讨论的核回归理论以及序列模型等价性, 可以完全给出固定核回归的泛化性能。假设回归函数 f^* 在特征函数基底下的展开为

$$f^* = \sum_{j=1}^{\infty} \theta_j^* e_j.$$

那么, 根据定理 3.1 以及 (3.37), 上述核梯度下降估计量的泛化误差为

$$\mathcal{R}_{\text{Ker}}(\hat{f}_{\lambda}, f^*) = (1 + o_{\mathbb{P}}(1)) \mathcal{R}_{\text{Seq}}(\hat{\theta}_{\lambda}, \theta^*), \quad (5.6)$$

其中

$$\begin{aligned} \mathcal{R}_{\text{Seq}}(\hat{\theta}_{\lambda}, \theta^*) &= \sum_{j=1}^{\infty} \left[\psi_{\lambda}^{\text{GF}}(\lambda_j) \right]^2 (\theta_j^*)^2 + \frac{\sigma^2}{n} \sum_{j=1}^{\infty} \left[\lambda_j \varphi_{\lambda}^{\text{GF}}(\lambda_j) \right]^2, \\ \varphi_{\lambda}^{\text{GF}}(z) &= \frac{1 - e^{-tz}}{z}, \quad \psi_{\lambda}^{\text{GF}}(z) = e^{-tz}, \quad t = \lambda^{-1}. \end{aligned}$$

本文将利用上述解析表达式, 在下面的若干个例子中计算固定核回归的收敛速率, 以说明固定核回归的局限性。

5.1.1 不同衰减速率的特征值

在核回归中, 通常的假设是回归函数 f^* 属于再生核希尔伯特空间 $\mathcal{H}_k$ 。更一般的假设是 f^* 属于 $\mathcal{H}_k$ 的某个插值空间^[35,96]。基于 Mercer 分解, $\mathcal{H}_k$ 的 $s \geq 0$ 阶插值空间可以表达为

$$[\mathcal{H}_k]^s := \left\{ \sum_{j=1}^{\infty} \beta_j \lambda_j^{\frac{s}{2}} e_j \mid (\beta_j)_{j \geq 1} \in \ell^2 \right\}, \quad (5.7)$$

其配备范数

$$\|f\|_{[\mathcal{H}_k]^s} = \|\beta\|_{\ell^2}, \quad f = \sum_{j=1}^{\infty} \beta_j \lambda_j^{\frac{s}{2}} e_j.$$

特别地, 插值空间 $[\mathcal{H}_k]^1$ 对应于再生核希尔伯特空间 $\mathcal{H}_k$ 本身。那么, 假设特征值衰减率为 $\lambda_j \asymp j^{-\gamma}$, 核回归中的经典结果 (参见例如 Yao et al.^[38], Li et al.^[95]) 表

明如下的极小极大最优收敛速率：

$$\inf_{\hat{f}} \sup_{f^* \in [\mathcal{H}_k]^s, \|f^*\|_{[\mathcal{H}_k]^s} \leq 1} \mathbb{E} \mathcal{R}_{\text{Ker}}(\hat{f}, f^*) \asymp n^{-\frac{s\gamma}{s\gamma+1}}. \quad (5.8)$$

不仅如此，(5.6) 也能够说明，如果回归函数 f^* “恰好” 属于 $[\mathcal{H}_k]^s$ ，那么核梯度下降估计量 $\hat{f}_t^{\text{GF}}$ 刚好达到该收敛速率，既不会表现得更好，也不会更差。

但是，再生核希尔伯特空间 $\mathcal{H}_k$ 乃至其插值空间 $[\mathcal{H}_k]^s$ 与核函数本身密切相关，也与核函数的特征分解紧密联系。所以，核回归方法的泛化性能在很大程度上取决于核的特征分解以及目标函数在基底下的展开的系数的性质。换句话说，核的选择应当会显著影响核方法的性能。为了定量地说明这一点，考虑以下示例。

众所周知，具有特定对称性的核（例如球面上的点积核或环面上的平移不变周期核）共享同一组特征函数（例如球面谐波函数或傅里叶基）。由此，可以考虑具有相同特征函数但不同特征值衰减率的核函数。固定一组特征函数 $\{e_j\}_{j \geq 1}$ 和真实函数 f^* 。假设有两个核函数 k_1 和 k_2 ，它们的特征函数均为 $\{e_j\}_{j \geq 1}$ ，而它们的特征值衰减率分别为 $\lambda_{j,1} \asymp j^{-\gamma_1}$ 和 $\lambda_{j,2} \asymp j^{-\gamma_2}$ 。可以得出：

$$f^* \in [\mathcal{H}_{k_1}]^{s_1} \iff f^* \in [\mathcal{H}_{k_2}]^{s_2} \quad \text{当且仅当} \quad \gamma_1 s_1 = \gamma_2 s_2.$$

鉴于收敛速率仅取决于乘积 $s\gamma$ ，相对于这两个核的收敛速率将完全相同。

注 5.1： 核函数的 Mercer 分解所给出的特征值衰减速率 $\lambda_j \asymp j^{-\gamma}$ 与其所对应的再生核希尔伯特空间 $\mathcal{H}_k$ 的正则性密切相关，它也与再生核希尔伯特空间的有效维度（effective dimension）或容量条件（capacity condition）有着密切联系^[14]。特别地，如果考虑光滑区域 $\Omega \subset \mathbb{R}^d$ 上的 Sobolev 空间 $H^r(\Omega)$ ，其中 $r > d/2$ ，则这个空间是一个再生核希尔伯特空间，与该空间相关联的核函数的特征值通常以 $\lambda_j \asymp j^{-2r/d}$ 衰减^[35]。因此， γ 表示再生核希尔伯特空间 $\mathcal{H}_k$ 中函数的“光滑度”。另一方面，插值空间 $[\mathcal{H}_k]^s$ 的指标 s 则代表了其中函数相对于 $\mathcal{H}_k$ 本身的额外正则性，它随着 s 的增加而增加，因此 s 也可以被视为“相对光滑性”的量度。以 Sobolev 空间为例，已知 $[\mathcal{H}_k]^s \cong H^{rs}(\Omega)$ ，因此 s 衡量了函数相对于基准空间 $\mathcal{H}_k = H^r(\Omega)$ 的相对光滑度。两者相结合，乘积 $s\gamma$ 则表示了函数的绝对光滑度。因此，上述结果实际上说明，当特征函数及其排序固定时，核回归的收敛速率仅取决于目标函数的绝对光滑度，而与核本身的特征值衰减速率无关。

上述结果似乎表明，当特征函数固定时，核回归方法在不同核上的表现相似。然而，需要注意的是，这种相似性源于两个核具有相同的特征值衰减阶数，这与预先设定的基底顺序相一致。事实上，如果真实函数的衰减顺序与核的特征值衰减顺序相偏离，即使特征函数基底保持不变，也可能导致显著不同的收敛速率。本章将在后面的两个子节中说明这一点。

5.1.2 低维结构

考虑环面 $\mathbb{T}^d = [-1, 1]^d$ 上的平移不变周期核，其特征函数由傅里叶基 $\phi_{\mathbf{m}}(\mathbf{x}) = e^{i\pi\langle \mathbf{m}, \mathbf{x} \rangle}$ 给出，其中 $\mathbf{m} \in \mathbb{Z}^d$ 。在此基底上，某个目标函数 $f^*(\mathbf{x})$ 可以表示为

$$f^* = \sum_{\mathbf{m} \in \mathbb{Z}^d} f_{\mathbf{m}} \phi_{\mathbf{m}}(\mathbf{x}).$$

如果假设 f^* 具有某种低维结构，那么其傅里叶系数 $f_{\mathbf{m}}$ 也会具有一定的额外结构。

例如，假设目标函数 f^* 仅依赖于 d_0 个变量，具体表现为 $f^*(\mathbf{x}) = g(x_1, \dots, x_{d_0})$ （对于某些 $d_0 < d$ ），并假设 g 属于 t 阶 Sobolev 空间 $H^t(\mathbb{T}^{d_0})$ 。此时，可以计算得到，系数 $f_{\mathbf{m}}$ 简化为：

$$f_{\mathbf{m}} = \begin{cases} g_{\mathbf{m}_1}, & \mathbf{m} = (\mathbf{m}_1, \mathbf{0}), \mathbf{m}_1 \in \mathbb{Z}^{d_0}, \\ 0, & \text{其余情况.} \end{cases}$$

现在，考虑平移不变周期核，它们的特征函数都由傅里叶基给出，所以只需关注特征值。考虑如下两组特征值所对应的核函数 k_1 和 k_2 ： k_1 由 $\lambda_{\mathbf{m},1} = (1 + \|\mathbf{m}\|^2)^{-r}$ 给出（对于某些 $r > d/2$ ），其所对应的再生核希尔伯特空间是全空间上的 Sobolev 空间 $H^r(\mathbb{T}^d)$ ； k_2 由 $\lambda_{\mathbf{m},2} = (1 + \|\mathbf{m}\|^2)^{-r}$ （当 $\mathbf{m} = (\mathbf{m}_1, \mathbf{0})$ 时）给出，其余情况 $\lambda_{\mathbf{m},2} = 0$ ，它所对应的再生核希尔伯特空间是 d_0 维子空间上的 Sobolev 空间 $H^r(\mathbb{T}^{d_0})$ 的扩展。

那么，上述目标函数 f^* 同时属于 $[\mathcal{H}_{k_1}]^s$ 和 $[\mathcal{H}_{k_2}]^s$ （其中 $s = t/r$ ）。在将特征值按降序重新排列后，这两个核的衰减率分别为 $\gamma_1 = 2r/d$ 和 $\gamma_2 = 2r/d_0$ 。因此，相对于这两个核的收敛速率分别为：

$$\frac{2t}{2t+d} \quad \text{与} \quad \frac{2t}{2t+d_0}.$$

由此可见，当 d 显著大于 d_0 时，第二个核的收敛速率明显优于第一个核。

该示例说明，即使特征函数相同，特征值也会显著影响收敛速率。在上述场景中，第二个核通过仅关注相关维度，从目标函数的低维结构中受益；而第一个核由于考虑了所有维度，从而受到了“维度灾难”的影响。从这个例子中可以得到的关键点是核与目标函数之间的对齐（alignment）。为了推广这一例子，可以考虑以下特征值顺序与目标函数系数顺序不一致的情况。

5.1.3 错位结构

固定一组特征函数 $(e_j)_{j \geq 1}$ ，并将真实函数展开为 $f^* = \sum_{j \geq 1} \theta_j^* e_j$ 。注意，通过给定 $(e_j)_{j \geq 1}$ ，已经定义了基底在 j 上的顺序，但真实函数的系数 θ_j^* 并不一定随 j 有序。假设索引序列 $\ell(j)$ 给出了 $|\theta_{\ell(j)}^*|$ 的降序排列，即 $|\theta_{\ell(1)}^*| \geq |\theta_{\ell(2)}^*| \geq \dots$ ，那

么可以通过 $\ell(j)$ 与 j 之间的差异来刻画这种“错位”。具体而言, 假设:

$$\left| \theta_{\ell(j)}^* \right| \asymp j^{-(p+1)/2} \quad \text{且} \quad \ell(j) \asymp j^q \quad \text{对于} \quad p > 0, q \geq 1, \quad (5.9)$$

其中较大的 q 表示更严重的错位。在特征值方面, 考虑 $\lambda_{j,1} \asymp j^{-\gamma}$ (其顺序与 j 一致), 而 $\lambda_{\ell(j),2} \asymp j^{-\gamma}$ (其顺序与 $\ell(j)$ 一致)。那么, 对应这两组系数序列的收敛速率分别为:

$$\frac{p}{p+q} \quad \text{与} \quad \frac{p}{p+1}.$$

因此, 如果错位很严重 (即当 q 较大时), 收敛速率可能会有巨大差异。

第 5.1.2 小节和第 5.1.3 小节说明了核的特征值与目标函数的结构相对齐是非常有益的。然而, 先验地选择合适的核通常很困难, 特别是当目标函数的结构未知时, 因此固定核回归可能会受到核本身的限制而表现不佳。受这些示例的启发, 本章希望探索“自适应特征”的想法, 即可以从数据中学习核。

5.2 对角参数化模型

受上一节示例的启发, 作为迈向自适应特征的第一步, 本章考虑在特征函数固定但核特征值可变的情况下的自适应特征方法。对应于第 4.1 节所述的自适应特征模型的框架 (4.3), 本章需要定义一个参数化的特征映射 Φ_a , 其中 a 参数化了核的特征值。

为此, 首先回顾固定核特征映射 (5.2) 以及核梯度流 (5.5)。为了学习特征值, 本章考虑由 $a = (a_j)_{j \geq 1}$ 参数化的特征映射及其对应的核:

$$\Phi_a(x) = (a_j e_j(x))_{j \geq 1}, \quad k_a(x, x') = \sum_{j \geq 1} a_j^2 e_j(x) e_j(x'). \quad (5.10)$$

对于预测器, 类似地考虑 $f(x) = \langle \Phi_a(x), \beta \rangle_{\ell^2}$, 其中 $\beta = (\beta_j)_{j \geq 1} \in \ell^2$ 是特征映射下的系数向量。然后, 考虑同时训练 a 和 β 的梯度流:

$$\begin{cases} \dot{\beta}(t) = -\nabla_{\beta} \mathcal{L}_n(f_t), & \beta(0) = \mathbf{0}; \\ \dot{a}(t) = -\nabla_a \mathcal{L}_n(f_t), & a_j(0) = \lambda_j^{1/2}, \end{cases} \quad (5.11)$$

其中 $f_t = \langle \Phi_{a(t)}(x), \beta(t) \rangle_{\ell^2}$, $\mathcal{L}_n$ 是 (4.2) 中定义的经验损失。这里, $(\lambda_j)_{j \geq 1}$ 是核的初始特征值。将梯度流 (5.11) 在时刻 t 得到的估计量记为 $\hat{f}_t^{\text{DA}} = f_t$ 。

5.2.1 更深层的过参数化

除了 (5.10) 中的参数化特征映射外, 还可以进一步考虑在特征映射中增加额外参数以构成多层模型。现在考虑更深层的参数化形式:

$$\Phi_{a,b}(x) = \left(a_j b_j^D e_j(x) \right)_{j \geq 1}, \quad k_{a,b}(x, x') = \sum_{j \geq 1} a_j^2 b_j^{2D} e_j(x) e_j(x'), \quad (5.12)$$

其中 $D > 0$ 是额外层的数量, $\mathbf{b} = (b_j)_{j \geq 1}$ 是额外参数。类似地考虑 $f(x) = \langle \Phi_{a,b}(x), \beta \rangle_{\ell^2}$ 以及参数 $\mathbf{a}$ 、 $\mathbf{b}$ 和 β 上的梯度流:

$$\begin{cases} \dot{\mathbf{a}} = -\nabla_{\mathbf{a}} \mathcal{L}_n(f_t), & a_j(0) = \lambda_j^{\frac{1}{2}}; \\ \dot{\mathbf{b}} = -\nabla_{\mathbf{b}} \mathcal{L}_n(f_t), & b_j(0) = b_0; \\ \dot{\beta} = -\nabla_{\beta} \mathcal{L}_n(f_t), & \beta(0) = \mathbf{0}. \end{cases} \quad (5.13)$$

这里, $b_0 > 0$ 是所有 b_j 的共同初始值。将此梯度流 (5.13) 在时刻 t 得到的估计量记为 $\hat{f}_t^{\text{DA}, D}$ 。

值得注意的是, 如果考虑形式为 $\Phi(x) = (a_j b_{j,1} \cdots b_{j,D} e_j(x))_{j \geq 1}$ 且具有相同初始化 $b_{j,k} = b_0, k = 1, \dots, D$ 的过参数化, 则由于对称性, $b_{j,k}$ 在训练过程中将保持相同, 因此在不计常数因子的意义下, 这等价于本章的参数化 $\theta_j = a_j b_j^D \beta_j$ 。

5.3 泛化能力的结果

对于回归函数, 由于 $(e_j(x))_{j \geq 1}$ 构成一组标准正交基, 可以将其展开为 $f^*(x) = \sum_{j \geq 1} \theta_j^* e_j(x)$ 。随后, 本文针对真实系数 $(\theta_j^*)_{j \geq 1}$ 定义以下量:

$$\begin{aligned} J_{\text{sig}}(\delta) &:= \left\{ j : |\theta_j^*| \geq \delta \right\}, \quad \Phi(\delta) := |J_{\text{sig}}(\delta)|, \\ \Psi(\delta) &= \sum_{j \notin J_{\text{sig}}(\delta)} (\theta_j^*)^2 = \sum_{j=1}^{\infty} (\theta_j^*)^2 \mathbf{1}\{(\theta_j^*)^2 < \delta^2\}. \end{aligned} \quad (5.14)$$

这两个量分别衡量了真实系数中显著分量的数量以及非显著分量的平方和, 它们描述了拟合回归函数 (或者这一组系数) 的难易程度。本章对真实系数的假设如下:

假设 5.1: 存在常数 B_{∞} 使得对于所有 $j \geq 1$ 有 $|\theta_j^*| \leq B_{\infty}$ 。此外, 存在常数 $\kappa \geq 1$ 和 $B_{\text{spn}} > 0$ 使得

$$\max J_{\text{sig}}(\delta) \leq B_{\text{spn}} \delta^{-\kappa}, \quad \forall \delta > 0, \quad (5.15)$$

且存在常数 $s_0, B_{\text{sig}} > 0$ 使得

$$\Phi(\delta) \leq B_{\text{sig}} \delta^{-(1-s_0)}. \quad (5.16)$$

在假设 5.1 中, 条件 (5.15) 表明显著分量 (即那些 $|\theta_j^*| \geq \delta$ 的分量) 的跨度随 $1/\delta$ 的增长速度至多为多项式阶。这一假设较为宽松, 适用于许多实际场景, 例如第 5.1.3 小节中考虑的情况 (对于第一个核, $\kappa = \frac{2q}{p+1}$)。从另一个角度看, 它对真实信号顺序与特征值顺序之间的“错位”程度施加了一个温和的限制, 其中 κ 衡量了 θ_j^* 的排序与 j 本身排序之间的偏差。需要说明的是, 条件 (5.15) 本身只是关于显著系数所在位置的跨度控制, 并不意味着 θ^* 是稀疏的, 也不推出 $(\theta_j^*)_{j \geq 1} \in \ell^1$ 。在 $\lambda_j \asymp j^{-\gamma}$ 的情形下, 传统源条件例如 $\sum_{j \geq 1} (\theta_j^*)^2 / \lambda_j^{2r} < \infty$ 可以推出 $\max J_{\text{sig}}(\delta) \lesssim \delta^{-1/(r\gamma)}$, 因此可看作 (5.15) 的一类充分条件; 但 (5.15) 只保留这一跨度后果, 并不与源条件等价。另一方面, 条件 (5.16) 要求显著分量的计数函数衰减足够快, 以保证 $(\theta_j^*)_{j \geq 1}$ 是可求和的, 这同样是一个温和的正规性条件。更加具体地, 本章可以引入如下具有错位的信号结构假设:

假设 5.2: 存在一个子列 $j(\ell)$ 以及 $p > 0$ 和 $q \geq 1$, 满足

$$\theta_{j(\ell)}^* \asymp \ell^{-\frac{p+1}{2}}, \quad \text{且对于其余分量 } \theta_k^* = 0, \quad (5.17)$$

以及 $j(\ell) \asymp \ell^q$ 。

假设 5.2 明确地规定了系数序列 $(\theta_j^*)_{j \geq 1}$ 的错位结构, 它与 (5.9) 中的设定相似。其中, p 衡量了回归函数或者其系数内蕴的衰减速率, 而 q 则衡量了回归函数系数衰减与给定的特征函数基底顺序之间的错位程度。当 $p > 1$ 时, 容易说明假设 5.2 满足假设 5.1 中的条件。

此外, 本章还需要引入以下假设: 特征函数一致有界且 (初始) 特征值呈多项式阶衰减。这一假设在核方法文献中非常常见^[34,36], 例如当特征函数为傅里叶基且核与 Sobolev 空间相关联时^[35], 该假设均成立。

假设 5.3: 假设 $\sup_{j \geq 1} \|e_j(x)\|_\infty \leq C_{\text{eigf}}$, 且对于某些 $\gamma > 1$, 有

$$c_{\text{eig}} j^{-\gamma} \leq \lambda_j \leq C_{\text{eig}} j^{-\gamma}, \quad (5.18)$$

其中 $C_{\text{eigf}}, c_{\text{eig}}, C_{\text{eig}} > 0$ 为常数。

基于上述假设, 本文现在给出对角自适应核回归的泛化误差刻画, 包括单层模型和多层模型两种情况。

定理 5.1: 如果假设 5.3 和假设 5.1 成立, 则对于任意 $s > 0$, 通过选择 $t \asymp n^{\frac{1}{2}}$, 当 n 足够大时, 以至少 $1 - C/n^2$ 的概率有

$$\|f^* - \hat{f}_t^{\text{DA}}\|_{L^2}^2 \lesssim \frac{1}{n} \left[\Phi(n^{-1/2} \sqrt{\ln n}) \ln n + n^{\frac{1+s}{2\gamma}} \right] + \Psi(n^{-1/2} \ln n). \quad (5.19)$$

此处, 常数 C 及隐含的常数可能取决于假设中的常数以及 s 的选择。

定理 5.2: 如果假设 5.3 和假设 5.1 成立, 则对于任意 $s > 0$, 通过选择 $t \asymp n^{\frac{D+1}{D+2}}$ 和

$b_0 \asymp n^{-\frac{1}{2(D+2)}}$, 当 n 足够大时, 以至少 $1 - C/n^2$ 的概率有

$$\|f^* - \hat{f}_t^{\text{DA}, D}\|_{L^2}^2 \lesssim \frac{1}{n} \left[\Phi(n^{-1/2} \sqrt{\ln n}) \ln n + n^{\frac{1+s}{(D+2)\gamma}} \right] + \Psi(n^{-1/2} \ln n). \quad (5.20)$$

此处, 常数 C 及隐含的常数可能取决于假设中的常数以及 s 的选择。

如果考虑假设 5.2 中具体给出的多项式错位结构, 有 $\Phi(\delta) \asymp \delta^{-\frac{2}{p+1}}, \Psi(\delta) \asymp \delta^{\frac{2p}{p+1}}$, 可以得到更明确的收敛速率如下。举一个更具体的例子, 如果考虑 (5.9) 的设定 (证明见 Li et al.^[101] 的 Lemma F.2), 则得到以下推论:

推论 5.1: 考虑 (5.11) 或 (5.13) 中的对角自适应特征模型。假设假设 5.3 和假设 5.2 成立, 其中 $p > 1$, 且 $\gamma > \frac{p+1}{D+2}$ 。则通过选择 $b_0 \asymp n^{-\frac{1}{2(D+2)}}$ (若 $D \neq 0$) 和 $t \asymp n^{\frac{D+1}{D+2}}$, 当 n 足够大时, 忽略对数因子, 有

$$\|f^* - \hat{f}_t^{\text{DA}, D}\|_{L^2}^2 = \tilde{O}_{\mathbb{P}}(n^{-\frac{p}{p+1}}). \quad (5.21)$$

5.3.1 讨论

从以下几个方面讨论定理 5.1 和定理 5.2 的启示。

5.3.1.1 泛化能力的提升

定理 5.1 和定理 5.2 展示了对角自适应特征模型的泛化性能。与普通的固定核函数的核梯度流相比 (见第 5.1 节), 对角自适应特征模型通过使特征值自适应于真实信号, 显著提高泛化性能。推论 5.1 表明, 在错位结构假设下, 对角自适应特征模型能够实现最优收敛速率 $n^{-\frac{p}{p+1}}$ (至多含对数因子), 而传统的核梯度流方法只能实现 $n^{-\frac{p}{p+q}}$ 的速率。当 q 较大时, 即真实信号的排序与特征值的排序之间存在严重错位时, 这种提升尤为显著。此外, 如果回到第 5.1.2 小节中使用各向同性核 k_1 的低维回归函数, 可以看到, 虽然固定核梯度流会受到收敛速率为 $\frac{2t}{2t+d}$ 的“维度灾难”影响, 但对角自适应特征模型却能实现与维度无关的速率 $\frac{2t}{2t+d_0}$ 。因此, 自适应模型显著改善了泛化性能。

5.3.1.2 自适应的早停时间的选择

自适应核方法的一个显著优势是它不需要关于真实参数的先验知识来选择停止时间。考虑 (5.9) 描述的情况, 传统梯度下降需要选择停止时间 $t \asymp n^{(q\gamma)/(p+q)}$ 才能达到最优收敛速率, 而这一选择关键取决于真实参数中未知的 p 和 q 。相比之下, 对角自适应特征模型只需选择停止时间 $t \asymp n^{\frac{D+1}{D+2}}$, 这并不依赖于未知的真实参数。

5.3.1.3 深度的影响

定理 5.2 的结果还表明，更深的过参数化可以进一步提高泛化性能。在两层过参数化中，定理 5.1 中出现的额外项 $n^{\frac{1+s}{2\gamma}}$ 是由于调整较大特征值时的局限性造成的。通过引入深度，即在模型中添加额外的 D 层并配合适当的初始化，这一项可以在定理 5.2 中改进为 $n^{\frac{1+s}{(D+2)\gamma}}$ 。这种改进表明，深度可以细化模型对特征值自适应的敏感度，从而实现对潜在信号结构更微妙的调整。这一发现强调了模型深度在增强学习过程中的重要性，也为深度学习模型的经验成功提供了理论依据。

5.3.1.4 与前人工作的比较

本章所考虑的对角自适应特征模型与现有关于过参数化梯度下降泛化性能的文献密切相关^[77-79]。从以下几个方面将本章的结果与现有文献进行比较：

- **问题设置：**现有文献^[77-79]主要研究高维线性回归领域，侧重于隐式正则化和稀疏性；而本研究则深入探讨非参数核回归，强调可训练核对潜在信号结构的自适应性。
- **过参数化设置：**现有工作 Zhao et al.^[77]考虑的是两层 Hadamard 积 $\theta = a \odot b$ 的过参数化设置，其中每个分量的初始化均相同，即 $a(0) = \alpha \mathbf{1}$ 且 $b(0) = \mathbf{0}$ 。相比之下，本章的工作考虑为每个分量初始化不同的特征值 $a_j(0) = \lambda_j^{1/2}$ 。此外，本章通过增加额外的 D 层将过参数化扩展到更深的模型。虽然 Vaškevičius et al.^[79] 的后续工作 Li et al.^[78] 考虑了更深的过参数化，但其形式为 $\theta = u^{\odot D} - v^{\odot D}$ 且 $u(0) = v(0) = \alpha \mathbf{1}$ 。虽然这种设置由于初始化均匀而易于分析，但它无法揭示特征值学习的过程，而这正是本章结果中的一个核心洞察。此外，对定理 5.2 的分析涉及不同初始化的 a_j 和 b_j 之间的相互作用，因此本章的分析比现有工作更为复杂。
- **对过参数化的诠释：**先前的工作主要将过参数化视为一种隐式正则化机制，而本章的工作提供了一个全新的视角，即过参数化通过学习特征值来适应真实信号的结构。本文将在第 5.4 节更加具体地展示，核的参数化表现出一种特征学习行为，其中特征值得以学习以匹配真实信号。
- **受限等距性质（RIP）条件：**现有文献^[77-79]依赖 RIP 条件来分析梯度动力学。相比之下，本章的工作通过利用特征函数的集中性和对梯度流动力学的精细分析，不需要 RIP 条件。此外，正如 Vaškevičius et al.^[79] 所讨论的，Zhao et al.^[77], Li et al.^[78] 中的 RIP 条件实际上对于独立同分布的随机样本来说过于苛刻，难以成立。

5.4 特征映射的改进

现在, 本文将结合第 4.2 节中定义的特征误差度量, 来解释为什么对角自适应核回归能够实现更好的泛化性能。由于特征基是固定的, 投影误差 $\mathcal{E}_{\text{Proj}}(\Phi, f^*)$ 是一个常数, 因此假设 f^* 被包含在 $\{e_j\}_{j \in N}$ 张成的空间中并不失一般性。此时, (4.5) 中的特征误差度量简化为:

$$\mathcal{E}(\delta; \epsilon^2; \Phi, f^*) = \mathcal{E}_{\text{Stat}}(\delta, \epsilon^2; \Phi, f^*) = \#\{j \in N : \lambda_j \geq \delta\} \epsilon^2 + \sum_{j \in N} (\theta_j^*)^2 \mathbf{1}\{\lambda_j < \delta\}, \quad (5.22)$$

其中 $f^* = \sum_{j \in N} \theta_j^* e_j$ 是在特征基下表示的真实函数, λ_j 是与特征映射 Φ 相关联的特征值序列。在固定特征基的设定下, 特征映射实际上是通过权重序列来指定学习索引的顺序。当特征映射给出的顺序与真实函数系数的顺序相一致时, 特征误差度量达到最小。因此, 特征误差度量正是描述真实函数与特征映射之间“错位”(misalignment)程度的一种度量, 这也是本文所感兴趣的。

具体地, 在第 5.1.3 小节所考虑的例子中, 如果 Φ 所对应的特征值满足 $\lambda_j \asymp j^{-\gamma}$, 可以得到

$$\mathcal{E}^*(\delta; \Phi, f^*) := \inf_{\delta \geq 0} \mathcal{E}(\delta, \epsilon^2; \Phi, f^*) = (\epsilon^2)^{\frac{p}{p+q}},$$

而如果 Φ^* 所对应的特征值满足 $\lambda_{\ell(j)} \asymp j^{-\gamma}$, 则有

$$\mathcal{E}^*(\delta; \Phi^*, f^*) = (\epsilon^2)^{\frac{p}{p+1}}.$$

5.4.1 序列损失下的结果

利用第 4.3 节中的讨论, 首先考察替换样本损失 $\hat{\mathcal{L}}_n$ 为序列损失 $\bar{\mathcal{L}}_n$ 的情况以呈现最核心的结果。此时, 所考虑的对角自适应特征模型的演化方程为:

$$\begin{cases} \dot{\beta}(t) = -\nabla_{\beta} \bar{\mathcal{L}}_n, & \beta_j(0) = 0; \\ \dot{\theta}(t) = -\nabla_{\theta} \bar{\mathcal{L}}_n, & \theta_j(0) = \lambda_j^{\frac{1}{2}}, \end{cases} \quad (5.23)$$

以及对应的深层参数化版本

$$\begin{cases} \dot{\beta}(t) = -\nabla_{\beta} \bar{\mathcal{L}}_n, & \beta(0) = \mathbf{0}; \\ \dot{\theta}(t) = -\nabla_{\theta} \bar{\mathcal{L}}_n, & \theta_j(0) = \lambda_j^{1/2}; \\ \dot{b}(t) = -\nabla_b \bar{\mathcal{L}}_n, & b_j(0) = b_0. \end{cases} \quad (5.24)$$

本文有如下的结果:

定理 5.3: 考虑在假设 5.2 下由 (5.23) 定义的对角自适应特征模型。令 $t_* = t_*(n) \asymp$

$\sqrt{n/(\log n)}$, 则以至少 $1 - Cn^{-2}$ 的概率有:

$$\mathcal{E}^*(n^{-1}; \Phi_{\theta(t)}, f^*) \text{ 在 } t \in [0, t_*] \text{ 上关于 } t \text{ 单调递减.} \quad (5.25)$$

此外, 如果 $\gamma > \frac{1}{2}(1 + \frac{p}{q})$, 则

$$\mathcal{E}^*(n^{-1}; \Phi_{\theta(0)}, f^*) \asymp n^{-\frac{p}{p+q}} \gg n^{-(1-\frac{1}{2\gamma})} + n^{-\frac{p}{p+1}} (\log n)^{\frac{2p}{p+1}} \gtrsim \mathcal{E}^*(n^{-1}; \Phi_{\theta(t_*)}, f^*).$$

定理 5.4: 考虑在假设 5.2 下由 (5.24) 定义的对角自适应特征模型。令 $t_* = t_*(n) \asymp n^{\frac{D+1}{D+2}} / \sqrt{\log n}$ 且 $b_0 \asymp n^{-\frac{1}{2(D+2)}}$, 则以至少 $1 - Cn^{-2}$ 的概率有:

$$\mathcal{E}^*(n^{-1}; \Phi_{\theta(t)}, f^*) \text{ 在 } t \in [0, t_*] \text{ 上关于 } t \text{ 单调递减.} \quad (5.26)$$

此外, 如果 $\gamma > \frac{1}{D+2}(1 + \frac{p}{q})$, 则

$$\mathcal{E}^*(n^{-1}; \Phi_{\theta(0)}, f^*) \asymp n^{-\frac{p}{p+q}} \gg n^{-(1-\frac{1}{(D+2)\gamma})} + n^{-\frac{p}{p+1}} (\log n)^{\frac{2p}{p+1}} \gtrsim \mathcal{E}^*(n^{-1}; \Phi_{\theta(t_*)}, f^*).$$

定理 5.3 和定理 5.4 展示了对角自适应特征模型在训练过程中持续改进特征映射、从而更好地对齐真实函数的能力。初始时的特征映射 $\Phi_{\theta(0)}$ 的特征误差度量为 $\mathcal{E}^*(n^{-1}; \Phi_{\theta(0)}, f^*) \asymp n^{-\frac{p}{p+q}}$, 这在很大程度上受到错位程度 $q > 1$ 的影响。在训练过程中, 自适应特征逐渐调整以匹配真实函数, 导致特征误差度量 $\mathcal{E}^*(n^{-1}; \Phi_{\theta(t)}, f^*)$ 随 t 单调递减。最终特征误差度量中的 $n^{-(1-\frac{1}{(D+2)\gamma})}$ 项源于无法完全纠正的初始错位。尽管如此, 只要初始衰减足够快 (即 $\gamma > \frac{p+1}{D+2}$), 最终的特征误差度量就能恢复到非参数最优速率 $n^{-\frac{p}{p+1}}$ (带有一个对数因子)。

此外, 正如在泛化能力的结果中观察到的, 定理 5.4 中深度 D 的引入可以通过放宽对初始快速衰减的要求来进一步优化特征误差度量。深度的益处体现在由错位引起的额外误差项 $n^{-(1-\frac{1}{(D+2)\gamma})}$ 上, 该项随 D 的增加而减小。这种改进源于深层参数化在训练过程中调整特征权重的灵活性增强。这种由深度增强的自适应性反映了深度神经网络的行为, 即通过多层结构逐步细化特征表示。然而, 这是以增加计算复杂度和延长最优停止时间 $t_* \asymp n^{\frac{D+1}{D+2}} / \sqrt{\log n}$ 为代价的, 体现了性能提升与训练效率之间的权衡。

5.4.2 样本损失下的结果

现在, 本文研究在经验损失下对角自适应特征模型的特征误差度量演化。本节额外假设 ϵ 是独立的次高斯 (sub-Gaussian) 噪声。

定理 5.5: 令假设 5.2 和假设 5.3 成立。考虑对角自适应特征模型 (5.11) 或其多层版本 (5.13)。若 $D \neq 0$, 取 $b_0 \asymp n^{-\frac{1}{2(D+2)}}$ 。令 $s > 0$ 为任意小的常数, 并定义 $q = 2^{\frac{2(D+1)}{D+2}}$ 。那么, 存在 $L \asymp (-\frac{1}{2} + s) \log n$, 一个满足 $\delta_L \leq n^{-\frac{1}{2}+s}$ 的递减序列 $\delta_l = Cq^{-l}$ ($l \leq L$),

以及满足 $t_l \lesssim \delta_l^{-l} \log n$ 的时间序列 $t_0 = 0 < t_1 < \dots < t_L = t_* \lesssim n^{\frac{D+1}{D+2}}$, 使得以至少 $1 - Cn^{-2}$ 的概率, 有:

$$\mathcal{E}^*(n^{-1}; \Phi_{\theta(t), b(t)}, f^*) \lesssim \delta_l^p + n^{-\frac{p}{p+1}} + n^{-(1-\frac{1+s}{(D+2)\gamma})} \quad \forall t \in [t_l, t_*], \forall l = 0, \dots, L. \quad (5.27)$$

特别地,

$$\mathcal{E}^*(n^{-1}; \Phi_{\theta(t_*), b(t_*)}, f^*) \lesssim n^{-\frac{p}{p+1}+s} + n^{-(1-\frac{1+s}{(D+2)\gamma})}.$$

与定理 5.3 和定理 5.4 类似, 定理 5.5 表明, 在经验损失下, 对角自适应特征方法同样会在训练过程中逐步改善特征误差度量。此外, 定理 5.5 呈现出一种渐进式的阶梯状下降模式, 而非单调递减, 这是由于在经验损失下不同系数之间的交互作用所致。尽管如此, 在定理 5.5 中仍可获得与序列模型相同的最终特征误差度量。

最后, 本文也给出如下结果, 更加定量地刻画在最终停止时间 t_* 时特征值如何根据真实信号进行调整。

命题 5.1: 在与定理 5.1 或定理 5.2 相同的条件下, 对于选定的停止时间 t_* , 特征学习项 $a_k(t)b_k^D(t)$ (对于定理 5.1 中单层的情况, 取 $D = 0$ 且令 $b_j^D = 1$) 满足:

- 信号分量: 对于满足 $|\theta_k^*| \geq C_1 n^{-1/2} \ln n$ 的 θ_k^* , 有

$$a_k(t_*)b_k^D(t_*) \geq c_1 |\theta_k^*|^{\frac{D+1}{D+2}}. \quad (5.28)$$

- 噪声分量: 对于满足 $|\theta_k^*| \leq n^{-1/2}$ 且 $\lambda_k \leq n^{-\frac{1+s}{D+2}}$ 的分量, 有

$$a_k(t_*)b_k^D(t_*) \leq C_2 a_k(0)b_k^D(0)e^{C_3(\sqrt{\ln n} + \sqrt{\ln k})}. \quad (5.29)$$

此处, c_1, C_1, C_2, C_3 均为不依赖于 n 或 k 的常数。

通过该命题可以看出, 对于信号分量, 学习到的特征值至少为真实信号强度某次幂的常数倍。因此, 过参数化梯度下降调整了特征值以匹配真实信号。在噪声分量的情况下, 虽然特征值在训练过程中仍会有所增长, 但只要 λ_j 相对较小, 特征值就不会超过初始值的常数倍。这一发现表明, 过参数化梯度下降能够有效地使特征值自适应于真实信号, 同时减轻对噪声的过拟合, 呈现出特征学习 (feature learning) 的行为。本文注意到, 当 λ_j 相对较大时, 该方法仍倾向于过拟合噪声分量, 在泛化误差中贡献了一个额外的 $n^{-\frac{1}{D+2}\frac{1}{\gamma}}$ 项, 但当 γ 较大时, 该项可以忽略不计。

综上所述, 本章证明了在固定特征基的前提下, 自适应特征模型在不同的统计模型中均能持续优化特征误差度量。结果表明, 自适应模型能够有效地学习出与真实函数对齐的“最优”特征映射, 使特征误差度量接近非参数最优速率。这些发现强调了自适应特征方案的灵活性和稳健性, 架起了经典统计方法与现代学习范式之间的桥梁。

5.5 数值实验

本节通过数值模拟来验证理论结果。在训练过程中，通过小步长的梯度下降法对梯度流进行离散化处理。

5.5.1 低维结构的设定

先关注第 5.1.2 小节中的设定，其中 $\mathbf{x} \sim \text{Unif}([-1, 1]^d)$ ，特征函数可以选取为傅里叶基。为了说明自适应核方法的训练动力学，考虑 $d = 2$ 的情况。将真实函数选取为 $f^*(\mathbf{x}) = \cos(7.5\pi x_1)$ ，其中 $\mathbf{x} = (x_1, x_2)$ 。此时可以计算出 f^* 在特征函数上的系数为 $\langle f^*, \cos(k\pi x_1) \rangle = \frac{30}{(4k^2 - 225)\pi}$ ， $\langle f^*, \sin(k\pi x_1) \rangle = 0$ 。将特征值选取为 $\lambda_m = (1 + \|\mathbf{m}\|_2^2)^{-3/2}$ 。在 $n = 400$ 个样本的设定下，在图 5.1 中绘制了泛化误差曲线和系数的演化过程。

从泛化误差图中可以看出，自适应核方法的表现优于固定核方法。此外，从系数图中可以观察到，固定核方法学习了无关方向上的噪声，而自适应核方法则集中于真实方向，这说明了自适应核方法在学习信号结构方面的优势。这与命题 5.1 中的理论结果相一致。

本文进行了进一步的实验，在各种设定下对比固定核方法与对角自适应核方法的泛化误差。参见图 5.2。结果表明，自适应核方法能够获得明显优于固定核方法的泛化误差。此外，还可以观察到在自适应核方法中引入额外层所带来的益处。总体而言，数值模拟支持了本章的理论结果。

5.5.2 错位结构的设定

现在，本文考察第 5.1.3 小节中的设定，并且直接考虑序列损失的情况以简化数值实验。针对真实参数 θ^* 、特征值衰减率 γ 和深度 D 的不同设置，本文考察了过参数化梯度下降（包括更深层的参数化）的泛化性能。结果如表 5.1 所示。本文发现泛化误差速率总体上与推论 5.1 中的理论结果一致，虽然由于有限样本量存在一些波动。对比 $\gamma = 1.1$ 时深度 $D = 0$ 、 $D = 1$ 和 $D = 2$ 的结果，可以看到 $D = 0$ 的方法收敛速率最慢，而 $D = 2$ 的方法收敛速率最快，这证明了更深的参数化能够提升泛化性能。总之，数值实验验证了本章的理论结果。

此外，本文研究了命题 5.1 中讨论的特征值项 $a_j(t)b_j^D(t)$ 随时间 t 的演化。结果如图 5.3 所示。本文发现，特征值项确实能够适应信号的潜在结构：对于大信号，特征值项随着训练的进行而趋近于信号值；而对于小信号，特征值项没有显著增加。此外，本文发现更深的过参数化减少了噪声分量对应的特征值项的波动，从而提升了模型的泛化性能。本节实验主要用于验证对角参数化自适应特征模型的理论

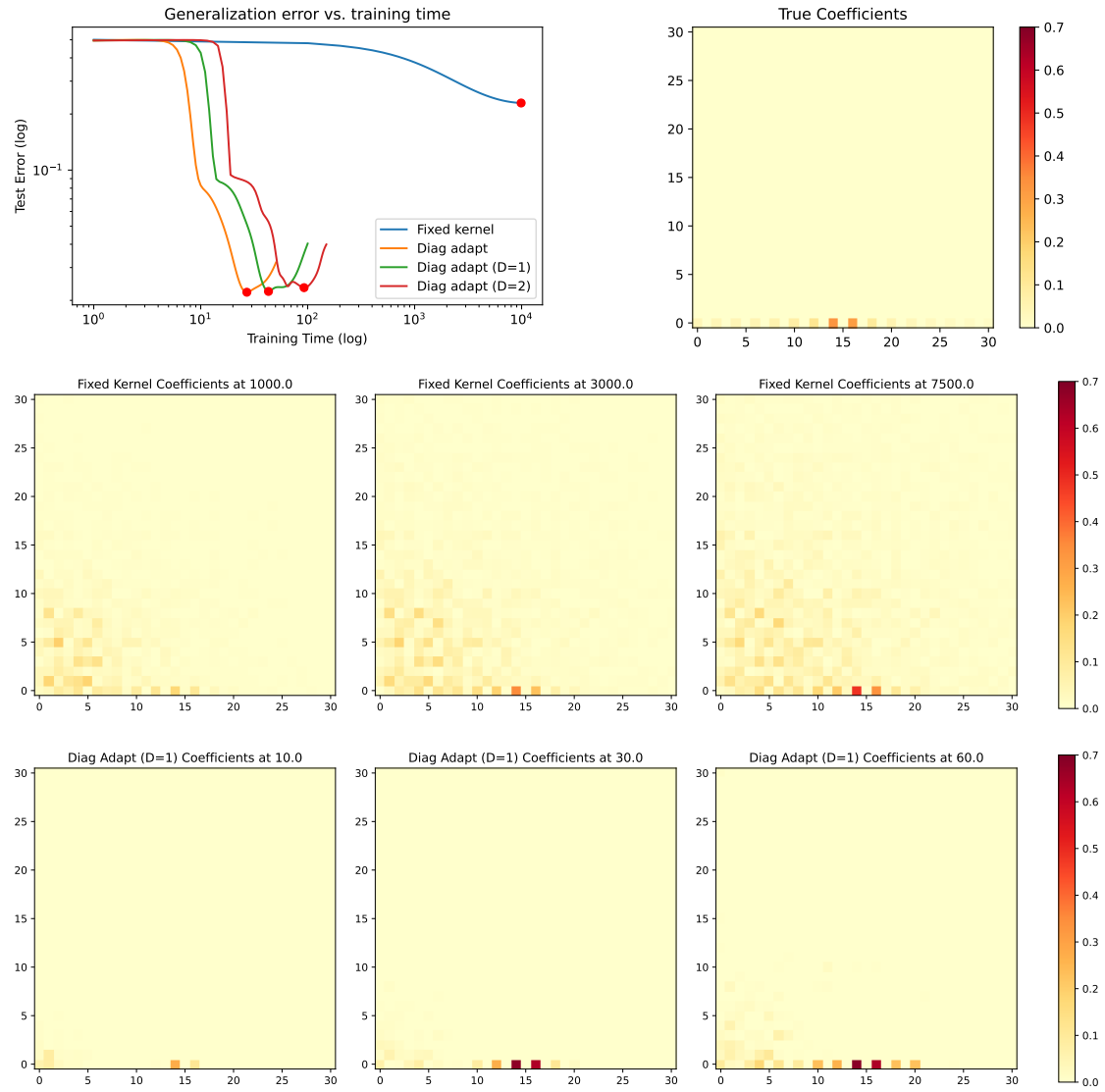


图 5.1 固定核方法与对角自适应核方法在具有低维结构的 $d = 2$ 示例上的对比。左上图显示了泛化误差曲线。右上图显示了真实系数，下两行显示了固定核方法和对角自适应核方法的系数演化。由于特征函数的指标是二维的，将系数绘制在二维网格中。

表 5.1 在真实参数 p, q 和特征值衰减率 γ 的不同设置下，对角自适应特征模型的泛化误差收敛速率，其中 r^* 为理想收敛速率。收敛速率是通过对样本量 n 在 2000, 2200, ..., 4000 范围内的泛化误差进行对数最小二乘拟合估计的，其中泛化误差为 256 次重复实验的均值。

	$p = 0.6 (r^* = 0.37)$			$p = 1 (r^* = 0.5)$			$p = 3 (r^* = 0.75)$		
γ	$q = 1$	$q = 1.5$	$q = 2$	$q = 1$	$q = 1.5$	$q = 2$	$q = 1$	$q = 1.5$	$q = 2$
1.1	0.38	0.40	0.50	0.50	0.45	0.48	0.78	0.69	0.67
2	0.36	0.41	0.52	0.49	0.46	0.50	0.80	0.73	0.72
3	0.36	0.41	0.52	0.48	0.46	0.50	0.76	0.73	0.74

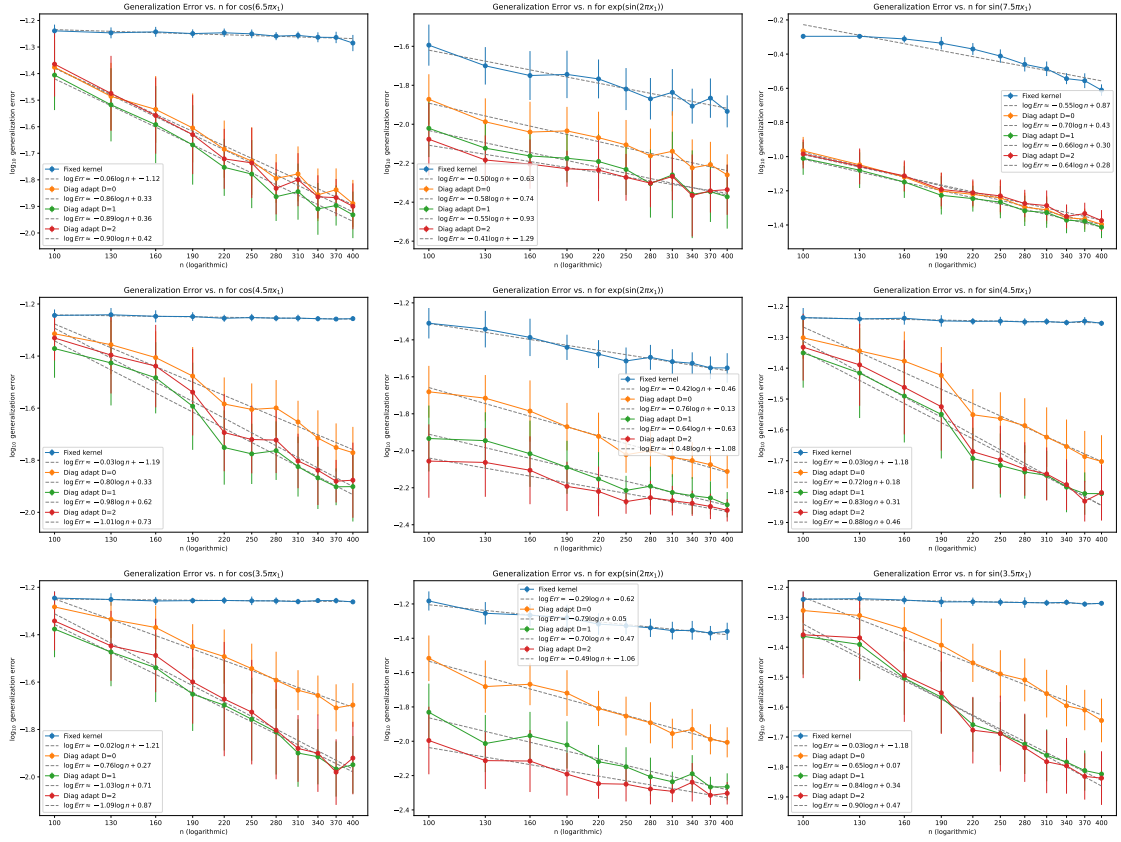


图 5.2 固定核方法与对角自适应核方法在各种设定下的对比。三行分别对应第 5.1.2 小节中维度 $d = 2, 3, 4$ 的情况。误差棒表示 32 次运行的标准差。

表 5.2 深度为 $D = 1$ 和 $D = 2$ 的对角自适应特征模型的泛化误差的收敛速率。设置与表 5.1 相同。

	γ	$p = 0.6 (r^* = 0.37)$			$p = 1 (r^* = 0.5)$			$p = 3 (r^* = 0.75)$		
		$q = 1$	$q = 1.5$	$q = 2$	$q = 1$	$q = 1.5$	$q = 2$	$q = 1$	$q = 1.5$	$q = 2$
$D = 1$	1.1	0.36	0.40	0.52	0.49	0.46	0.49	0.79	0.73	0.72
	2	0.36	0.40	0.52	0.48	0.46	0.50	0.76	0.73	0.73
	3	0.30	0.32	0.38	0.45	0.41	0.44	0.75	0.73	0.74
$D = 2$	1.1	0.34	0.39	0.49	0.46	0.44	0.48	0.76	0.74	0.75
	2	0.35	0.40	0.51	0.47	0.45	0.49	0.74	0.73	0.73
	3	0.36	0.40	0.51	0.48	0.46	0.50	0.74	0.73	0.73

预言。进一步将自适应核梯度流与直接训练的神经网络、经典自适应非参数回归方法进行系统比较，需要针对具体网络结构、优化算法和数据生成机制设计额外实验，本文将其留作后续工作。

5.6 单层参数化的泛化能力证明

在本节中给出定理 5.1 的证明，而定理 5.2 的证明虽然类似但更为复杂，将推迟至第 5.7 节。

证明的核心思想是通过一个对角化版本来近似动力学过程 (5.31) (即 (5.11) 的显式形式)，从而本文可以分别分析每个分量的动力学。接着，本文分别处理信号分量和噪声分量，证明信号分量以最优速率收敛到真实参数，且噪声分量得到了良好的控制。这里的主要挑战在于处理一个额外的扰动项：本章分析了带扰动的对角化方程的动力学，并利用命题 5.2 进行迭代论证，从而逐渐缩小扰动项。

本节其余部分的结构安排如下：在第 5.6.1 小节中推导了梯度流的显式形式 (5.31)。在第 5.6.2 小节和第 5.6.3 小节中建立了原始动力学与对角化动力学之间的近似关系。在第 5.6.5 小节中分析了带扰动的一维方程的动力学。最后，在第 5.6.4 小节中证明了定理 5.1。

在本章余下的证明中，使用 s 表示一个可以取为任意小的正数。记 $\ln^+(x) = \max(\ln x, 0)$ 。用

$$E(x) = (e_j(x))_{j \geq 1}^\top$$

表示基函数给出的（无穷维）列向量。

5.6.1 梯度流的显式形式

对于 (5.11) 中的自适应核回归，记 $\theta(t) = (\theta_j)_{j \geq 1} = a(t) \odot \beta(t)$ 。那么，有

$$f_t(x) = \langle \Phi_{a(t)}(x), \beta(t) \rangle = \langle E(x), a(t) \odot \beta(t) \rangle = \langle E(x), \theta(t) \rangle = E(x)^\top \theta(t),$$

作为对比，真实函数类似地给出为 $f^*(x) = \sum_{j \geq 1} \theta_j^* e_j(x) = E(x)^\top \theta^*$ 。

此外，记

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n E(x_i) E(x_i)^\top, \quad r = \frac{1}{n} \sum_{i=1}^n \varepsilon_i E(x_i). \quad (5.30)$$

直接计算表明

$$\nabla_{\beta} \mathcal{L}_n(f_t) = \frac{1}{n} \sum_{i=1}^n D_{\beta(t)} E(x_i) [E(x_i)^\top (\theta(t) - \theta^*) - \varepsilon_i]$$

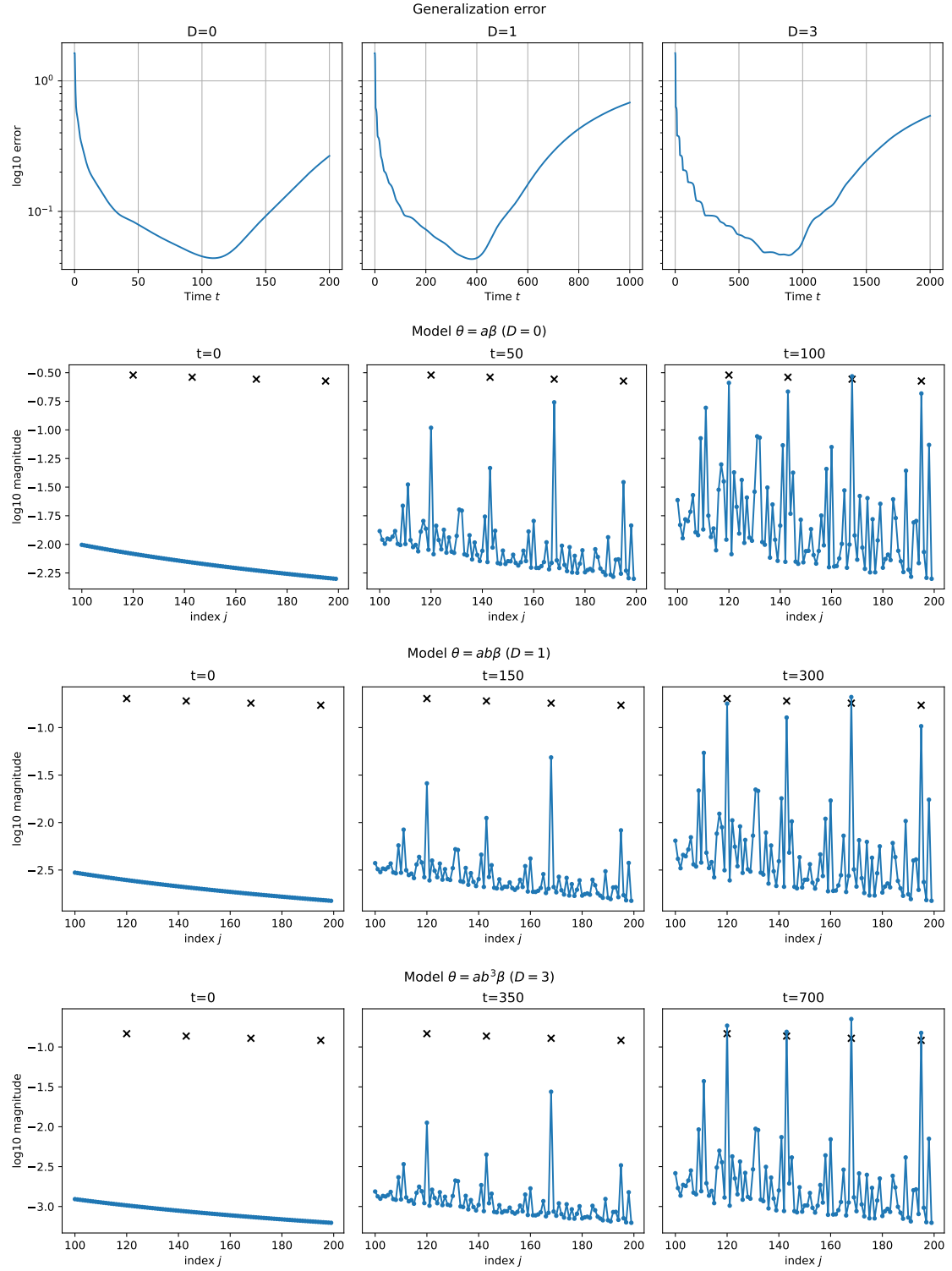


图 5.3 泛化误差以及特征值项 $a_j(t)b_j^D(t)$ 随时间 t 的演化。第一行显示了三种参数化设置 $D = 0, 1, 3$ 的泛化误差随训练时间 t 的变化。其余行显示了特征值项 $a_j(t)b_j^D(t)$ 随时间 t 的演化。为了便于展示，选择了索引 $j = 100$ 到 200 。蓝色线条表示特征值项，黑色标记表示根据命题 5.1 缩放后的非零信号。在设置方面，取 $p = 1$ ， $q = 2$ 和 $\gamma = 2$ 。

$$\begin{aligned}
 &= D_{\beta(t)} \left[\frac{1}{n} \sum_{i=1}^n E(x_i) E(x_i)^\top (\theta(t) - \theta^*) - \frac{1}{n} \sum_{i=1}^n \varepsilon_i E(x_i) \right] \\
 &= D_{\beta(t)} [\hat{\Sigma}(\theta(t) - \theta^*) - \mathbf{r}].
 \end{aligned}$$

类似地,

$$\nabla_a \mathcal{L}_n(f_t) = D_{a(t)} [\hat{\Sigma}(\theta(t) - \theta^*) - \mathbf{r}].$$

因此, 推导出梯度流的显式形式为

$$\begin{cases} \dot{\beta}(t) = -\nabla_{\beta} \mathcal{L}_n = \mathbf{a} \odot [\hat{\Sigma}(\theta^* - \theta) + \mathbf{r}], & \beta(0) = \mathbf{0}, \\ \dot{a}(t) = -\nabla_a \mathcal{L}_n = \beta \odot [\hat{\Sigma}(\theta^* - \theta) + \mathbf{r}], & a_j(0) = \lambda_j^{1/2}. \end{cases} \quad (5.31)$$

5.6.2 分量划分

将 θ 的分量划分为两个集合: 信号分量 $S \subset \mathbb{N}^+$ 和噪声分量 $R = S^c$ 。按如下方式选择 S :

$$S = S_1 \cup S_2 = \left\{ j \geq 1 : |\theta_j^*| \geq n^{-1/2} \sqrt{\ln n} \right\} \cup \left\{ j \geq 1 : \lambda_j \geq n^{-1/2} \right\}. \quad (5.32)$$

那么, 根据假设 5.1, 有

$$|S_1| \leq \Phi(n^{-1/2}) \leq Cn^{1/2-s} \quad \max S_1 \leq Cn^{\kappa/2}.$$

同样, 根据假设 5.3, 有 $|S_2| \leq Cn^{1/(2\gamma)}$ 。注意到 $\gamma > 1$, 得到

$$|S| \leq Cn^{(1-s_0)/2} \quad \max S \leq Cn^{\kappa/2}. \quad (5.33)$$

此外, 假设 5.3 还给出 $S_2^c \geq cn^{1/(2\gamma)}$, 因此

$$L := \min R \geq cn^{1/(2\gamma)}. \quad (5.34)$$

近似分解.

本文通过以下分解, 利用对角化的 $\theta^* - \theta$ 来近似 $\hat{\Sigma}(\theta^* - \theta) + \mathbf{r}$:

$$\hat{\Sigma}(\theta^* - \theta) + \mathbf{r} = (\theta^* - \theta) + \mathbf{p} + \mathbf{q} + \mathbf{r} = (\theta^* - \theta) + \mathbf{h}, \quad (5.35)$$

其中 $\mathbf{h} = \mathbf{p} + \mathbf{q} + \mathbf{r}$,

$$\mathbf{p} = (\hat{\Sigma}_{\cdot S} - I_{\cdot S})(\theta^* - \theta)_S, \quad \mathbf{q} = (\hat{\Sigma}_{\cdot R} - I_{\cdot R})(\theta^* - \theta)_R, \quad (5.36)$$

且 I 是无限维单位矩阵。具体而言, 通过计算以下两项之差得到 (5.35):

$$\begin{aligned}
 \hat{\Sigma}(\theta^* - \theta) &= \hat{\Sigma}_{\cdot S}(\theta^* - \theta)_S + \hat{\Sigma}_{\cdot R}(\theta^* - \theta)_R, \\
 \theta^* - \theta &= I_{\cdot S}(\theta^* - \theta)_S + I_{\cdot R}(\theta^* - \theta)_R.
 \end{aligned}$$

5.6.3 集中不等式

以下引理给出了 (5.35) 中误差项的初步上界。

引理 5.1: 在假设 5.3 下, 存在常数 $C_1, C_2 > 0$, 使得以至少 $1 - C_1/n^2$ 的概率, 对于所有 $k \geq 1$, 有

$$|r_k| \leq C_2 \sqrt{\frac{\ln(nk)}{n}} \quad (5.37)$$

$$|(\hat{\Sigma}_{kR} - I_{kR})\theta_R^*| \leq C_2 \sqrt{\frac{\|\theta_R^*\|_1}{n} \ln(nk)}, \quad (5.38)$$

$$\|\hat{\Sigma}_{kS} - I_{kS}\|_2 \leq C_2 \sqrt{\frac{|S|}{n} \ln(nk)} \quad (5.39)$$

$$\left\| (\hat{\Sigma}_{kR} - I_{kR}) \Lambda_R^{\frac{1}{2}} \right\|_2 \leq C_2 \sqrt{\frac{\text{Tr } \Lambda_R}{n} \ln(nk)}. \quad (5.40)$$

证明 首先为固定的 k 建立不等式, 最后通过联合概率界得到对所有 k 成立的结果。对于第一个不等式, 回顾:

$$r_k = \frac{1}{n} \sum_{i=1}^n \varepsilon_i e_k(x_i).$$

由于 (x_i, ε_i) 是独立同分布的, 且 $|e_k(x_i)| \leq C_{\text{eigf}}$, ε_i 是次高斯 (sub-Gaussian) 的, 可以应用 Hoeffding 不等式引理 5.23 得到: 对于固定的 $k \geq 1$,

$$\mathbb{P} \left\{ |r_k| \geq C_2 \sqrt{\frac{\ln(nk)}{n}} \right\} \leq \frac{C}{k^2 n^2}.$$

对于第二个不等式, 注意到

$$\hat{\Sigma}_{kR} \theta_R^* = \frac{1}{n} \sum_{i=1}^n e_k(x_i) \sum_{j \in R} e_j(x_i) \theta_j^*.$$

令 $\xi_i = e_k(x_i) \sum_{j \in R} e_j(x_i) \theta_j^*$, 有

$$|\xi_i| \leq C_{\text{eigf}}^2 \sum_{j \in R} |\theta_j^*| = \|\theta_R^*\|_1,$$

且 $\mathbb{E} \xi_i = I_{kR} \theta_R^*$ 。因此, Hoeffding 不等式给出

$$\mathbb{P} \left\{ |(\hat{\Sigma}_{kR} - I_{kR}) \theta_R^*| \geq C_2 \sqrt{\frac{\|\theta_R^*\|_1}{n} \ln(nk)} \right\} \leq \frac{C}{k^2 n^2}.$$

对于第三个不等式, 注意到

$$\hat{\Sigma}_{kS} = \frac{1}{n} \sum_{i=1}^n (e_k(x_i) e_j(x_i))_{j \in S}.$$

定义随机向量 $v = (e_k(x_i)e_j(x_i))_{j \in S}$, 有

$$\|v\|_2^2 = \sum_{j \in S} |e_k(x_i)e_j(x_i)|^2 \leq C_{\text{eigf}}^2 |S|.$$

因此, 利用希尔伯特空间上的 Hoeffding 不等式引理 5.24, 得到

$$\mathbb{P} \left\{ \|\hat{\Sigma}_{kS} - I_{kS}\|_2 \geq 2\sqrt{\frac{|S|}{n} \ln(nk)} \right\} \leq \frac{2}{n^2 k^2}.$$

对于第四个不等式, 类似地, 有

$$\hat{\Sigma}_{kR} \Lambda_R^{\frac{1}{2}} = \frac{1}{n} \sum_{i=1}^n \left(e_k(x_i) \lambda_j^{\frac{1}{2}} e_j(x_i) \right)_{j \in R}.$$

定义随机向量 $w = \left(e_k(x_i) \lambda_j^{\frac{1}{2}} e_j(x_i) \right)_{j \in R}$, 有

$$\|w\|_2^2 = \sum_{j \in R} \left| e_k(x_i) \lambda_j^{\frac{1}{2}} e_j(x_i) \right|^2 \leq C_{\text{eigf}}^2 \sum_{j \in R} \lambda_j = \text{Tr } \Lambda_R,$$

因此同样有

$$\mathbb{P} \left\{ \left\| (\hat{\Sigma}_{kR} - I_{kR}) \Lambda_R^{\frac{1}{2}} \right\|_2 \geq 2\sqrt{\frac{\text{Tr } \Lambda_R}{n} \ln(nk)} \right\} \leq \frac{2}{n^2 k^2}.$$

最后, 由于尾概率关于 k 是可求和的, 可以应用联合概率界得到结果。 ■

5.6.4 定理 5.1 的证明

对于稍后确定的某个 $E > 0$, 将过拟合噪声之前的时间定义为

$$T^{\text{err}} = \inf \left\{ t \geq 0 : \exists k \in R, |\theta_k(t)| > 2\lambda_k e^{E\sqrt{\ln n + \ln k}} \right\}, \quad (5.41)$$

其中若上式中的集合为空, 则约定 $T^{\text{err}} = +\infty$ 。在此先声称始终有 $T^{\text{err}} \geq \bar{C}\sqrt{n}$, 因此对于随后考虑的所有时间都有 $t \leq T^{\text{err}}$; 将在本小节末尾的第 5.6.4.2 小节中证明这一声明。在下文中, 即使不明确提及, 也始终假设 $t \leq T^{\text{err}}$ 。

现在, 开始利用命题 5.1 中的集中性结果来控制 (5.35) 中的误差项 p, q, r 。对于 p_k , 有

$$\begin{aligned} |p_k| &\leq \|\hat{\Sigma}_{kS} - I_{kS}\|_2 \|(\theta^* - \theta)_S\|_2 \stackrel{(5.39)}{\leq} C \sqrt{\frac{|S|}{n} \ln(kn)} \cdot \sqrt{|S|} \|(\theta^* - \theta)_S\|_\infty \\ &= C |S| \sqrt{\frac{\ln(kn)}{n}} \cdot \|(\theta^* - \theta)_S\|_\infty, \end{aligned}$$

另一方面, 对于 q_k , 有

$$q_k = (\hat{\Sigma}_{kR} - I_{kR})\theta_R^* - (\hat{\Sigma}_{kR} - I_{kR})\theta_R.$$

对于第一项, 使用 (5.38) 以及假设 5.1 得到

$$|(\hat{\Sigma}_{kR} - I_{kR})\theta_R^*| \leq C \sqrt{\frac{\|\theta_R^*\|_1}{n} \ln(nk)} \leq C \sqrt{\frac{\ln(nk)}{n}}.$$

对于第二项, 使用 (5.40):

$$\begin{aligned} |(\hat{\Sigma}_{kR} - I_{kR})\theta_R| &= \left| (\hat{\Sigma}_{kR} - I_{kR})\Lambda_R^{\frac{1}{2}}\Lambda_R^{-\frac{1}{2}}\theta_R \right| \leq \left\| (\hat{\Sigma}_{kR} - I_{kR})\Lambda_R^{\frac{1}{2}} \right\|_2 \left\| \Lambda_R^{-\frac{1}{2}}\theta_R \right\|_2 \\ &\leq C \sqrt{\frac{\text{Tr } \Lambda_R}{n} \ln(nk)} \cdot \left\| \Lambda_R^{-\frac{1}{2}}\theta_R \right\|_2, \end{aligned}$$

现在, 当 $t \leq T^{\text{err}}$ 时, 有

$$\begin{aligned} \left\| \Lambda_R^{-\frac{1}{2}}\theta_R(t) \right\|_2^2 &\leq C \sum_{k \in R} \lambda_k e^{2E\sqrt{\ln n + \ln k}} \\ &\leq C e^{2E\sqrt{\ln n}} \sum_{k \geq L} \lambda_k e^{2E\sqrt{\ln k}} \\ &\leq C n^s \sum_{k \geq L} k^{-\gamma+s} \leq C_E n^s L^{-(\gamma-1-s)} \\ &\leq C_E n^s n^{-(\gamma-1-s)/(2\gamma)} = C_E n^{-(\gamma-1)/(2\gamma)+s} \leq C, \end{aligned}$$

其中 C_E 是一个额外依赖于 E 的常数, 可以让 n 足够大, 使得所得常数 C 不依赖于 E 。因此, 有

$$|(\hat{\Sigma}_{kR} - I_{kR})\theta_R| \leq C \sqrt{\frac{\text{Tr } \Lambda_R}{n} \ln(nk)} \leq C \sqrt{\frac{\ln(nk)}{n}}.$$

结合 (5.37), 将 p_k, q_k, r_k 的上界总结为

$$|p_k| \leq \eta_k \|(\theta^* - \theta)_S\|_\infty, \quad |q_k| \leq C n^{-1/2} \sqrt{\ln(nk)}, \quad |r_k| \leq C n^{-1/2} \sqrt{\ln(nk)}, \quad (5.42)$$

其中使用了 (5.33) 得到

$$\eta_k = C |S| n^{-1/2} \sqrt{\ln(nk)} \leq C n^{-s} \sqrt{\ln(nk)}.$$

5.6.4.1 信号分量的收缩动力学

对于信号 $k \in S$, 利用 $k \leq \max S \leq C n^{\kappa/2}$ 将 (5.42) 重写为

$$\|p_S\|_\infty \leq \eta \|(\theta^* - \theta)_S\|_\infty, \quad \|q_S + r_S\|_\infty \leq \varepsilon := C n^{-1/2} \sqrt{\ln n}. \quad (5.43)$$

其中可以选择 $\eta = C n^{-s_1}$ 对于某个小的 $s_1 > 0$ 。然而, 这里的挑战在于, 通过 $\|(\theta^* - \theta)_S\|_\infty \leq C$ 对 $\|p_S\|_\infty$ 进行的粗略限制并不能给出所需的最佳误差界

$n^{-1/2}\sqrt{\ln n}$ 。尽管如此, 如果 $\|(\theta^* - \theta)_S\|_\infty$ 减小, $\|p_S\|_\infty$ 也会减小。因此, 需要以下命题来确保 $\|(\theta^* - \theta)_S\|_\infty$ 的减小, 该命题与 Vaškevičius et al.^[79] 的第 B.3 节中的证明思路相似。

命题 5.2 (收缩单调性与收缩时间): 假设对于常数 $\eta \leq 1/8$ 和 $\varepsilon > 0$, 有

$$\|p_S\|_\infty \leq \eta \|\theta_S^* - \theta_S\|_\infty, \quad \|q_S + r_S\|_\infty \leq \varepsilon.$$

假设对于某个 $t_0 \geq 0$, 有

$$\|\theta_S^* - \theta_S(t_0)\|_\infty \leq M.$$

那么,

$$\|\theta_S^* - \theta_S(t)\|_\infty \leq M \vee 2\varepsilon, \quad \forall t \geq t_0. \quad (5.44)$$

此外,

- 如果 $\eta M \geq \varepsilon$, 那么

$$\|\theta_S^* - \theta_S(t)\|_\infty \leq \frac{1}{2}M, \quad \forall t \geq t_0 + \bar{T}^{\text{half}}(M), \quad (5.45)$$

其中

$$\bar{T}^{\text{half}}(M) = 4M^{-1} \left[2 + \frac{1}{2} \max_{k \in S} \left(\ln^+ \frac{M}{\lambda_k} + \ln^+ \frac{|\theta_k^*|}{\lambda_k} \right) \right]. \quad (5.46)$$

- 如果 $\eta M < \varepsilon$, 那么对于满足 $|\theta_k^*| \geq 4\varepsilon$ 的每个 $k \in S$, 有

$$|\theta_k^* - \theta_k(t)| \leq 4\varepsilon, \quad \forall t \geq t_0 + \bar{T}_k^{\text{fin}}, \quad (5.47)$$

其中

$$\bar{T}_k^{\text{fin}} \leq 4|\theta_k^*|^{-1} \left(2 + \frac{1}{2} \ln^+ \frac{M}{\lambda_k} + \frac{1}{2} \ln^+ \frac{|\theta_k^*|}{4\lambda_k} + \ln^+ \frac{|\theta_k^*|}{2\varepsilon} \right). \quad (5.48)$$

另一方面, 对于所有 $k \in S$, 有

$$|\theta_k^* - \theta_k(t)| \leq 2 \max(|\theta_k^*|, 4\varepsilon), \quad \forall t \geq t_0 + (2\varepsilon)^{-1}. \quad (5.49)$$

证明 证明依赖于第 5.6.5 小节中建立的一维方程分析。定义

$$t = \inf \{s \geq t_0 : \|\theta_S^* - \theta_S(s)\|_\infty > M \vee 2\varepsilon\}.$$

那么, 必然有 $\|\theta_S^* - \theta_S(t)\|_\infty = M \vee 2\varepsilon$, 这意味着对于某个索引 k , 有 $|\theta_k^* - \theta_k(t)| = M \vee 2\varepsilon$ 。由此给出

$$\|h_S(t)\|_\infty \leq \kappa := \eta(M \vee 2\varepsilon) + \varepsilon < M \vee 2\varepsilon.$$

注意到 $\eta < 1/2$ 。与引理 5.2 的证明类似, 回顾 (5.58):

$$\dot{\theta}_k(t) = (a_k^2(t) + \beta_k^2(t))(z - \theta_k(t) + h_k(t)).$$

因此, $\dot{\theta}_k(t)$ 与 $z - \theta_k(t)$ 符号相同, 从而 $|\theta_k^* - \theta_k(t)|$ 在 t 处必然是减小的, 这与 t 的定义矛盾。

现在证明第二部分。利用已获得的控制 (5.44), 得到

$$\|h_S(t)\|_\infty \leq \kappa, \quad \forall t \geq t_0.$$

如果 $\varepsilon \leq \eta M$, 有 $\kappa < 2\eta M \leq M/4$ 。在引理 5.3 和推论 5.2 中取 $\kappa' = M/4$, 在以下时间的最大值之后, 有 $|\theta_k^* - \theta_k(t)| \leq 2\kappa' = M/2$:

$$(\kappa')^{-1} \left[2 + \frac{1}{2} \ln^+ \frac{M}{\lambda_k} + \frac{1}{2} \ln^+ \frac{|\theta_k^*| - 2\kappa'}{\lambda_k} \right] \leq 4M^{-1} \left[2 + \frac{1}{2} \ln^+ \frac{M}{\lambda_k} + \frac{1}{2} \ln^+ \frac{|\theta_k^*|}{\lambda_k} \right]$$

或

$$(\kappa')^{-1} \left(1 + \ln^+ \frac{M}{\kappa'} \right) \leq 4M^{-1} \ln 4 \leq 8M^{-1}.$$

这表明有

$$\bar{T}^{\text{half}}(M) \leq 4M^{-1} \left[2 + \frac{1}{2} \max_{k \in S} \left(\ln^+ \frac{M}{\lambda_k} + \ln^+ \frac{|\theta_k^*|}{\lambda_k} \right) \right].$$

如果 $\varepsilon > \eta M$, 有 $\kappa < 2\varepsilon$ 。取 $\kappa' = 2\varepsilon$, 当 $|\theta_k^*| \geq 2\kappa'$ 时, 结果遵循推论 5.3 (其中取 $\delta = 2\varepsilon$)。对于最后一个陈述, 应用推论 5.4。 ■

基于前述命题, 可以考虑以下迭代收缩动力学。令 $\nu_1 = \varepsilon/\eta = Cn^{-1/2+s_1} \sqrt{\ln n}$ 。对于 $i = 0, 1, \dots, I+1$, 定义 $M_i = 2^{-i} B_\infty$, 其中 $I = \lfloor \log_2(B_\infty/\nu_1) \rfloor$ 。显然 $I \leq \log_2 \sqrt{n} + C$ 。由于 $\|\theta_S^* - \theta_S(0)\|_\infty = \|\theta_S^*\|_\infty \leq B_\infty$, 迭代地使用命题 5.2 得到

$$\|\theta_S^* - \theta_S(t)\|_\infty \leq M_{i+1}, \quad \forall t \geq \sum_{j=0}^i T^{(j)},$$

其中

$$T^{(i)} = 4M_i^{-1} \left[2 + \frac{1}{2} \max_{k \in S} \left(\ln^+ \frac{M_i}{\lambda_k} + \ln^+ \frac{|\theta_k^*|}{\lambda_k} \right) \right] \leq CM_i^{-1} \ln n,$$

因为 $\lambda_k \asymp k^{-\gamma}$ 且 $\max S \leq Cn^{\kappa/2}$ 。现在, 有

$$\bar{T} := \sum_{i=0}^I T^{(i)} \leq C \sum_{i=0}^I M_i^{-1} \ln n \leq C\nu_1^{-1} \ln n \leq Cn^{1/2-s_1} \ln n \leq Cn^{1/2}. \quad (5.50)$$

最后, 当 $t \geq \bar{T}$ 时, 有 $\eta M_{I+1} < \varepsilon$ 。选择 $\nu_2 = n^{-1/2} \ln n$, 利用命题 5.2 的第二

部分得到, 对于所有 $k \in J_{\text{sig}}(v_2)$,

$$|\theta_k^* - \theta_k(t)| \leq 4\varepsilon, \quad \forall t \geq \bar{T} + T^{\text{fin}},$$

其中

$$\begin{aligned} T^{\text{fin}} &\leq C \max_{k \in J_{\text{sig}}(v_2)} 4 |\theta_k^*|^{-1} \left[2 + \frac{1}{2} \ln^+ \frac{M_{I+1}}{\lambda_k} + \frac{1}{2} \ln^+ \frac{|\theta_k^*|}{4\lambda_k} + \ln^+ \frac{|\theta_k^*|}{2\varepsilon} \right] \\ &\leq C v_2^{-1} \ln n \leq C n^{1/2}. \end{aligned}$$

另一方面, 对于 $k \in S \setminus J_{\text{sig}}(v_2)$, 使用 (5.49) 得到

$$|\theta_k^* - \theta_k(t)| \leq 2 \max(|\theta_k^*|, 4\varepsilon),$$

前提是 $t \geq \bar{T} + C n^{1/2} \geq \bar{T} + (2\varepsilon)^{-1}$ 。回顾 (5.32) 中 S 的选择, 发现

$$\begin{aligned} \sum_{k \in S \setminus J_{\text{sig}}(v_2)} |\theta_k^* - \theta_k(t)|^2 &= \sum_{k \in S_1 \setminus J_{\text{sig}}(v_2)} |\theta_k^* - \theta_k(t)|^2 + \sum_{k \in S_2 \setminus J_{\text{sig}}(v_2)} |\theta_k^* - \theta_k(t)|^2 \\ &\leq C \sum_{k \in S_1 \setminus J_{\text{sig}}(v_2)} |\theta_k^*|^2 + C \sum_{k \in S_2 \setminus (S_1 \cup J_{\text{sig}}(v_2))} \varepsilon^2 \\ &\leq C \Psi(v_2) + C n^{\frac{1}{2\gamma}} \varepsilon^2. \end{aligned}$$

因此, 得出结论, 在 $t \geq \underline{C} n^{1/2}$ 之后, 有

$$\begin{aligned} \|\theta_S^* - \theta_S(t)\|_2^2 &= \sum_{k \in J_{\text{sig}}(v_2)} |\theta_k^* - \theta_k(t)|^2 + \sum_{k \in S \setminus J_{\text{sig}}(v_2)} |\theta_k^* - \theta_k(t)|^2 \\ &\leq C \varepsilon^2 |S| + C \Psi(v_2) + C n^{\frac{1}{2\gamma}} \varepsilon^2 \\ &= C \frac{\ln n}{n} \left(\Phi(n^{-1/2} \sqrt{\ln n}) + n^{\frac{1}{2\gamma}} \right) + C \Psi(n^{-1/2} \ln n). \end{aligned} \quad (5.51)$$

此处, 注意到常数 $\underline{C}$ 不依赖于 E 。

5.6.4.2 噪声项的界

在这一部分中, 将使用引理 5.8 来证明: 如果 E 选择得当, 对于某个常数 $\bar{C} > 0$, 有 $T^{\text{err}} \geq \bar{C} \sqrt{n}$ 。由于 $|\theta_k^*| \leq n^{-1/2} \sqrt{\ln n}$, 只需对项 $\int_0^{\bar{C} \sqrt{n}} |h_k(t)| \, dt$ 进行限制即可, 其中回顾 $h_k = p_k + q_k + r_k$ 。

记 $\bar{T}^{(i)} = \sum_{j=0}^i T^{(j)}$ 。现在, 对于 $t \in [\bar{T}^{(i)}, \bar{T}^{(i+1)}]$, 利用前一部分的结果, 有 $\|\theta_S^* - \theta_S(t)\|_\infty \leq M_i$, 因此 (5.42) 中的第一个不等式给出

$$|p_k(t)| \leq C M_i n^{-s} \sqrt{\ln k}, \quad t \in [\bar{T}^{(i)}, \bar{T}^{(i+1)}].$$

因此,

$$\begin{aligned} \int_0^{\bar{T}} |p_k(t)| \, dt &\leq C \sum_{i=0}^I M_i n^{-s} \sqrt{\ln k} \cdot T^{(i)} \leq C \sum_{i=0}^I M_i n^{-s} \sqrt{\ln k} \cdot M_i^{-1} \ln n \\ &\leq |I| n^{-s} \sqrt{\ln k} \ln n \leq C \sqrt{\ln k}. \end{aligned}$$

此外, 对于 $t \geq \bar{T}$, 有

$$|p_k(t)| \leq \eta_k \|(\theta^* - \theta)_S\|_\infty \leq C n^{-s} \sqrt{\ln(nk)} \cdot M_{I+1} \leq C n^{-1/2} \sqrt{\ln k},$$

故

$$\int_{\bar{T}}^{\bar{C}\sqrt{n}} |p_k(t)| \, dt \leq C \bar{C} \sqrt{\ln k}.$$

另一方面, (5.42) 中的另外两个不等式给出 $|q_k + r_k| \leq C n^{-1/2} \sqrt{\ln n + \ln k}$, 因此

$$\int_0^{\bar{C}\sqrt{n}} |q_k(t) + r_k(t)| \, dt \leq C \bar{C} \sqrt{\ln n + \ln k}.$$

结合这三个估计就证明了该声明。

5.6.4.3 泛化误差

根据 (5.51) 和第 5.6.4.2 小节, 可以选择一个常数 $\bar{C} \geq \underline{C}$, 然后固定常数 E . 现在, 当 $\underline{C}\sqrt{n} \leq t \leq \bar{C}\sqrt{n} \leq T^{\text{err}}$ 时, 有

$$\begin{aligned} \|\theta_R(t)\|_2^2 &\leq C \sum_{k \in R} \lambda_k^2 e^{2E\sqrt{\ln n + \ln k}} \\ &\leq C e^{2E\sqrt{\ln n}} \sum_{k \in R} \lambda_k^2 e^{2E\sqrt{\ln k}} \\ &\leq C e^{2E\sqrt{\ln n}} \sum_{k \geq L} \lambda_k^2 e^{2E\sqrt{\ln k}} \\ &\leq C e^{2E\sqrt{\ln n}} L^{1-2\gamma+s} \leq C n^{-1+\frac{1+s}{2\gamma}}. \end{aligned}$$

因此,

$$\|\theta_R^* - \theta_R(t)\|_2^2 \leq 2 \|\theta_R^*\|_2^2 + 2 \|\theta_R(t)\|_2^2 \leq \Psi(n^{-1/2} \sqrt{\ln n}) + C n^{-1+\frac{1+s}{2\gamma}}. \quad (5.52)$$

结合 (5.51) 和 (5.52), 得到所需结果:

$$\|f^* - \hat{f}_t^{\text{EigGD}}\|_{L^2}^2 = \|\theta^* - \theta(t)\|_2^2 \lesssim \frac{1}{n} \left[\Phi(n^{-1/2} \sqrt{\ln n}) \ln n + n^{\frac{1+s}{2\gamma}} \right] + \Psi(n^{-1/2} \ln n). \quad (5.53)$$

5.6.5 一维动力学

在本小节中, 将重点讨论带扰动的梯度流的一维动力学。令 $h(t) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ 为一个连续扰动函数。现在考虑以下受扰动的一维动力学:

$$\begin{cases} \dot{\beta}(t) = a(t)[z - \theta(t) + h(t)], & \beta(0) = 0, \\ \dot{a}(t) = \beta(t)[z - \theta(t) + h(t)], & a(0) = \lambda^{1/2}, \end{cases} \quad (5.54)$$

其中 $\theta(t) = a(t)\beta(t)$ 。

守恒量. 计算 $a^2(t)$ 和 $\beta^2(t)$ 的时间导数, 有

$$\frac{d}{dt}a^2(t) = \frac{d}{dt}\beta^2(t) = 2a(t)\beta(t)[z - \theta(t) + h(t)],$$

因此有守恒量

$$a^2(t) - \beta^2(t) = a^2(0) - \beta^2(0) = \lambda, \quad (5.55)$$

即 $a(t) = (\lambda + \beta^2(t))^{1/2}$, 从而

$$\max(\lambda^{1/2}, |\beta(t)|) \leq a(t) \leq \sqrt{2} \max(\lambda^{1/2}, |\beta(t)|), \quad (5.56)$$

$$\max(\lambda^{1/2}, |\beta(t)|) |\beta(t)| \leq |\theta(t)| \leq \sqrt{2} \max(\lambda^{1/2}, |\beta(t)|) |\beta(t)|. \quad (5.57)$$

θ 的演变. 直接计算可得

$$\begin{aligned} \dot{\theta} &= \dot{a}\beta + a\dot{\beta} = [\beta^2 + a^2](z - \theta + h) \\ &= \theta^2(a^{-2} + \beta^{-2})(z - \theta + h). \end{aligned} \quad (5.58)$$

引理 5.2 (扰动界): 考虑 (5.54)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 都有 $|h(t)| \leq \kappa$ 。那么, 对于 $t \geq t_0$,

(i) 只要 $|z - \theta(t)| \geq \kappa$, $|z - \theta(t)|$ 就是减小的。

(ii) 一旦对于某个 $t_1 \geq t_0$ 有 $|z - \theta(t_1)| \leq \kappa$, 则对于所有 $t \geq t_1$ 都有 $|z - \theta(t)| \leq \kappa$ 。

证明 不失一般性, 假设 $z \geq 0$ 。在 (5.58) 中回顾:

$$\dot{\theta} = \theta^2(a^{-2} + \beta^{-2})(z - \theta + h).$$

现在, 如果 $|z - \theta(t)| \geq \kappa$, 有 $\theta(t) \leq z - \kappa$ 或 $\theta(t) \geq z + \kappa$ 。对于第一种情况, $\dot{\theta} \geq 0$, 故 $\theta(t)$ 是增加的, 这意味着 $|z - \theta(t)| = z - \theta(t)$ 是减小的。同样, 对于第二种情况, $\dot{\theta} \leq 0$, 因此 $|z - \theta(t)|$ 也是减小的。由此, 得到了命题的第一部分。第二部分仅是第一部分的推论。 ■

引理 5.3 (从下方趋近, 大 κ): 考虑方程 (5.54) 且 $z \geq 0$ (对于 $z \leq 0$ 有类似结果)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 有 $|h(t)| \leq \kappa$, 且存在某

个 $M \geq 0$ 使得

$$z - M \leq \theta(t_0) \leq z.$$

那么, 有

$$|z - \theta(t)| \leq 2\kappa, \quad \forall t \geq t_0 + \bar{T}^{\text{app}},$$

其中

$$\bar{T}^{\text{app}} \leq \kappa^{-1} \left[2 + \frac{1}{2} \ln^+ \frac{M}{\lambda} + \frac{1}{2} \ln^+ \frac{z - 2\kappa}{\lambda} \right]. \quad (5.59)$$

证明 根据引理 5.2 可知, $|z - \theta(t)|$ 一直减小直到达到 κ , 随后保持在 κ 之下。因此, 只需给出 $\theta(t) = z - 2\kappa$ 的到达时间的估计。进一步定义

$$T^{\text{neg}} = \inf \{s \geq 0 : \beta(t_0 + s) \geq -\lambda^{1/2}\}, \quad T^{\text{esc}} = \inf \{s \geq 0 : \beta(t_0 + s) \geq \lambda^{1/2}\}$$

$$T^{\text{app}} = \inf \{s \geq 0 : |z - \theta(t_0 + s)| \leq 2\kappa\}.$$

同时回顾

$$\dot{\beta}(t) = a(t) [z - \theta(t) + h(t)] \geq \kappa a(t), \quad \text{对于 } t \leq T^{\text{app}}. \quad (5.60)$$

在下文中, 假设 $0 < T^{\text{neg}} < T^{\text{esc}} < T^{\text{app}}$ 并推导 T^{app} 的上界。容易看出, 所得估计对其他情况 (如 $T^{\text{esc}} = 0$ 或 $T^{\text{app}} \leq T^{\text{esc}}$) 同样有效。

阶段 1: 调整符号 有 $\beta(t_0) \leq -\lambda^{1/2}$, 因此 (5.56) 给出

$$z - M \leq \theta(t_0) \leq -\sqrt{2}\beta^2(t_0), \implies \beta(t_0) \geq -\sqrt{M}.$$

同时, 结合 (5.60) 和 (5.56) 得到

$$\dot{\beta}(t_0 + s) \geq -\kappa \beta(t_0 + s), \quad \text{对于 } s \in [0, T^{\text{neg}}],$$

因此

$$\beta(t_0 + s) \geq \beta(t_0) e^{-\kappa s} \geq -\sqrt{M} e^{-\kappa s}, \quad \text{对于 } s \in [0, T^{\text{neg}}],$$

从而

$$T^{\text{neg}} \leq \frac{1}{2} \kappa^{-1} \ln \frac{M}{\lambda}.$$

阶段 2: 脱离原点 现在考虑 $t \geq T^{\text{neg}}$ 。将 $a \geq \lambda^{1/2}$ 代入 (5.60) 推导出

$$\dot{\beta} \geq \lambda^{\frac{1}{2}} \kappa,$$

这意味着

$$T^{\text{esc}} - T^{\text{neg}} \leq 2\kappa^{-1}.$$

阶段 3: 趋近 z 现在考虑 $t \geq T^{\text{esc}}$ 。利用 $a \geq \beta$ 得到

$$\dot{\beta} \geq \kappa \beta,$$

这意味着

$$\beta(t_0 + T^{\text{esc}} + s) \geq \lambda^{1/2} e^{\kappa s}, \quad \forall s \in [0, T^{\text{app}} - T^{\text{esc}}].$$

注意到 $\theta = a\beta \geq \beta^2$, 有

$$\beta \geq \sqrt{z - 2\kappa} \implies \theta \geq z - 2\kappa,$$

从而

$$T^{\text{app}} - T^{\text{esc}} \leq \kappa^{-1} \ln \frac{z - 2\kappa}{\lambda^{1/2}}.$$

综上所述, 有

$$T^{\text{esc}} \leq \kappa^{-1} \left[2 + \frac{1}{2} \ln^+ \frac{M}{\lambda} + \frac{1}{2} \ln^+ \frac{z - 2\kappa}{\lambda} \right].$$

■

引理 5.4 (从下方达到信号时间, 小 κ): 考虑与引理 5.3 相同的设置。进一步假设 $z \geq 2\kappa$ 。那么,

$$\theta(t) \geq z/4, \quad \forall t \geq t_0 + \bar{T}^{\text{sig}},$$

其中

$$\bar{T}^{\text{sig}} = 4z^{-1} \left[2 + \frac{1}{2} \ln^+ \frac{M}{\lambda} + \frac{1}{2} \ln^+ \frac{z}{4\lambda} \right].$$

证明 证明类似于引理 5.3, 但对该项使用不同的控制:

$$z - \theta(t) + h(t) \geq \frac{3}{4}z - \kappa \geq \frac{1}{4}z, \quad \text{对于 } t \in [t_0, T^{\text{sig}}],$$

其中

$$T^{\text{sig}} = \inf \{s \geq 0 : \theta(t_0 + s) \geq z/4\}.$$

遵循引理 5.3 的证明, 有

$$T^{\text{neg}} \leq 2z^{-1} \ln \frac{M}{\lambda}, \quad T^{\text{esc}} - T^{\text{neg}} \leq 8z^{-1}, \quad T^{\text{sig}} - T^{\text{esc}} \leq 2z^{-1} \ln^+ \frac{z}{4\lambda},$$

由此得出所需结果。 ■

引理 5.5 (从上方趋近): 考虑方程 (5.54) 且 $z \geq 0$ (对于 $z \leq 0$ 有类似结果)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 有 $|h(t)| \leq \kappa$, 且 $\theta(t_0) \geq z$ 。那么, 有

$$|z - \theta(t)| \leq 2 \max(z, 2\kappa), \quad \forall t \geq t_0 + \max(z, 2\kappa)^{-1}.$$

证明 类似于引理 5.3 的证明, 根据引理 5.2, 只需考虑 $M > 2 \max(z, 2\kappa)$ 的情况, 并给出 $\theta(t_0 + s) = z + 2 \max(z, 2\kappa)$ 的到达时间 (记为 T^{sig}) 的控制。

回顾 (5.58) 以及 $\theta \geq \beta^2$ 的事实得出

$$\dot{\theta} = \theta^2(a^{-2} + \beta^{-2})(z - \theta + h) \leq -\theta(\theta - z - \kappa).$$

注意到在 T^{sig} 之前, 有

$$\frac{1}{2}\theta \geq \frac{1}{2}z + \max(z, 2\kappa) \geq z + \kappa,$$

从而 $\theta - z - \kappa \geq \frac{1}{2}\theta$ 。由此得出

$$\dot{\theta}(t_0 + s) \leq -\frac{1}{2}\theta^2(t_0 + s), \quad \text{对于 } s \in [0, T^{\text{app}}],$$

这意味着

$$T^{\text{sig}} \leq 2 \left((z + 2 \max(z, 2\kappa))^{-1} - (z + M)^{-1} \right) \leq \max(z, 2\kappa)^{-1}.$$

■

引理 5.6 (z 附近的近似时间): 考虑方程 (5.54) 且 $z \geq 0$ (对于 $z \leq 0$ 有类似结果)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 有 $|h(t)| \leq \kappa$ 。同时假设

$$\frac{1}{4}z \leq \theta(t_0) \leq 3z.$$

那么, 对于任何 $\delta > 0$, 有

$$|z - \theta(t)| \leq \kappa + \delta, \quad \forall t \geq t_0 + 4z^{-1} \ln^+ \frac{|z - \theta(t_0)| - \kappa}{\delta}.$$

证明 定义

$$T^{\text{app}} = \inf \{ s \geq 0 : |z - \theta(t_0 + s)| \leq \kappa + \delta \}.$$

利用引理 5.2, 只需在 T^{app} 不为零时证明其上界, 因此根据单调性, 对于所有 $t \geq t_0$, 都有 $\frac{1}{4}z \leq \theta(t) \leq 3z$ 。

现在, (5.57) 中的下界 $\beta^2 \leq \theta$ 给出 $\beta^{-2} \geq \theta^{-1}$ 。将这些代入 (5.58), 有

$$\theta^2(a^{-2} + \beta^{-2}) \geq \theta \geq z/4,$$

从而

$$\begin{aligned} \dot{\theta} &\geq \frac{1}{4}z(z - \kappa - \theta), \quad \text{若 } \theta(t_0) \leq z - \kappa, \\ \dot{\theta} &\leq -\frac{1}{4}z(\theta - z - \kappa), \quad \text{若 } \theta(t_0) \geq z + \kappa. \end{aligned}$$

这表明对于 $s \in [0, T^{\text{app}}]$,

$$z - \kappa - \theta(t_0 + s) \leq (z - \kappa - \theta(t_0))e^{-\frac{1}{4}zs}, \quad \text{若 } \theta(t_0) \leq z - \kappa,$$

$$\theta(t_0 + s) - z - \kappa \leq (\theta(t_0) - z - \kappa)e^{-\frac{1}{4}zs}, \quad \text{若 } \theta(t_0) \geq z + \kappa,$$

因此

$$T^{\text{app}} \leq 4z^{-1} \ln \frac{|z - \theta(t_0)| - \kappa}{\delta}.$$

■

推论 5.2: 在与引理 5.5 相同的设置下, 有

$$|z - \theta(t)| \leq 4\kappa, \quad \forall t \geq t_0 + \kappa^{-1} \left(1 + \ln^+ \frac{|z - \theta(t_0)|}{\kappa} \right).$$

证明 引理 5.5 给出: 对于 $t \geq t_1 = t_0 + \max(z, 2\kappa)^{-1}$, 有

$$|z - \theta(t)| \leq 2 \max(z, 2\kappa).$$

现在, 如果 $z \leq 2\kappa$, 则结论成立。否则, 有 $z \geq 2\kappa$, 可以进一步应用引理 5.6 得到所需结果, 其中注意到单调性也意味着 $|z - \theta(t_1)| \leq \max(|z - \theta(t_0)|, \kappa)$ 。 ■

推论 5.3: 考虑方程 (5.54) 且 $z \geq 0$ (对于 $z \leq 0$ 有类似结果)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 有 $|h(t)| \leq \kappa$, 且 $z \geq 2\kappa$ 。那么,

$$|z - \theta(t)| \leq \kappa + \delta, \quad \forall t \geq t_0 + \bar{T}^{\text{app}}(\delta),$$

其中

$$\bar{T}^{\text{app}}(\delta) \leq 4z^{-1} \left[2 + \frac{1}{2} \ln^+ \frac{|z - \theta(t_0)|}{\lambda} + \frac{1}{2} \ln^+ \frac{z}{4\lambda} + \ln^+ \frac{2z}{\delta} \right].$$

证明 这是引理 5.4、引理 5.5 和引理 5.6 的推论。 ■

引理 5.7 (从负值回撤): 考虑方程 (5.54) 且 $z \geq 0$ (对于 $z \leq 0$ 有类似结果)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 都有 $|h(t)| \leq \kappa$ 。那么, 有

$$\theta(t) \geq -2\kappa, \quad \forall t \geq t_0 + \kappa^{-1}.$$

证明 证明类似于引理 5.5, 但现在处理 $\theta(t) \leq 0$ 的情况。首先回顾 (5.58)。 $|\theta| \geq \beta^2$ 给出 $\theta^2(a^{-2} + \beta^{-2}) \geq |\theta|$ 。当 $\theta(t) \leq -2\kappa$ 时, 有 $z - \theta(t) - \kappa \geq \theta(t)/2$, 因此

$$\dot{\theta} \geq |\theta(t)| (z - \theta(t) - \kappa) \geq \frac{1}{2} \theta^2(t),$$

这意味着

$$\theta(t_0 + s) \geq -[s/2 - \theta(t_0)^{-1}]^{-1},$$

从而对于所有 $s \geq \kappa^{-1}$, 有 $\theta(t_0 + s) \geq -2\kappa$ 。 ■

推论 5.4 (回撤至 κ): 考虑方程 (5.54)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于

所有 $t \geq t_0$ 都有 $|h(t)| \leq \kappa$ 。那么,

$$|z - \theta(t)| \leq 2 \max(|z|, 2\kappa), \quad \forall t \geq t_0 + \kappa^{-1}.$$

证明 若 $\theta(t_0) \geq z$, 使用引理 5.5。若 $\theta(t_0) \leq 0$, 使用引理 5.7 并注意到 $|z - \theta(t)| \leq |z| + |\theta(t)|$ 。 ■

引理 5.8 (误差控制): 考虑方程 (5.54)。有

$$|\beta(t)| \leq \lambda^{1/2} e^{\sqrt{2} \int_0^t (|h(s)| + |z|) ds}, \quad (5.61)$$

$$|\theta(t)| \leq \sqrt{2} \lambda e^{2\sqrt{2} \int_0^t (|h(s)| + |z|) ds}. \quad (5.62)$$

证明 不失一般性, 假设 $z \geq 0$ 。回顾方程 (5.54):

$$\dot{\beta}(t) = a(t) [z - \theta(t) + h(t)].$$

为了给出 $|\beta|$ 的控制, 注意到 $a(t)$ 始终为正, 因此 $\theta(t)$ 与 $\beta(t)$ 符号相同。现在, 如果 $\theta(t) \geq 0$, 则有 $|z - \theta(t) + h(t)| \leq z + |h(t)|$ 。另一方面, 如果 $\theta(t) \leq 0$ 从而 $\beta(t) \leq 0$, 则有 $z - \theta(t) + h(t) \geq -|h(t)|$ 。因此, 得出结论:

$$|\beta(t)| \leq \int_0^t a(s) (z + |h(s)|) ds.$$

现在考虑

$$T^{\text{esc}} = \inf \{t \geq 0 : |\beta(t)| \geq \lambda^{1/2}\}.$$

当 $t \leq T^{\text{esc}}$ 时, (5.55) 意味着 $a \leq \sqrt{2} \lambda^{1/2}$ 。因此有

$$|\beta(t)| \leq \sqrt{2} \lambda^{1/2} \int_0^t (|h(s)| + |z|) ds, \quad t \in [0, T^{\text{esc}}],$$

从而

$$T^{\text{esc}} \geq \inf \left\{ t \geq 0 : \sqrt{2} \int_0^t (|h(s)| + |z|) ds = 1 \right\}.$$

在 T^{esc} 之后, 应用 $a \leq \sqrt{2} |\beta|$ 得到

$$|\beta(t)| \leq \lambda^{1/2} + \sqrt{2} \int_{T^{\text{esc}}}^t |\beta(s)| (z + |h(s)|) ds, \quad t \geq T^{\text{esc}},$$

从而

$$|\beta(t)| \leq \lambda^{1/2} e^{\sqrt{2} \int_{T^{\text{esc}}}^t (|h(s)| + |z|) ds}, \quad t \geq T^{\text{esc}}.$$

利用 $\theta = a\beta \leq \sqrt{2} \beta^2$, 得到对于 $t \geq T^{\text{esc}}$:

$$|\theta(t)| \leq \sqrt{2} \lambda e^{2\sqrt{2} \int_{T^{\text{esc}}}^t (|h(s)| + |z|) ds} \leq \sqrt{2} \lambda e^{2\sqrt{2} \int_0^t (|h(s)| + |z|) ds}.$$

最后, 注意到该不等式对 $t \leq T^{\text{esc}}$ 同样有效, 因为此时 $|\theta(t)| \leq \sqrt{2} \lambda^{1/2}$ 。

■

5.7 多层参数化的泛化能力证明

在本节中提供多层模型的证明，它类似于第 5.6 节中两层模型的证明，但更为复杂。记

$$\theta(t) = (\theta_j)_{j \geq 1} = a(t) \odot b^D(t) \odot \beta(t), \quad (5.63)$$

其中用 b^D 表示 b 的逐元素幂。那么，类似于两层模型，有

$$f_t(x) = \langle \Phi_{a(t), b(t)}(x), \beta(t) \rangle = \langle E(x), a(t) \odot b^{\odot D}(t) \odot \beta(t) \rangle = \langle E(x), \theta(t) \rangle = E(x)^\top \theta(t),$$

接着，遵循第 5.6.1 小节，可以推导出多层模型的梯度流：

$$\begin{cases} \dot{a} = b^D \odot \beta \odot \Delta, & a_j(0) = \lambda_j^{\frac{1}{2}} \\ \dot{b} = Da \odot b^{D-1} \odot \beta \odot \Delta, & b_j(0) = b_0 \\ \dot{\beta} = a \odot b^D \odot \Delta, & \beta(0) = 0, \end{cases} \quad (5.64)$$

其中

$$\Delta := \hat{\Sigma}(\theta^* - \theta) + r, \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n E(x_i)E(x_i)^\top, \quad r = \frac{1}{n} \sum_{i=1}^n \varepsilon_i E(x_i).$$

5.7.1 多层模型的方程分析

考虑多层模型的受扰动一维方程。令 $h(t) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ 为连续扰动函数。现在考虑以下受扰动的一维动力学：

$$\begin{cases} \dot{a} = b^D \beta (z - \theta + h), & a(0) = \lambda^{1/2} > 0, \\ \dot{b} = Dab^{D-1} \beta (z - \theta + h), & b(0) = b_0, \\ \dot{\beta} = ab^D (z - \theta + h), & \beta(0) = 0, \end{cases} \quad (5.65)$$

其中 $\theta(t) = a(t)b^D(t)\beta(t)$ 。

守恒量. 易见

$$\frac{d}{dt} a^2 = \frac{1}{D} \frac{d}{dt} b^2 = \frac{d}{dt} \beta^2 = 2ab^D \beta (z - \theta + h).$$

因此，有

$$a^2(t) - \beta^2(t) \equiv \lambda, \quad b^2(t) - D\beta^2(t) \equiv b_0^2. \quad (5.66)$$

利用这一点得到

$$a(t) = (\beta^2(t) + \lambda)^{1/2} > 0, \quad b(t) = (D\beta^2(t) + b_0^2)^{1/2} > 0. \quad (5.67)$$

利用这些守恒量, 本文可以证明以下关于 β 的估计:

$$\begin{aligned} \max(\lambda^{\frac{1}{2}}, |\beta|) &\leq a \leq \sqrt{2} \max(\lambda^{\frac{1}{2}}, |\beta|), \\ \max(b_0, \sqrt{D} |\beta|) &\leq b \leq \sqrt{2} \max(b_0, \sqrt{D} |\beta|), \end{aligned} \quad (5.68)$$

这也意味着 $|\theta| = |ab^D \beta| \geq D^{D/2} |\beta|^{D+2}$ 。

θ 的演变. 直接计算可得

$$\begin{aligned} \dot{\theta} &= \dot{a}b^D \beta + aDb^{D-1} \dot{b} \beta + ab^D \dot{\beta} \\ &= [(b^D \beta)^2 + (Dab^{D-1} \beta)^2 + (ab^D)^2] (z - \theta + h) \\ &= \theta^2 (a^{-2} + Db^{-2} + \beta^{-2}) (z - \theta + h). \end{aligned} \quad (5.69)$$

方程的以下估计将在证明中频繁使用. 利用 (5.68), 有

$$|\theta| = |ab^D \beta| \geq D^{D/2} |\beta|^{D+2}, \implies |\beta| \leq (D^{-D/2} |\theta|)^{1/(D+2)}.$$

因此,

$$\theta^2 (a^{-2} + Db^{-2} + \beta^{-2}) \geq \theta^2 \beta^{-2} \geq D^{-\frac{D}{D+2}} |\theta|^{\frac{2D+2}{D+2}}. \quad (5.70)$$

引理 5.9 (扰动界): 考虑方程 (5.65). 令 $t_0 \geq 0$. 假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 都有 $|h(t)| \leq \kappa$. 那么, 对于 $t \geq t_0$,

- (i) 只要 $|z - \theta(t)| \geq \kappa$, $|z - \theta(t)|$ 就是减小的。
- (ii) 一旦对于某个 $t_1 \geq t_0$ 有 $|z - \theta(t_1)| \leq \kappa$, 则对于所有 $t \geq t_1$ 都有 $|z - \theta(t)| \leq \kappa$ 。

证明 与引理 5.2 相同. ■

引理 5.10 (从下方趋近, 大 κ): 考虑方程 (5.65) 且 $z \geq 0$ (对于 $z \leq 0$ 有类似结果). 令 $t_0 \geq 0$. 假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 有 $|h(t)| \leq \kappa$, 且存在某个 $M \geq 0$ 使得

$$z - M \leq \theta(t_0) \leq z.$$

那么, 有

$$|z - \theta(t)| \leq 2\kappa, \quad \forall t \geq t_0 + \bar{T}^{\text{app}},$$

其中

$$\bar{T}^{\text{app}} \leq 4 [D^{D/2} \kappa R^D]^{-1} \ln \frac{R}{r}, \quad (5.71)$$

且 $r = \min(\lambda^{1/2}, b_0/\sqrt{D})$, $R = \max(\lambda^{1/2}, b_0/\sqrt{D})$ 。

证明 证明类似于引理 5.3, 但由于存在 b 项而更为复杂. 记

$$r = \min(\lambda^{1/2}, b_0/\sqrt{D}), \quad R = \max(\lambda^{1/2}, b_0/\sqrt{D}),$$

进一步定义

$$\begin{aligned} T^{\text{neg},2} &= \inf \{s \geq 0 : \beta(t_0 + s) \geq -R\}, \quad T^{\text{neg},1} = \inf \{s \geq 0 : \beta(t_0 + s) \geq -r\}, \\ T^{\text{esc},1} &= \inf \{s \geq 0 : \beta(t_0 + s) \geq r\}, \quad T^{\text{esc},2} = \inf \{s \geq 0 : \beta(t_0 + s) \geq R\}, \\ T^{\text{app}} &= \inf \{s \geq 0 : |z - \theta(t_0 + s)| \leq 2\kappa\}, \end{aligned}$$

并且假设以下情况时，只需给出 T^{app} 的估计：

$$0 < T^{\text{neg},2} < T^{\text{neg},1} < T^{\text{esc},1} < T^{\text{esc},2} < T^{\text{app}}.$$

还回顾

$$\dot{\beta}(t) = a(t)b^D(t)[z - \theta(t) + h(t)] \geq \kappa a(t)b^D(t), \quad \text{对于 } t \leq T^{\text{app}}. \quad (5.72)$$

阶段 $0 \leq s \leq T^{\text{neg},2}$ 在 $T^{\text{neg},2}$ 之前，有 $\beta(t) \leq -R$ ，因此 (5.68) 给出

$$z - M \leq \theta(t_0) \leq -D^{\frac{D}{2}} 2^{\frac{D+1}{2}} |\beta(t_0)|^{D+2}, \quad \Rightarrow \beta(t_0) \geq -\left(D^{-\frac{D}{2}} 2^{-\frac{D+1}{2}} M\right)^{\frac{1}{D+2}}.$$

此外，结合 (5.72) 和 (5.68)（即 $a \geq |\beta|$ 且 $b \geq \sqrt{D}|\beta|$ ）得到

$$\dot{\beta}(t_0 + s) \geq D^{\frac{D}{2}} \kappa |\beta(t_0 + s)|^{D+1}, \quad \text{对于 } s \in [0, T^{\text{neg},2}],$$

故

$$|\dot{\beta}(t_0 + s)| \leq \left[|\beta(t_0)|^{-D} + DD^{\frac{D}{2}} \kappa s\right]^{-\frac{1}{D}}, \quad \text{对于 } s \in [0, T^{\text{neg},2}],$$

这意味着

$$T^{\text{neg},2} \leq \left(D^{\frac{D+2}{2}} R^D \kappa\right)^{-1}.$$

阶段 $T^{\text{neg},2} \leq s \leq T^{\text{neg},1}$ 在 $T^{\text{neg},2}$ 之后，考虑 $\lambda^{1/2} \leq b_0/\sqrt{D}$ 和 $\lambda^{1/2} \geq b_0/\sqrt{D}$ 两种情况。可以利用 (5.68) 中的替代下界得到

$$\dot{\beta}(t_0 + s) \geq \kappa b_0^D |\beta(t_0 + s)|,$$

或

$$\dot{\beta}(t_0 + s) \geq \kappa \lambda^{1/2} D^{\frac{D}{2}} |\beta(t_0 + s)|^D.$$

因此，对于第一种情况，有

$$T^{\text{neg},1} - T^{\text{neg},2} \leq (\kappa b_0^D)^{-1} \ln \frac{R}{r}.$$

对于第二种情况，有

$$T^{\text{neg},1} - T^{\text{neg},2} \leq \begin{cases} (\kappa \lambda^{1/2})^{-1} \ln \frac{R}{r}, & \text{若 } D = 1, \\ \left[(D-1)\kappa D^{\frac{D}{2}} \lambda^{\frac{D}{2}}\right]^{-1}, & \text{若 } D \geq 2. \end{cases}$$

结合这两种情况, 有

$$T^{\text{esc},2} - T^{\text{esc},1} \leq (\kappa(D^{1/2}R)^D)^{-1} \ln \frac{R}{r}.$$

阶段 $T^{\text{neg},1} \leq s \leq T^{\text{esc},1}$ 现在利用估计

$$\dot{\beta}(t_0 + s) \geq \kappa \lambda^{1/2} b_0^D,$$

就有

$$T^{\text{esc},1} - T^{\text{neg},1} \leq 2r (\kappa \lambda^{1/2} b_0^D)^{-1}.$$

阶段 $T^{\text{esc},1} \leq s \leq T^{\text{esc},2}$ 这一阶段与 $T^{\text{neg},2} \leq s \leq T^{\text{neg},1}$ 阶段非常相似, 有

$$T^{\text{esc},2} - T^{\text{esc},1} \leq (\kappa(D^{1/2}R)^D)^{-1} \ln \frac{R}{r}.$$

阶段 $T^{\text{esc},2} \leq s \leq T^{\text{app}}$ 这一阶段类似于 $0 \leq s \leq T^{\text{neg},2}$ 阶段, 可以得到

$$\beta(t_0 + T^{\text{esc},2} + s_1) \geq \left[\beta(t_0 + T^{\text{esc},2}) - D^{\frac{D+2}{2}} \kappa s_1 \right]^{-1/D},$$

从而

$$T^{\text{app}} - T^{\text{esc},2} \leq \left(D^{\frac{D+2}{2}} R^D \kappa \right)^{-1}.$$

最后, 注意到主导项是 $(\kappa(D^{1/2}R)^D)^{-1} \ln(R/r)$.

■

引理 5.11 (从下方达到信号时间, 小 κ): 考虑与引理 5.10 相同的设置。进一步假设 $z \geq 2\kappa$ 。那么,

$$\theta(t) \geq z/4, \quad \forall t \geq t_0 + \bar{T}^{\text{sig}},$$

其中

$$\bar{T}^{\text{sig}} = 4 [D^{D/2} z R^D]^{-1} \ln \frac{R}{r},$$

且 $r = \min(\lambda^{1/2}, b_0/\sqrt{D})$, $R = \max(\lambda^{1/2}, b_0/\sqrt{D})$ 。

证明 证明类似于引理 5.10, 但对该项使用不同的估计:

$$z - \theta(t) + h(t) \geq \frac{3}{4}z - \kappa \geq \frac{1}{4}z, \quad \text{对于 } t \in [t_0, T^{\text{sig}}],$$

其中

$$T^{\text{sig}} = \inf \{s \geq 0 : \theta(t_0 + s) \geq z/4\},$$

故将 κ 替换为 $z/4$ 即可得到结果。

■

引理 5.12 (从上方趋近): 考虑方程 (5.65) 且 $z \geq 0$ (对于 $z \leq 0$ 有类似结果)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 有 $|h(t)| \leq \kappa$, 且 $\theta(t_0) \geq z$ 。那么,

记 $m = 2 \max(z, 2\kappa)$, 有

$$|z - \theta(t)| \leq m, \quad \forall t \geq t_0 + \frac{D+2}{D+1} D^{\frac{D}{D+2}} m^{-\frac{2D+2}{D+2}}.$$

证明 证明类似于引理 5.5。定义

$$T^{\text{sig}} = \inf \left\{ s \geq 0 : \theta(t_0 + s) \leq z + 2 \max(z, 2\kappa) \right\},$$

有

$$\frac{1}{2} \theta(t_0 + s) \geq \frac{1}{2} z + \max(z, 2\kappa) \geq z + \kappa, \quad \text{对于 } s \in [0, T^{\text{sig}}],$$

因此 $\theta - z - \kappa \geq \frac{1}{2} \theta$ 。将此以及 (5.70) 代入 (5.69), 得到

$$\begin{aligned} \dot{\theta} &= \theta^2(a^{-2} + Db^{-2} + \beta^{-2})(z - \theta + h) \\ &\leq -D^{-\frac{D}{D+2}} \theta^{\frac{2D+2}{D+2}} \cdot \frac{1}{2} \theta \\ &= -\frac{1}{2} D^{-\frac{D}{D+2}} \theta^{1+\frac{2D+2}{D+2}}, \end{aligned}$$

从而通过命题 5.6 得到结果。 ■

引理 5.13 (z 附近的近似时间): 考虑方程 (5.65) 且 $z \geq 0$ (对于 $z \leq 0$ 有类似结果)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 有 $|h(t)| \leq \kappa$ 。同时假设

$$\frac{1}{4}z \leq \theta(t_0) \leq 3z.$$

那么, 对于任何 $\delta > 0$, 有

$$|z - \theta(t)| \leq \kappa + \delta, \quad \forall t \geq t_0 + 4^{\frac{2D+2}{D+2}} D^{\frac{D}{D+2}} z^{-\frac{2D+2}{D+2}} \ln^+ \frac{|z - \theta(t_0)| - \kappa}{\delta}.$$

证明 证明类似于引理 5.6: 定义

$$T^{\text{app}} = \inf \left\{ s \geq 0 : |z - \theta(t_0 + s)| \leq \kappa + \delta \right\},$$

并在 T^{app} 不为零时提供其上界。根据单调性, 对于所有 $t \geq t_0$ 都有 $\frac{1}{4}z \leq \theta(t) \leq 3z$ 。

现在, 回顾 (5.69) 和 (5.70):

$$\dot{\theta} = \theta^2(a^{-2} + Db^{-2} + \beta^{-2})(z - \theta + h),$$

且

$$\theta^2(a^{-2} + Db^{-2} + \beta^{-2}) \geq D^{-\frac{D}{D+2}} \theta^{\frac{2D+2}{D+2}} \geq 4^{-\frac{2D+2}{D+2}} D^{-\frac{D}{D+2}} z^{\frac{2D+2}{D+2}} =: k,$$

这里利用了 $\theta \geq z/4$ 的事实。

将其代入 $\theta(t)$ 的方程并考虑 $\theta(t_0) \leq z - \kappa$ 和 $\theta(t_0) \geq z + \kappa$ 两种情况, 有

$$z - \kappa - \theta(t_0 + s) \leq (z - \kappa - \theta(t_0))e^{-ks}, \quad \text{若 } \theta(t_0) \leq z - \kappa,$$

$$\theta(t_0 + s) - z - \kappa \leq (\theta(t_0) - z - \kappa)e^{-ks}, \quad \text{若 } \theta(t_0) \geq z + \kappa,$$

从而

$$T^{\text{app}} \leq k^{-1} \ln \frac{|z - \theta(t_0)| - \kappa}{\delta}.$$

■

推论 5.5: 在与引理 5.12 相同的设置下, 有

$$|z - \theta(t)| \leq 4\kappa, \quad \forall t \geq t_0 + C_D \kappa^{-\frac{2D+2}{D+2}} \left(1 + \ln^+ \frac{|z - \theta(t_0)|}{\kappa} \right),$$

其中 C_D 是取决于 D 的常数。

证明 类似于推论 5.2 的证明, 但使用引理 5.12 和引理 5.13。 ■

推论 5.6: 考虑方程 (5.65) 且 $z \geq 0$ (对于 $z \leq 0$ 有类似结果)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 有 $|h(t)| \leq \kappa$, 且 $z \geq 2\kappa$ 。那么,

$$|z - \theta(t)| \leq \kappa + \delta, \quad \forall t \geq t_0 + \bar{T}^{\text{app}}(\delta),$$

其中

$$\bar{T}^{\text{app}}(\delta) \leq C_{D,1} z^{-1} R^{-D} \ln \frac{R}{r} + C_{D,2} z^{-\frac{2D+2}{D+2}} \ln^+ \frac{2z}{\delta},$$

$r = \min(\lambda^{1/2}, b_0/\sqrt{D})$, $R = \max(\lambda^{1/2}, b_0/\sqrt{D})$ 且 $C_{D,1}, C_{D,2}$ 是仅取决于 D 的两个常数。

证明 这是引理 5.11、引理 5.12 和引理 5.13 的推论。 ■

引理 5.14 (从负值回撤): 考虑方程 (5.65) 且 $z \geq 0$ (对于 $z \leq 0$ 有类似结果)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 都有 $|h(t)| \leq \kappa$ 。那么, 有

$$\theta(t) \geq -2\kappa, \quad \forall t \geq t_0 + \frac{D+2}{D+1} D^{\frac{D}{D+2}} (2\kappa)^{-\frac{2D+2}{D+2}}.$$

证明 通过利用引理 5.12 的证明对引理 5.7 的证明进行修改即可完成。 ■

推论 5.7 (回撤至 κ): 考虑方程 (5.65)。令 $t_0 \geq 0$ 。假设存在某个 $\kappa > 0$ 使得对于所有 $t \geq t_0$ 都有 $|h(t)| \leq \kappa$ 。那么,

$$|z - \theta(t)| \leq 2 \max(|z|, 2\kappa), \quad \forall t \geq t_0 + \frac{D+2}{D+1} D^{\frac{D}{D+2}} (2\kappa)^{-\frac{2D+2}{D+2}}.$$

证明 若 $\theta(t_0) \geq z$, 使用引理 5.12。若 $\theta(t_0) \leq 0$, 使用引理 5.14 并注意到 $|z - \theta(t)| \leq |z| + |\theta(t)|$ 。 ■

引理 5.15 (误差控制): 考虑方程 (5.65)。只要

$$2^{\frac{D+1}{2}} \lambda^{1/2} b_0^D \int_0^t (z + |h(s)|) \, ds \leq \min(\lambda^{1/2}, b_0/\sqrt{D}),$$

就得到

$$|\theta(t)| \leq 2^{\frac{D+1}{2}} \lambda b_0^D \int_0^t (|z| + |h(s)|) \, ds.$$

此外, 如果 $\lambda^{1/2} \leq b_0/\sqrt{D}$ 且

$$2^{\frac{D+1}{2}} b_0^D \int_0^t (|z| + |h(s)|) \, ds \leq \ln \frac{b_0/\sqrt{D}}{\lambda^{1/2}}, \quad (5.73)$$

有

$$|\beta(t)| \leq \lambda^{1/2} e^{2^{\frac{D+1}{2}} b_0^D \int_0^t (|z| + |h(\tau)|) \, d\tau} \quad (5.74)$$

$$|\theta(t)| \leq 2^{\frac{D+1}{2}} \lambda b_0^D e^{2^{\frac{D+3}{2}} b_0^D \int_0^t (|z| + |h(s)|) \, ds}. \quad (5.75)$$

证明 不失一般性, 假设 $z \geq 0$ 。类似于引理 5.8 证明中的论据, 从 (5.65) 得到

$$|\beta(t)| \leq \int_0^t a(s) b^D(s) (z + |h(s)|) \, ds. \quad (5.76)$$

记

$$r = \min(\lambda^{1/2}, b_0/\sqrt{D}), \quad R = \max(\lambda^{1/2}, b_0/\sqrt{D}),$$

定义

$$T^{(1)} = \inf \{t \geq 0 : |\beta(t)| \geq r\}, \quad T^{(2)} = \inf \{t \geq 0 : |\beta(t)| \geq R\}.$$

当 $t \leq T^{(1)}$ 时, 利用 (5.68) 得到 $a(t) \leq \sqrt{2}\lambda^{1/2}$ 且 $b(t) \leq \sqrt{2}b_0$, 故

$$|\beta(t)| \leq 2^{\frac{D+1}{2}} \lambda^{1/2} b_0^D \int_0^t (z + |h(s)|) \, ds,$$

这表明

$$T^{(1)} \geq \inf \left\{ t \geq 0 : 2^{\frac{D+1}{2}} \lambda^{1/2} b_0^D \int_0^t (z + |h(s)|) \, ds \geq r \right\}.$$

当 $t > T^{(1)}$ 时, 考虑 $\lambda^{1/2} \leq b_0/\sqrt{D}$ 的情况。那么有 $a(t) \leq \sqrt{2}|\beta(t)|$ 且 $b(t) \leq \sqrt{2}b_0$, 故

$$|\beta(t)| \leq \lambda^{1/2} + 2^{\frac{D+1}{2}} b_0^D \int_{T^{(1)}}^t |\beta(s)| (z + |h(s)|) \, ds,$$

从而

$$|\beta(T^{(1)} + s)| \leq \lambda^{1/2} e^{2^{\frac{D+1}{2}} b_0^D \int_{T^{(1)}}^{T^{(1)}+s} (z + |h(\tau)|) \, d\tau}, \quad s \leq T^{(2)} - T^{(1)},$$

这意味着

$$T^{(2)} - T^{(1)} \geq \inf \left\{ t \geq T^{(1)} : 2^{\frac{D+1}{2}} b_0^D \int_{T^{(1)}}^{T^{(1)}+s} (z + |h(\tau)|) \, d\tau = \ln \frac{b_0/\sqrt{D}}{\lambda^{1/2}} \right\}.$$

再利用 $a \leq \sqrt{2}|\beta|$ 和 $b \leq \sqrt{2}b_0$, 现在得到 $\theta(t)$ 的控制:

$$|\theta(T^{(1)} + s)| = |ab^D \beta| \leq 2^{\frac{D+1}{2}} b_0^D |\beta|^2 \leq 2^{\frac{D+1}{2}} b_0^D \lambda e^{2^{\frac{D+3}{2}} b_0^D \int_{T^{(1)}}^{T^{(1)}+s} (z+|h(\tau)|) d\tau}.$$

最后, 引理陈述中的上界更为粗略, 且同样适用于 $t \leq T^{(1)}$. ■

5.7.2 定理 5.2 的证明

在下文中, 将选择

$$c_b n^{-\frac{D}{2(D+2)}} \leq b_0^D \leq C_b n^{-\frac{D}{2(D+2)}},$$

其中 $c_b, C_b > 0$ 是待确定的常数, 并将被追踪.

对于某个小的 $s' > 0$, 按如下方式选择信号分量 S :

$$S = S_1 \cup S_2 = \left\{ j \geq 1 : |\theta_j^*| \geq n^{-1/2} \sqrt{\ln n} \right\} \cup \left\{ j \geq 1 : \lambda_j \geq n^{-\frac{1+s'}{D+2}} \right\}. \quad (5.77)$$

那么, 类似于 (5.33), 也有

$$|S| \leq C n^{(1-s_0)/2}, \quad \max S \leq C n^{k/2}. \quad (5.78)$$

对于噪声分量 $R = S^c$, 有

$$L := \min R \geq c n^{\frac{1+s'}{(D+2)\gamma}}. \quad (5.79)$$

类似于第 5.6.4 小节中的证明, 引入误差时间

$$T^{\text{err}} = \inf \left\{ t \geq 0 : \exists k \in R, |\theta_k(t)| > 2^{\frac{D+2}{2}} \lambda_k b_0^D e^{E \sqrt{\ln n + \ln k}} \right\}, \quad (5.80)$$

其中 E 也是待确定的常数, 并同样约定空集时 $T^{\text{err}} = +\infty$.

现在, 可以利用与第 5.6.4 小节类似的论据来展示 (5.42), 为了可读性, 在此重复:

$$|p_k| \leq \eta_k \|(\theta^* - \theta)_S\|_\infty, \quad |q_k| \leq C n^{-1/2} \sqrt{\ln(nk)}, \quad |r_k| \leq C n^{-1/2} \sqrt{\ln(nk)}, \quad (5.81)$$

其中 $\eta_k = C |S| n^{-1/2} \sqrt{\ln(nk)} \leq C n^{-s} \sqrt{\ln(nk)}$.

5.7.2.1 信号收缩

遵循第 5.6.4.1 小节, 仍然有 (5.43)。现在, 提供以下命题作为命题 5.2 的多层扩展。

命题 5.3 (收缩动力学, 多层): 假设

$$\|p_S\|_\infty \leq \eta \|\theta_S^* - \theta_S\|_\infty, \quad \|q_S + r_S\|_\infty \leq \varepsilon,$$

对于常数 $\eta \leq 1/8$ 和 $\varepsilon > 0$ 。假设对于某个 $t_0 \geq 0$ 有

$$\|\theta_S^* - \theta_S(t_0)\|_\infty \leq M.$$

那么,

$$\|\theta_S^* - \theta_S(t)\|_\infty \leq M \vee 2\varepsilon, \quad \forall t \geq t_0. \quad (5.82)$$

此外,

- 如果 $\eta M \geq \varepsilon$, 那么

$$\|\theta_S^* - \theta_S(t)\|_\infty \leq \frac{1}{2}M, \quad \forall t \geq t_0 + \bar{T}^{\text{half}}(M), \quad (5.83)$$

其中

$$\bar{T}^{\text{half}}(M) = C_{D,1} M^{-\frac{2D+2}{D+2}} + C_{D,2} M^{-1} \max_{k \in S} R_k^{-D} \ln \frac{R_k}{r_k}, \quad (5.84)$$

这里 $r_k = \min(\lambda_k^{1/2}, b_0/\sqrt{D})$ 且 $R_k = \max(\lambda_k^{1/2}, b_0/\sqrt{D})$ 。

- 如果 $\eta M < \varepsilon$, 那么对于满足 $|\theta_k^*| \geq 4\varepsilon$ 的每个 $k \in S$, 有

$$|\theta_k^* - \theta_k(t)| \leq 4\varepsilon, \quad \forall t \geq t_0 + \bar{T}_k^{\text{fin}}, \quad (5.85)$$

其中

$$\bar{T}_k^{\text{fin}} \leq C_{D,3} |\theta_k^*|^{-1} R_k^{-D} \ln \frac{R}{r} + C_{D,4} |\theta_k^*|^{-\frac{2D+2}{D+2}} \ln^+ \frac{|\theta_k^*|}{\varepsilon}. \quad (5.86)$$

这里 $C_{D,3}$ 和 $C_{D,4}$ 是仅取决于 D 的常数。另一方面, 对于所有 $k \in S$, 有

$$|\theta_k^* - \theta_k(t)| \leq 2 \max(|\theta_k^*|, 4\varepsilon), \quad \forall t \geq t_0 + C_{D,5} \varepsilon^{-\frac{2D+2}{D+2}}. \quad (5.87)$$

此处, $C_{D,1}, \dots, C_{D,5}$ 是仅取决于 D 的绝对常数。

证明 证明类似于命题 5.2, 其中在第一部分使用引理 5.10 和推论 5.5, 在第二部分使用推论 5.6 和推论 5.7。 ■

令 $v_1 = \varepsilon/\eta = Cn^{-1/2+s_1} \sqrt{\ln n}$ 。对于 $i = 0, 1, \dots, I+1$, 定义 $M_i = 2^{-i}B$, 其中 $I = \lfloor \log_2(B/v_1) \rfloor$ 。显然 $I \leq \log_2 \sqrt{n} + C$ 。由于 $\|\theta_S^* - \theta_S(0)\|_\infty = \|\theta_S^*\|_\infty \leq B$, 迭代地使用命题 5.3 得到

$$\|\theta_S^* - \theta_S(t)\|_\infty \leq M_{i+1}, \quad \forall t \geq \sum_{j=0}^i T^{(j)},$$

其中

$$T^{(i)} = CM_i^{-\frac{2D+2}{D+2}} + CM_i^{-1} \max_{k \in S} R_k^{-D} \ln \frac{R_k}{r_k} \leq CM_i^{-\frac{2D+2}{D+2}} + CM_i^{-1} b_0^{-D} \ln n,$$

这里利用 $\lambda_k \asymp k^{-\gamma}$ 和 $\max S \leq Cn^{\kappa/2}$ 得到

$$\max_{k \in S} R_k^{-D} \ln \frac{R_k}{r_k} \leq C b_0^{-D} \ln n.$$

现在, 有

$$\begin{aligned} \bar{T} &:= \sum_{i=0}^I T^{(i)} \leq C \sum_{i=0}^I M_i^{-\frac{2D+2}{D+2}} + C \sum_{i=0}^I M_i^{-1} \ln n \\ &\leq C v_1^{-1} b_0^{-D} \ln n + C v_1^{-\frac{2D+2}{D+2}} \\ &\leq C n^{\frac{D+1}{D+2}}. \end{aligned}$$

最后, 当 $t \geq \bar{T}$ 时, 有 $\eta M_{I+1} < \varepsilon$ 。选择 $v_2 = n^{-1/2} \ln n$, 利用命题 5.3 的第二部分得到, 对于所有 $k \in J_{\text{sig}}(v_2)$,

$$|\theta_k^* - \theta_k(t)| \leq 4\varepsilon, \quad \forall t \geq \bar{T} + T^{\text{fin}},$$

其中

$$\begin{aligned} T^{\text{fin}} &\leq C \max_{k \in J_{\text{sig}}(v_2)} \left[|\theta_k^*|^{-1} R_k^{-D} \ln \frac{R}{r} + |\theta_k^*|^{-\frac{2D+2}{D+2}} \ln^+ \frac{|\theta_k^*|}{\varepsilon} \right] \\ &\leq C v_2^{-1} b_0^{-D} + C v_2^{-\frac{2D+2}{D+2}} \ln n \\ &\leq C(c_b^{-1} + 1)n^{\frac{D+1}{D+2}}. \end{aligned}$$

另一方面, 对于 $k \in S \setminus J_{\text{sig}}(v_2)$, 只要 $t \geq \bar{T} + C\varepsilon^{-\frac{2D+2}{D+2}}$, 便利用 (5.87) 得到

$$|\theta_k^* - \theta_k(t)| \leq 2 \max(|\theta_k^*|, 4\varepsilon).$$

类似于两层情况的论据,

$$\begin{aligned} \sum_{k \in S \setminus J_{\text{sig}}(v_2)} |\theta_k^* - \theta_k(t)|^2 &= \sum_{k \in S_1 \setminus J_{\text{sig}}(v_2)} |\theta_k^* - \theta_k(t)|^2 + \sum_{k \in S_2 \setminus J_{\text{sig}}(v_2)} |\theta_k^* - \theta_k(t)|^2 \\ &\leq C \sum_{k \in S_1 \setminus J_{\text{sig}}(v_2)} |\theta_k^*|^2 + C \sum_{k \in S_2 \setminus (S_1 \cup J_{\text{sig}}(v_2))} \varepsilon^2 \\ &\leq C\Psi(v_2) + Cn^{\frac{1+s'}{(D+2)\gamma}} \varepsilon^2. \end{aligned}$$

因此, 得出结论, 在 $t \geq \underline{C}n^{\frac{D+1}{D+2}} = C(c_b^{-1} + 1)n^{\frac{D+1}{D+2}}$ 之后, 有

$$\|\theta_S^* - \theta_S(t)\|_2^2 \leq C \frac{\ln n}{n} \left[\Phi\left(n^{-1/2} \sqrt{\ln n}\right) + n^{\frac{1+s'}{(D+2)\gamma}} \right] + C\Psi(n^{-1/2} \ln n).$$

5.7.2.2 噪声项的界

本文将利用引理 5.15 的第二部分来证明：如果 E 选择得当，(5.80) 中定义的误差时间 T^{err} 满足 $T^{\text{err}} \geq \bar{C} n^{\frac{D+1}{D+2}}$ （对于某个常数 $\bar{C}$ ）。

为了应用该引理，首先必须证明对于所有 $k \in R$ 都有 (5.88)，即

$$2^{\frac{D+1}{2}} b_0^D \int_0^t (|\theta_k^*| + |h_k(s)|) \, ds \leq \ln \frac{b_0/\sqrt{D}}{\lambda_k^{1/2}}, \quad (5.88)$$

其中 $h_k(s) = p_k(s) + q_k(s) + r_k(s)$ 。

令 $T^* = \bar{C} n^{\frac{D+1}{D+2}}$ ，其中 $\bar{C}$ 为常数。 b_0 的选择给出

$$b_0^D T^{*n^{-1/2}} \leq \bar{C} C_b. \quad (5.89)$$

首先，有

$$b_0^D \int_0^{T^*} |\theta_k^*| \, ds = b_0^D T^* |\theta_k^*| \leq b_0^D T^{*n^{-1/2}} \sqrt{\ln n} \leq \bar{C} C_b \sqrt{\ln n}.$$

记 $\bar{T}^{(i)} = \sum_{j=0}^i \bar{T}^{(j)}$ 。现在，对于 $t \in [\bar{T}^{(i)}, \bar{T}^{(i+1)}]$ ，利用前一部分的结果，有 $\|\theta_S^* - \theta_S(t)\|_\infty \leq M_i$ ，故 (5.81) 中的第一个不等式给出

$$|p_k(t)| \leq C M_i n^{-s} \sqrt{\ln k}, \quad t \in [\bar{T}^{(i)}, \bar{T}^{(i+1)}].$$

从而

$$\begin{aligned} \int_0^{\bar{T}} |p_k(t)| \, dt &\leq C \sum_{i=0}^I M_i n^{-s} \sqrt{\ln k} \cdot T^{(i)} \\ &\leq C \sum_{i=0}^I M_i n^{-s} \sqrt{\ln k} \cdot \left(M_i^{-\frac{2D+2}{D+2}} + M_i^{-1} b_0^{-D} \ln n \right) \\ &\lesssim \left(\sum_{i=0}^I M_i^{-\frac{D}{D+2}} n^{-s} \sqrt{\ln k} + \sum_{i=0}^I b_0^{-D} n^{-s} \sqrt{\ln k} \right) \\ &\lesssim v_1^{-\frac{D}{D+2}} + b_0^{-D} n^{-s} \sqrt{\ln k} \log v_1^{-1} \\ &\lesssim v_1^{-\frac{D}{D+2}} + b_0^{-D} \sqrt{\ln k}, \end{aligned}$$

且

$$b_0^D \int_0^{\bar{T}} |p_k(t)| \, dt \lesssim C_b \sqrt{\ln k}.$$

此外，对于 $t \geq \bar{T}$ ，有

$$|p_k(t)| \leq \eta_k \|(\theta^* - \theta)_S\|_\infty \leq C n^{-s} \sqrt{\ln(nk)} \cdot M_{I+1} \leq C n^{-1/2} \sqrt{\ln k},$$

因此结合 (5.89),

$$b_0^D \int_{\bar{T}}^{T^*} |p_k(t)| \, dt \leq C \bar{C} C_b \sqrt{\ln k}.$$

此外, (5.81) 中的另外两个不等式给出

$$|q_k + r_k| \leq C n^{-1/2} \sqrt{\ln n + \ln k},$$

故

$$b_0^D \int_0^{T^*} |q_k(t) + r_k(t)| \, dt \leq \bar{C} C_b \sqrt{\ln n + \ln k}.$$

结合所有这些项, 得到

$$2^{\frac{D+1}{2}} b_0^D \int_0^{T^*} (|\theta_k^*| + |h_k(t)|) \, dt \leq C \bar{C} C_b (\sqrt{\ln n} + \sqrt{\ln k}). \quad (5.90)$$

现在, 根据对 $R = S^c$ 的选择, 有 $\ln \lambda_k^{-1} \geq \frac{1+s'}{D+2} \ln n$ (同时 $k \geq L \asymp n^{\frac{1+s'}{(D+2)\gamma}}$), 从而

$$\begin{aligned} \ln \frac{b_0/\sqrt{D}}{\lambda_k^{1/2}} &\geq -\frac{1}{2(D+2)} \ln n + \frac{1}{2} \ln \lambda_k^{-1} - C \\ &\geq -\frac{1}{2(D+2)} \ln n + \frac{1}{2} \frac{1+s'/2}{1+s'} \ln \lambda_k^{-1} + \frac{1}{2} \frac{s'/2}{1+s'} \ln \lambda_k^{-1} - C \\ &\geq \frac{s'}{4(D+2)} \ln n + \frac{s'\gamma}{4(1+s')} \ln k - C. \end{aligned}$$

因此, 只要 n 足够大, 对于所有 $k \in R$ 都有 (5.88)。现在, 将 (5.90) 代入引理 5.15 中的上界 (5.75), 即可证明该断言。

5.7.2.3 泛化误差

现在, 可以固定 c_b, C_b, E 并将它们视为常数。现在, 当 $t \in [\underline{C} n^{\frac{D+1}{D+2}}, \bar{C} n^{\frac{D+1}{D+2}}]$ 时, 误差项由下式限制:

$$\begin{aligned} \|\theta_R(t)\|_2^2 &\leq C \sum_{k \in R} \lambda_k^2 b_0^{2D} e^{C \bar{C} (\sqrt{\ln n} + \sqrt{\ln k})} \\ &\leq C n^{-\frac{D}{D+2}} e^{C \bar{C} \sqrt{\ln n}} \sum_{k \in R} \lambda_k^2 e^{2C \sqrt{\ln k}} \\ &\leq C n^{-\frac{D}{D+2}} e^{C \bar{C} \sqrt{\ln n}} \sum_{k \geq L} \lambda_k^2 e^{2C \sqrt{\ln k}} \\ &\leq C n^{-\frac{D}{D+2}} e^{C \bar{C} \sqrt{\ln n}} L^{1-2\gamma+s} \\ &\leq C n^{-1+\frac{1+s}{(D+2)\gamma}}. \end{aligned}$$

因此,

$$\|\theta_R^* - \theta_R(t)\|_2^2 \leq 2\|\theta_R^*\|_2^2 + 2\|\theta_R(t)\|_2^2 \leq \Psi(n^{-1/2}\sqrt{\ln n}) + Cn^{-1+\frac{1+s}{(D+2)\gamma}}.$$

最后, 通过结合信号分量和噪声分量即可得到结果。

5.8 特征改善的证明

在下文中, 固定特征基 $\{e_j\}_{j \geq 1}$ 以及真实函数 f^* , 从而固定系数 $\{\theta_j^*\}_{j \geq 1}$ 。此时, 特征误差度量仅与权重 $\lambda = (\lambda_j)_{j \geq 1}$ 相关。为了简化符号, 记

$$\mathcal{E}(\delta, \epsilon^2; \lambda) = \mathcal{E}(\delta, \epsilon^2; \Phi, f^*) = \#\{j \in N : \lambda_j \geq \delta\} \epsilon^2 + \sum_{j \in N} (\theta_j^*)^2 \mathbf{1}\{\lambda_j < \delta\},$$

以及

$$\mathcal{E}^*(\epsilon^2; \lambda) = \inf_{\delta \geq 0} \mathcal{E}(\delta, \epsilon^2; \lambda), \quad \delta^*(\epsilon^2; \lambda) \in \arg \min_{\delta \geq 0} \mathcal{E}(\delta, \epsilon^2; \lambda).$$

从特征误差度量的表达式可以明显看出, $\mathcal{E}^*(\epsilon^2; \lambda)$ 仅取决于由权重 λ 诱导的索引顺序。此外, 可以选择 $\delta^*(\epsilon^2; \lambda) = \lambda_j$ 对于某个 $j \geq 1$ 。进一步地, 有局部条件:

$$\sum_{\lambda_j = \delta^*(\epsilon^2; \lambda)} (\theta_j^*)^2 \geq \#\{j \in N : \lambda_j = \delta^*(\epsilon^2; \lambda)\} \epsilon^2, \quad (5.91)$$

否则可以通过增大 δ 来获得更小的误差。

5.8.1 特征误差度量的基本性质

将信号和噪声的索引集定义为

$$I_s(\epsilon^2) = \{j : (\theta_j^*)^2 \geq \epsilon^2\}, \quad I_n(\epsilon^2) = \{j : (\theta_j^*)^2 < \epsilon^2\}. \quad (5.92)$$

以下命题刻画了特征误差度量保持非增的充分条件。

命题 5.4: 令 λ 为权重序列, λ' 为仅将 λ_j 修改为 λ'_j 后的结果。假设 $\lambda_r = \delta^*(\epsilon^2; \lambda)$ 。那么, 仅在以下情况下才可能出现 $\mathcal{E}^*(\epsilon^2; \lambda') > \mathcal{E}^*(\epsilon^2; \lambda)$:

- (a) $\lambda_j < \lambda_r$, $\lambda'_j \geq \lambda_r$ 且 $(\theta_j^*)^2 < \epsilon^2$;
- (b) $\lambda_j \geq \lambda_r$, $\lambda'_j < \lambda_r$ 且 $(\theta_j^*)^2 > \epsilon^2$, 并且存在某个 λ_l 使得 $(\theta_l^*)^2 < \epsilon^2$ 且 $\lambda_r > \lambda_l \geq \lambda'_j$ 。

证明 列举以下情况并逐一考虑:

- (1) $\lambda_j \geq \lambda_r$ 且 $\lambda'_j \geq \lambda_r$, 或者 $\lambda_j < \lambda_r$ 且 $\lambda'_j < \lambda_r$;
- (2) $\lambda_j < \lambda_r$, $\lambda'_j \geq \lambda_r$;
- (3) $\lambda_j \geq \lambda_r$, $\lambda'_j < \lambda_r$ 。

对于情况 (1), 始终有

$$\mathcal{E}^*(\epsilon^2; \lambda') \leq \mathcal{E}(\lambda_r, \epsilon^2; \lambda') = \mathcal{E}(\lambda_r, \epsilon^2; \lambda) = \mathcal{E}^*(\epsilon^2; \lambda).$$

对于情况 (2), 如果 $(\theta_j^*)^2 \geq \epsilon^2$, 发现

$$\mathcal{E}^*(\epsilon^2; \lambda) - \mathcal{E}(\lambda_r, \epsilon^2; \lambda') = \mathcal{E}(\lambda_r, \epsilon^2; \lambda) - \mathcal{E}(\lambda_r, \epsilon^2; \lambda') = (\theta_j^*)^2 - \epsilon^2 \geq 0,$$

所以只有当 $(\theta_j^*)^2 < \epsilon^2$ 时, 才可能出现 $\mathcal{E}^*(\epsilon^2; \lambda') > \mathcal{E}^*(\epsilon^2; \lambda)$, 即情况 (a)。

对于情况 (3), 如果 $(\theta_j^*)^2 \leq \epsilon^2$, 类似于前一种情况, 有

$$\mathcal{E}^*(\epsilon^2; \lambda) - \mathcal{E}(\lambda_r, \epsilon^2; \lambda') = \epsilon^2 - (\theta_j^*)^2 \geq 0.$$

现在, 如果 $(\theta_j^*)^2 > \epsilon^2$, 但不存在 (b) 中指定的 λ_l , 有

$$\mathcal{E}^*(\epsilon^2; \lambda) - \mathcal{E}(\lambda'_j, \epsilon^2; \lambda') = \mathcal{E}(\lambda_r, \epsilon^2; \lambda) - \mathcal{E}(\lambda'_j, \epsilon^2; \lambda') = \sum_{l: \lambda_r > \lambda_l \geq \lambda'_j} [(\theta_l^*)^2 - \epsilon^2] \geq 0.$$

■

推论 5.8: 在与命题 5.4 相同的设置下, $\mathcal{E}^*(\epsilon^2; \lambda') > \mathcal{E}^*(\epsilon^2; \lambda)$ 仅在发生“上穿 (up-crossing)”时才可能。即, 存在某些索引 j, k 使得: (1) $\lambda_j \geq \lambda_r > \lambda_k$; (2) $(\theta_j^*)^2 \geq \epsilon^2$ 且 $(\theta_k^*)^2 < \epsilon^2$; (3) $\lambda'_j \leq \lambda'_k$ 。

证明 对于情况 (b), 该条件对于对 (j, l) 已经成立。对于情况 (a), 利用局部条件 (5.91), 可以找到某个满足 $\lambda_k = \lambda_r$ 且 $(\theta_k^*)^2 \geq \epsilon^2$ 的 k 。那么, 对 (k, j) 满足上述条件。

■

从命题 5.4 中, 发现如果 λ 改变后满足以下条件, 则 $\mathcal{E}^*(\epsilon^2; \lambda)$ 是非增的:

引理 5.16: 令 $\lambda(t), t \in [0, T]$ 为权重的连续流。令 $N = N_1 \sqcup N_2$ 为索引集 N 的一个划分。进一步假设:

- (1) 对于每个 $j \in N_1 \cap \mathcal{I}_s(\epsilon^2)$ 和 $k \in N_1 \cap \mathcal{I}_n(\epsilon^2)$, 如果存在某个 t_0 使得 $\lambda_j(t_0) \geq \lambda_k(t_0)$, 那么对于所有 $t \geq t_0$ 都有 $\lambda_j(t) \geq \lambda_k(t)$ 。
- (2) 对于每个 $j \in N_2$, 有 $\lambda_j(t) < \delta^*(\epsilon^2; \lambda(t))$ 。

那么, $\mathcal{E}^*(\epsilon^2; \lambda(t))$ 关于 t 是非增的。

证明 利用权重的连续性以及 $\mathcal{E}^*(\epsilon^2; \lambda)$ 仅取决于由权重诱导的索引顺序, 可以将 $\lambda(t)$ 的连续动力学简化为一次仅改变一个权重的离散步骤 (如果 N 是无限的, 可以取一个有限但足够大的子集)。那么, 结果遵循推论 5.8: 对于 $j \in N_2 \cap \mathcal{I}_s(\epsilon^2)$, 由于始终有 $\lambda_j(t) < \delta^*(\epsilon^2; \lambda(t))$, 上穿不可能发生; 对于 $j \in N_1 \cap \mathcal{I}_s(\epsilon^2)$, 条件 (1) 确保了对于 $k \in N_1 \cap \mathcal{I}_n(\epsilon^2)$ 不会发生上穿, 而条件 (2) 确保了对于 $k \in N_2 \cap \mathcal{I}_n(\epsilon^2)$ 也不会发生上穿。

■

5.8.2 一维动力学补充结果

承接第 5.6.5 小节, 给出对角自适应特征模型的一维动力学的一些额外结果。考虑无扰动的一维梯度流方程:

$$\begin{cases} \dot{\theta}(t) = \beta(t)(z(t) - w(t)), & \theta(0) = \lambda^{\frac{1}{2}} > 0, \\ \dot{\beta}(t) = \theta(t)(z(t) - w(t)), & \beta(0) = 0, \end{cases} \quad (5.93)$$

其中 $z(t)$ 是一个连续函数, $\lambda > 0$ 是常数, 且 $w(t) = \theta(t)\beta(t)$ 。那么, 也可以计算 $w(t)$ 的动力学:

$$\dot{w}(t) = (\theta^2(t) + \beta^2(t))(z(t) - w(t)), \quad w(0) = 0.$$

在下文中, 记 $\tilde{\lambda}(t) = \theta(t)^2$, 且 $\tilde{\lambda}(0) = \lambda$ 。

与第 5.6.5 小节中的分析相同, 可以计算出:

$$\frac{1}{2} \frac{d}{dt} \theta^2 = \frac{1}{2} \frac{d}{dt} \beta^2 = \theta\beta(z - w),$$

因此有:

$$\theta^2(t) - \beta^2(t) = \theta(0)^2 - \beta(0)^2 = \lambda. \quad (5.94)$$

这也表明 $\theta(t) \geq \lambda^{\frac{1}{2}}$ 。

此外, 如果 $z(t)$ 不改变符号, β 将与 z 具有相同的符号。而且, 如果 $z(t) \equiv z$ 是一个常数, 可知 $\theta(t)$ 和 $|\beta(t)|$ 是单调增加的。

引理 5.17: 考虑 (5.93) 的两个实例, 分别具有 $\theta, \beta, z, \lambda$ 和 $\theta', \beta', z', \lambda'$ 。假设 $\min_t |z(t)| \geq \max_t |z'(t)|$ 。那么, 如果对于某个 $t_0 \geq 0$ 有 $\theta(t_0) \geq \theta'(t_0)$, 则对于所有 $t \geq t_0$ 都有 $\theta(t) \geq \theta'(t)$ 。

证明 不失一般性, 可以假设 $\min_t z(t) \geq \max_t |z'(t)| \geq 0$ 。首先, 对于 $\lambda \geq \lambda'$ 的情况, 利用比较原理容易看出对于所有 $t \geq 0$ 都有 $\theta(t) \geq \theta'(t)$ 且 $\beta(t) \geq |\beta'(t)|$ 。现在, 如果 $\lambda < \lambda'$, 利用 (5.94), 有:

$$\beta(t_0)^2 = \theta^2(t_0) - \lambda \geq (\theta'(t_0))^2 - \lambda' = (\beta'(t_0))^2,$$

所以再次利用比较原理, 对于所有 $t \geq t_0$ 都有 $\beta(t) \geq |\beta'(t)|$ 且 $\theta(t) \geq \theta'(t)$ 。 ■

引理 5.18: 记 $M = \max_t |z|$ 。有:

$$\theta(t) \leq \sqrt{2\lambda}^{1/2}, \quad |w(t)| \leq \sqrt{2\lambda} \quad \forall t \leq \frac{1}{\sqrt{2}M}$$

以及

$$\theta(t) \leq \lambda^{1/2} \left[1 + e^{\sqrt{2}tM} \right].$$

证明 可以使用与引理 5.8 相同的证明方法。对于 θ 的上界, 使用 $\theta(t) \leq \sqrt{\lambda + \beta^2(t)}$ 。

引理 5.19: 假设 $m = \min_t |z| > 0$ 。有:

$$\theta(t)^2 \geq \frac{1}{2}m, \quad \text{对于 } t \geq m^{-1} \left(2 + \log^+ \frac{m}{2\lambda} \right).$$

证明 为了简化符号, 去掉下标 j 。定义:

$$T^{\text{esc}} = \inf \{ t \geq 0 : |\beta(t)| \geq \lambda^{1/2} \}, \quad T^{\text{sig}} = \inf \{ t \geq 0 : |w(t)| \geq m/2 \}.$$

注意到如果 $|w(t)| = |\theta(t)\beta(t)| \geq m/2$, 那么:

$$\theta(t)^2 \geq |\theta(t)\beta(t)| = |w(t)| \geq m/2.$$

因此, 只需考虑 $T^{\text{sig}} > 0$ 的情况并给出 T^{sig} 的上界。不失一般性, 假设 $z(t) > 0$ 。当 $t \leq T^{\text{esc}} \wedge T^{\text{sig}}$ 时, 有:

$$\dot{\beta}(t) \geq \frac{1}{2} \lambda^{\frac{1}{2}} m, \quad t \leq T^{\text{esc}} \wedge T^{\text{sig}}.$$

故 $T^{\text{esc}} \wedge T^{\text{sig}} \leq 2/m$ 。如果 $T^{\text{sig}} = T^{\text{esc}} \wedge T^{\text{sig}} \leq 2/m$, 已经证明了结果。对于另一种情况, 有:

$$\dot{w} = (\theta^2(t) + \beta^2(t))(z(t) - w(t)) \geq 2(\theta(t)\beta(t)) \cdot \frac{1}{2}m = w(t)m, \quad t \in [T^{\text{esc}}, T^{\text{sig}}].$$

结合 $w(T^{\text{esc}}) \geq \lambda$, 得出:

$$T^{\text{sig}} - T^{\text{esc}} \leq \frac{1}{m} \log \frac{m}{2\lambda}.$$

5.8.3 序列损失下的单层参数化

考虑序列模型下的两层对角自适应核方法, 其中 $z_j = \theta_j^* + \varepsilon_j$, 且 $\varepsilon_j \stackrel{i.i.d.}{\sim} N(0, 1/n)$ 。梯度流方程的显式形式给出如下:

$$\begin{cases} \dot{\theta}_j = -\nabla_{\theta_j} \mathcal{L}_n = \beta_j(t) [z_j - f_j(t)], & \theta_j(0) = \lambda_j^{\frac{1}{2}}, \\ \dot{\beta}_j = -\nabla_{\beta_j} \mathcal{L}_n = \theta_j(t) [z_j - f_j(t)], & \beta_j(0) = 0, \end{cases}$$

其中 $f_j(t) = \theta_j(t)\beta_j(t)$ 。该方程与 $z(t) \equiv z = \theta_j^* + \varepsilon_j$ 且 $w(t) = f_j(t)$ 的一维梯度流方程 (5.93) 一致。

5.8.3.1 定理 5.3 的证明

初始时刻, 由于 $\varepsilon_j \sim N(0, 1/n)$, 以至少 $1 - Cn^{-2}$ 的概率, 有:

$$|\varepsilon_j| \lesssim \sqrt{\frac{\log(jn)}{n}}, \quad \forall j \geq 1.$$

单调性。首先证明：如果 $\mathcal{E}(n^{-1}; \lambda(t)) \leq \mathcal{E}(n^{-1}; \lambda(0))$ ，那么：

$$\delta^*(n^{-1}; \lambda(t)) \geq \delta_0 = Cn^{-\frac{q\gamma}{p+q}}. \quad (5.95)$$

首先计算初始误差度量：

$$\mathcal{E}(\delta, n^{-1}; \lambda) = \#\{j \in N : \lambda_j \geq \delta\} n^{-1} + \sum_{j \in N} (\theta_j^*)^2 \mathbf{1}\{\lambda_j < \delta\}.$$

对于第一项，利用 $\lambda_j \asymp j^{-\gamma}$ ，有：

$$\#\{j \in N : \lambda_j \geq \delta\} n^{-1} \asymp \delta^{-1/\gamma} n^{-1}.$$

对于第二项，利用 $f_{j(\ell)}^* \asymp \ell^{-(p+1)/2}$ ， $j(\ell) \asymp \ell^q$ 发现：

$$\sum_{j \in N} (\theta_j^*)^2 \mathbf{1}\{\lambda_j < \delta\} \asymp \delta^{p/(q\gamma)}.$$

平衡这两项，发现：

$$\delta^*(n^{-1}; \lambda) \asymp n^{-\frac{q\gamma}{p+q}}, \quad \mathcal{E}^*(n^{-1}; \lambda) \asymp n^{-\frac{p}{p+q}}.$$

由于 $\lambda_j(t)$ 是单调增加的，有：

$$\mathcal{E}(\delta, n^{-1}; \lambda(t)) \geq \#\{j \in N : \theta_j(t)^2 \geq \delta\} n^{-1} \geq \#\{j \in N : \lambda_j \geq \delta\} n^{-1} \geq C\delta^{-1/\gamma} n^{-1}.$$

因此，如果 $\mathcal{E}(n^{-1}; \lambda(t)) \leq \mathcal{E}(n^{-1}; \lambda(0))$ ，有 $\mathcal{E}^*(\delta, n^{-1}; \lambda(t)) \geq cn^{-\frac{p}{p+q}}$ ，从而：

$$\delta^*(n^{-1}; \lambda(t)) \geq Cn^{-\frac{q\gamma}{p+q}}.$$

现在，对于满足 (5.95) 的时间区间 $[0, T]$ ，应用引理 5.16 并验证条件 (1) 和 (2)。取 $L \asymp (n/\log n)^{\frac{q}{p+1}}$ ，并设 $N_1 = \{j < L\}$ 和 $N_2 = \{j \geq L\}$ 。对于每个 $j \geq L$ ，有：

$$|z_j| \leq |\theta_j^*| + |\varepsilon_j| \lesssim L^{-\frac{p+1}{2q}} + \sqrt{\frac{\log(jn)}{n}} \lesssim \sqrt{\frac{\log(jn)}{n}}.$$

利用引理 5.18，对于 $t \leq c\sqrt{n/(\log n)}$ ，有：

$$\lambda_j(t) = \theta_j(t)^2 \leq \lambda_j \left[1 + e^{\sqrt{2t}|z|}\right]^2 \leq C\lambda_j e^{c\sqrt{\log j}} < \delta_0 = Cn^{-\frac{q\gamma}{p+q}}, \quad j \geq L,$$

因为当 $q > 1$ 时， $L \asymp (n/\log n)^{\frac{q}{p+1}} \gg n^{\frac{q}{p+q}}$ 。因此，验证了引理 5.16 中的条件 (2)。另一方面，由于 $L \lesssim n^{\frac{q}{p+1}}$ ，有：

$$|\varepsilon_j| \lesssim \sqrt{\frac{C \log(jn)}{n}} \leq C\sqrt{\frac{\log n}{n}}, \quad j \leq L.$$

现在，对于满足 $j \in \mathcal{I}_s(n^{-1})$ 的 $j < L$ ，有：

$$|\theta_j^*| \gtrsim L^{-\frac{p+1}{2q}} \gtrsim \sqrt{\frac{\log n}{n}}.$$

因此, 将 L 的常数因子取足够小, 有:

$$|z_j| \geq |\theta_j^*| - |\varepsilon_j| \geq \frac{1}{2} |\theta_j^*| > C \sqrt{\frac{\log n}{n}}.$$

同时, 对于 $k < L$ 且 $k \in \mathcal{I}_n(n^{-1})$, 有 $|z_k| = |\varepsilon_k| \leq C \sqrt{\frac{\log n}{n}}$. 因此, 利用引理 5.17 就证明了引理 5.16 中的条件 (1)。

最后, 通过连续性论据证明时间区间 $[0, T]$ 实际上可以覆盖 $T = c\sqrt{n/(\log n)}$ 。假设 $\mathcal{E}^*(n^{-1}; \lambda(t))$ 在 t_0 时刻发生跳变, 那么由于 $\lambda(t)$ 的连续性以及 $\lambda_j(t), \lambda_{j'}(t)$ 以概率 1 不会重合, 它最多增加 ϵ^2 。因此, 仍然有 $\mathcal{E}^*(\delta, n^{-1}; \lambda(t)) \lesssim n^{-\frac{p}{p+q}}$ 且 $\delta^*(n^{-1}; \lambda(t)) \geq Cn^{-\frac{q\gamma}{p+q}}$, 虽然常数因子可能增大。尽管如此, 该论据的第二部分对于修改后的常数仍然成立, 因此在 t_0 时刻不会发生上穿, 所以 $\mathcal{E}^*(n^{-1}; \lambda(t))$ 在 t_0 时刻不能增加。

特征误差度量。 在前一部分已经证明了 $\mathcal{E}^*(n^{-1}; \lambda) \asymp n^{-\frac{p}{p+q}}$ 。对于 $t = t_* \asymp \sqrt{n/(\log n)}$, 取 $\delta_* = Cn^{-\frac{1}{2}}$ 。首先考虑 $\mathcal{E}_V(\delta_*, n^{-1}; \lambda)$ 。如前所述, 对于 $j \geq L$ 有 $\lambda_j(t_*) < \delta_*$ 。对于满足 $|\theta_j^*| \lesssim n^{-\frac{1}{2}}(\log n)^{\frac{1}{2}}$ 的 $j < L$, 应用引理 5.18 可发现 $\lambda_j(t) = \theta_j(t)^2 \leq 2\lambda_j$ 。因此:

$$\begin{aligned} & \# \{j \in N : \theta_j(t_*)^2 \geq \delta_*\} n^{-1} \\ &= \# \{j < L : \theta_j(t_*)^2 \geq \delta_*\} n^{-1} \\ &\leq n^{-1} \left[\# \{j < L : \lambda_j \geq \delta_*/2\} + \# \{j < L : |\theta_j^*| \gtrsim n^{-\frac{1}{2}}(\log n)^{\frac{1}{2}}\} \right] \\ &\lesssim \delta_*^{-1/\gamma} n^{-1} + (n \log n)^{\frac{1}{p+1}} n^{-1} \\ &\lesssim n^{-(1-1/(2\gamma))} + n^{-\frac{p}{p+1}} (\log n)^{-\frac{1}{p+1}}, \end{aligned}$$

现在考虑 $\mathcal{E}_B(\delta_*, n^{-1}; \lambda)$ 。对于满足 $|\theta_j^*| \gtrsim n^{-\frac{1}{2}}(\log n)^{3/2}$ 的 j (这蕴含 $j < L$), 应用引理 5.19 发现:

$$\theta_j(t)^2 \geq \frac{1}{2} |z_j| \geq \frac{1}{4} |\theta_j^*| \gtrsim \sqrt{\frac{\log n}{n}}.$$

因此:

$$\sum_{j \in N} (\theta_j^*)^2 \mathbf{1} \{ \lambda_j < \delta_* \} \leq \sum_{j \in N} (\theta_j^*)^2 \mathbf{1} \left\{ |\theta_j^*| \gtrsim n^{\frac{1}{2}} (\log n)^{3/2} \right\} \lesssim n^{-\frac{p}{p+1}} (\log n)^{\frac{2p}{p+1}}.$$

因此有:

$$\mathcal{E}^*(n^{-1}; \lambda(t_*)) \leq \mathcal{E}(\delta_*, n^{-1}; \lambda(t_*)) \lesssim n^{-(1-1/(2\gamma))} + n^{-\frac{p}{p+1}} (\log n)^{\frac{2p}{p+1}}.$$

5.8.4 序列损失下的更深层过参数化

在本小节中, 考虑序列模型下的更深层过参数化。回顾梯度流动力学给出如下:

$$\begin{cases} \dot{\beta}_j(t) = -\nabla_{\beta_j} \mathcal{L}_n = \theta_j b_j^D (z_j - f_j), & \beta_j(0) = 0; \\ \dot{\theta}_j(t) = -\nabla_{\theta_j} \mathcal{L}_n = b_j^D \beta_j (z_j - f_j), & \theta_j(0) = \lambda_j^{\frac{1}{2}}; \\ \dot{b}_j(t) = -\nabla_{b_j} \mathcal{L}_n = D \theta_j b_j^{D-1} \beta_j (z_j - f_j), & b_j(0) = b_0, \end{cases} \quad (5.96)$$

其中 $f_j = \theta_j b_j^D \beta_j$ 。在此情形下, 记:

$$\tilde{\lambda} = (\tilde{\lambda}_j)_{j \geq 1}, \quad \tilde{\lambda}_j(t) = \left(\theta_j(t) b_j^D(t) \right)^2, \quad \tilde{\lambda}_j(0) = \lambda_j b_0^{2D}.$$

这种梯度流动力学已在前文研究, 现在引入一些额外的引理。

命题 5.5: 考虑 (5.96) 并假设 $b_0/\sqrt{D} \leq \lambda_j^{\frac{1}{2}} \leq 1$, 有:

$$\tilde{\lambda}_j^{\frac{1}{2}} = \theta_j(t) b_j^D(t) \leq C_D \max(\lambda_j^{\frac{1}{2}} b_0^D, |z|^{\frac{D+1}{D+2}}, b_0^{-1} |z|).$$

证明 为了简化符号, 省略下标 j 。遵循守恒量 (5.66), 得到

$$\min(\lambda^{\frac{1}{2}}, |\beta|) \leq \theta \leq \sqrt{2} \max(\lambda^{\frac{1}{2}}, |\beta|), \quad \min(b_0, \sqrt{D} |\beta|) \leq b \leq \sqrt{2} \max(b_0, \sqrt{D} |\beta|).$$

如果 $\beta \leq \min(\lambda^{\frac{1}{2}}, b_0/\sqrt{D})$, 立即有:

$$\tilde{\lambda} = \theta b^D \leq 2^{\frac{D+1}{2}} \lambda^{\frac{1}{2}} b_0^D.$$

如果 $\beta \geq \max(\lambda^{\frac{1}{2}}, b_0/\sqrt{D})$, 有:

$$|z| \geq |f| \geq D^{\frac{D}{2}} |\beta|^{D+2} \implies |\beta| \leq D^{-\frac{D}{2(D+2)}} |z|^{\frac{1}{D+2}},$$

因此:

$$\tilde{\lambda} = \theta b^D \leq 2^{\frac{D+1}{2}} D^{\frac{D}{2}} |\beta|^{D+1} \leq 2^{\frac{D+1}{2}} D^{\frac{D}{2(D+2)}} |z|^{\frac{D+1}{D+2}}.$$

否则, 如果 $b_0/\sqrt{D} \leq \beta \leq \lambda^{\frac{1}{2}}$, 利用:

$$|z| \geq |f| \geq \lambda^{\frac{1}{2}} \cdot (\sqrt{D} |\beta|)^D \cdot |\beta| \implies |\beta| \leq (\lambda^{\frac{1}{2}} D^{\frac{D}{2}})^{-\frac{1}{D+1}} |z|^{\frac{1}{D+1}},$$

故:

$$\tilde{\lambda} \leq 2^{\frac{D+1}{2}} \lambda^{\frac{1}{2}} D^{\frac{D}{2}} |\beta|^D \leq 2^{\frac{D+1}{2}} (\lambda^{\frac{1}{2}} D^{\frac{D}{2}})^{\frac{1}{D+1}} |z|^{\frac{D}{D+1}} \leq 2^{\frac{D+1}{2}} D^{\frac{D}{2(D+1)}} \lambda^{\frac{1}{2(D+1)}} |z|^{\frac{D}{D+1}}.$$

此外, 还有:

$$|z| \geq \lambda^{\frac{1}{2}} \cdot (\sqrt{D} |\beta|)^D \cdot |\beta| \geq \lambda^{\frac{1}{2}} b_0^{D+1} D^{-\frac{1}{2}} \implies \lambda^{\frac{1}{2}} \leq D^{\frac{1}{2}} b_0^{-(D+1)} |z|.$$

将此代入前一不等式, 发现:

$$\tilde{\lambda} \leq 2^{\frac{D+1}{2}} D^{\frac{1}{2}} b_0^{-1} |z|.$$

■

引理 5.20 (比较原理): 考虑 (5.96) 并假设 $|z_j| \geq |z_k|$ 。那么, 如果对于某个 $t_0 \geq 0$ 有 $\theta_j(t_0) \geq \theta_k(t_0)$, 则对于所有 $t \geq t_0$ 都有 $\theta_j(t) \geq \theta_k(t)$ 。

证明 类似于引理 5.17 的证明, 并注意到所有 j 的初始化 $b_j(0) = b_0$ 是相同的。 ■

引理 5.21 (噪声情形): 对于梯度流 (5.96), 假设 $\lambda_j^{\frac{1}{2}} \leq b_0/\sqrt{D}$ 。记 $T_j^{(1)} = \left(2^{\frac{D+1}{2}} b_0^D |z_j|\right)^{-1}$, 有:

$$\theta_j(t) b_j^D(t) \leq 2^{\frac{D+1}{2}} \lambda_j^{\frac{1}{2}} b_0^D, \quad \text{对于 } t \leq T_j^{(1)}. \quad (5.97)$$

以及

$$\begin{aligned} \theta_j(t) b_j^D(t) &\leq 2^{\frac{D+1}{2}} \lambda_j^{\frac{1}{2}} b_0^D e^{2^{\frac{D+1}{2}} b_0^D |z_j| (t - T_j^{(1)})^+}, \\ \text{对于 } t &\leq \left(1 + \log \frac{b_0}{\sqrt{D} \lambda_j^{\frac{1}{2}}}\right) T_j^{(1)}. \end{aligned} \quad (5.98)$$

证明 这是引理 5.15 的直接推论。 ■

引理 5.22 (信号情形): 对于梯度流 (5.96), 记:

$$T_j^{\text{eig}} = \inf \left\{ t \geq 0 : \theta_j(t) b_j^D(t) \geq |z_j|^{\frac{D+1}{D+2}} \right\}.$$

有:

- 如果 $\lambda_j^{\frac{1}{2}} \leq b_0/\sqrt{D}$, 那么:

$$T_j^{\text{eig}} \leq 2(b_0^D |z_j|)^{-1} \left[1 + \log^+ \frac{(D^{-\frac{D}{2}} |z_j|/2)^{\frac{1}{D+2}}}{\lambda_j^{\frac{1}{2}}} \right], \quad (5.99)$$

- 如果 $\lambda_j^{\frac{1}{2}} \geq b_0/\sqrt{D}$, 那么:

$$T_j^{\text{eig}} \leq 2 \left(\sqrt{D} \lambda_j^{\frac{1}{2}} b_0^{D-1} |z_j| \right)^{-1} (1 + R_j), \quad (5.100)$$

其中

$$R_j = \begin{cases} \log^+ \frac{(D|z_j|/2)^{\frac{1}{D+2}}}{b_0}, & D = 1, \\ \frac{1}{D-1}, & D > 1. \end{cases}$$

证明 定义 $T_j^{\text{sig}} = \inf \{ t \geq 0 : |f_j(t)| \geq |z_j|/2 \}$. 利用守恒量, 发现 $\theta_j(t) \geq |\beta_j(t)|$ 且

$b_j(t) \geq \sqrt{D} |\beta_j(t)|$, 故:

$$|f_j(t)| = |\theta_j(t) b_j^D(t) \beta_j(t)| \leq |\theta_j(t) b_j^D(t)|^{\frac{D+2}{D+1}},$$

从而 $T_j^{\text{eig}} \leq T_j^{\text{sig}}$ 。其余部分可以仿照推论 5.6 的证明进行。 ■

5.8.4.1 定理 5.4 的证明

证明过程类似于第 5.8.3.1 小节中两层情形的证明, 但还必须额外考虑 $b_j(t)$ 项的相互作用。回顾 $b_0 = b_0(n) \asymp n^{-\frac{1}{2(D+2)}}$ 且 $t \leq t_* \asymp b_0^{-D} \sqrt{n/(\log n)} \asymp n^{\frac{D+1}{D+2}} / \sqrt{\log n}$ 。定义:

$$J = \min \left\{ j \geq 1 : \lambda_j^{\frac{1}{2}} \leq b_0 / \sqrt{D} \right\} \asymp b_0^{-2/\gamma} \asymp n^{\frac{1}{(D+2)\gamma}}.$$

遵循第 5.8.3.1 小节的证明, 可以计算出:

$$\delta^*(n^{-1}; \tilde{\lambda}) \geq \delta_0 \asymp b_0^{2D} n^{-\frac{q\gamma}{p+q}} \asymp n^{-\frac{D}{D+2}} n^{-\frac{q\gamma}{p+q}}.$$

与第 5.8.3.1 小节的证明相同, 分别考虑 $j < L$ 和 $j \geq L$ 。对于 $j < L$, 有仍然可以应用比较原理引理 5.20 来证明上穿不可能发生。

现在考虑 $j \geq L$ 的情形并证明 $\tilde{\lambda}_j(t) < \delta_0$ 。对于 $j \geq \max(L, J)$, 利用 $|z_j| \lesssim \sqrt{\log(jn)/n}$ 且 $t \leq t_*$, 发现:

$$b_0^D |z_j| t \lesssim \sqrt{\frac{\log(jn)}{\log n}} \lesssim 1 + \sqrt{\log j / \log n}.$$

然后, 应用 (5.98) 得到:

$$\tilde{\lambda}_j(t) \lesssim b_0^{2D} \lambda_j e^{C b_0^D |z_j| t} \leq b_0^{2D} \lambda_j e^{C \sqrt{\log j}} < \delta_0 = b_0^{2D} n^{-\frac{q\gamma}{p+q}},$$

但必须验证 (5.98) 中的条件。更详细地说明。由于 $j \geq J$, 可以写成:

$$\lambda_j^{\frac{1}{2}} = \kappa b_0 / \sqrt{D}, \quad \kappa \leq 1, \quad \log j \lesssim \log C (\kappa b_0 / \sqrt{D})^{-2/\gamma} \lesssim 1 + \log \kappa^{-1} + \log n.$$

一方面, 有:

$$t/T_j^{(1)} = C b_0^D |z_j| t \lesssim 1 + \sqrt{\log j / \log n} \lesssim 1 + \sqrt{\log \kappa^{-1} / \log n},$$

另一方面, 有:

$$1 + \log \frac{b_0}{\sqrt{D} \lambda_j^{\frac{1}{2}}} = 1 + \log \kappa^{-1} \gtrsim 1 + \sqrt{\log \kappa^{-1} / \log n}.$$

因此, 取 t 中的常数因子足够小, 该条件得到满足。

现在, 只剩下 $J > L$ 且 $j \in [L, J]$ 的情形。应用命题 5.5, 发现:

$$\tilde{\lambda}_j(t) \lesssim \max(\lambda_j^{\frac{1}{2}} b_0^D, |z_j|^{\frac{D+1}{D+2}}, b_0^{-1} |z_j|)^2.$$

第一项如前所述已经满足 $\lambda_j^{\frac{1}{2}} b_0^D \ll \delta_0$ 。为了控制第二项和第三项, 利用 $J > L$ 得到:

$$n^{\frac{1}{(D+2)\gamma}} \gtrsim (n/\log n)^{\frac{q}{p+1}}, \implies n^{\frac{q\gamma}{p+1}} \lesssim (\log n)^{\frac{q\gamma}{p+1}} n^{\frac{1}{D+2}},$$

所以由于 $q > 1$, 存在某个 $s > 0$ 使得:

$$\delta_0 = b_0^{2D} n^{-\frac{q\gamma}{p+q}} \gtrsim b_0^{2D} n^{-\frac{1}{D+2}} n^s \asymp n^{-\frac{D+1}{D+2}} n^s.$$

回到各项数值, 由于 J 仍然关于 n 呈多项式阶, 有 $|z_j| \lesssim \sqrt{\log(jn)/n} \lesssim \sqrt{(\log n)/n}$, 故:

$$|z_j|^{\frac{2(D+1)}{D+2}} \lesssim n^{-\frac{D+1}{D+2}} (\log n)^{\frac{D+1}{D+2}} \ll n^{-\frac{D+1}{D+2}} n^s,$$

且:

$$(b_0^{-1} |z_j|)^2 \lesssim n^{\frac{1}{D+2}} n^{-1} \log n = n^{-\frac{D+1}{D+2}} \log n \ll n^{-\frac{D+1}{D+2}} n^s.$$

因此, 在此情形下得出 $\tilde{\lambda}_j(t) < \delta_0$ 。

最后, 对于特征误差度量, 除了应用引理 5.21 和引理 5.22 外, 可以遵循与第 5.8.3.1 小节相同的论据。

特征误差度量。 现在, 考虑 $t = t_*$ 时刻的特征误差度量。选择 $\delta_* = C n^{-\frac{D+1}{2(D+2)}}$ 。首先考虑 $\mathcal{E}_V(\delta_*, n^{-1}; \tilde{\lambda})$ 。如前所述, 对于 $j \geq L$ 有 $\tilde{\lambda}_j < \delta_0 \leq \delta_*$ 。此外, 对于满足 $|\theta_j^*| \lesssim \sqrt{(\log n)/n}$ 且 $\lambda_j^{\frac{1}{2}} \leq b_0/\sqrt{D}$ 的 $j < L$, 应用引理 5.21 发现 $\tilde{\lambda}_j(t_*) \leq C \tilde{\lambda}_j(0)$ 。因此:

$$\begin{aligned} \#\{j \in N : \tilde{\lambda}_j(t_*) \geq \delta_*\} &\leq \#\left\{j < L : |\theta_j^*| \gtrsim (n \log n)^{\frac{1}{2}}\right\} \\ &\quad + \#\left\{j < L : \lambda_j \geq b_0/\sqrt{D}\right\} \\ &\quad + \#\{j < L : C \tilde{\lambda}_j(0) \geq \delta_*\} \\ &\lesssim n^{\frac{1}{p+1}} + n^{\frac{1}{(D+2)\gamma}}. \end{aligned}$$

现在, 对于 $\mathcal{E}_B(\delta_*, n^{-1}; \tilde{\lambda})$, 利用引理 5.22 发现, 如果 $|\theta_j^*| \gtrsim n^{-\frac{1}{2}} (\log n)^{3/2}$, 那么:

$$\tilde{\lambda}_j(t) \geq \left|\theta_j^*\right|^{\frac{D+1}{D+2}} \gtrsim n^{-\frac{D+1}{2(D+2)}} (\log n)^{\frac{3}{2} \frac{D+1}{D+2}} \geq \delta_*,$$

所以:

$$\sum_{j \in N} (\theta_j^*)^2 \mathbf{1}\{\tilde{\lambda}_j(t_*) \geq \delta_*\} \leq \sum_{j \in N} (\theta_j^*)^2 \mathbf{1}\left\{\left|\theta_j^*\right| \geq n^{\frac{1}{2}}(\log n)^{3/2}\right\} \lesssim n^{-\frac{p}{p+1}}(\log n)^{\frac{2p}{p+1}}.$$

结合这两个部分, 本文证明了特征误差度量的界。

5.8.5 样本损失的情况, 定理 5.5 的证明

下文考虑定理的通用版本。为了以最大的普适性描述真实函数的性质, 引入有关真实系数 $(\theta_j^*)_{j \geq 1}$ 的以下量:

$$\mathcal{N}(\delta; f^*) := \#\{j : |f_j^*| \geq \delta\}, \quad \mathcal{R}(\delta; f^*) := \sum_{j=1}^{\infty} (\theta_j^*)^2 \mathbf{1}\{|\theta_j^*| < \delta\}. \quad (5.101)$$

这两个量分别衡量了真实函数中显著系数的数量和残差项的总和。此外, 它们可以被视为真实函数的最佳特征误差度量, 此时特征映射的权重与真实系数一致。注意到, 由于 $|f_j^*|$ 可能不随 j 递减 (例如, 考虑假设 5.2), 所以这两个量不能简单地通过划分前 L 项和剩余项来获得。在假设 5.2 下, 有 $\mathcal{N}(\delta; f^*) \asymp \delta^{-\frac{2}{p+1}}$ 且 $\mathcal{R}(\delta; f^*) \asymp \delta^p$ 。

那么, 定理 5.5 是以下两个定理的直接推论。

定理 5.6: 考虑在假设 5.3 和假设 5.1 下的对角自适应特征模型 (5.11)。令 $s > 0$ 为一个任意小的常数。那么, 可以选择 $L \asymp (-\frac{1}{2} + s) \log n$, 一个满足 $\delta_L \leq n^{-\frac{1}{2}+s}$ 的递减序列 $\delta_l = C2^{-l}$ (对于 $l \leq L$), 以及满足 $t_l \lesssim \delta_l^{-l} \log n$ 的时间点 $t_0 = 0 < t_1 < \dots < t_L = t_* \lesssim \sqrt{n}$, 使得以至少 $1 - Cn^{-2}$ 的概率, 有:

$$\mathcal{E}^*(n^{-1}; \Phi_{\theta(t)}, f^*) \leq \mathcal{R}(\delta_l; f^*) + \left(\mathcal{N}(n^{-\frac{1}{2}}; f^*) + n^{\frac{1}{2\gamma}} \right) n^{-1}, \quad \forall t \in [t_l, t_*], \forall l = 0, \dots, L, \quad (5.102)$$

特别地,

$$\mathcal{E}^*(n^{-1}; \Phi_{\theta(t_*)}, f^*) \leq \mathcal{R}(n^{-\frac{1}{2}+s}; f^*) + \left(\mathcal{N}(n^{-\frac{1}{2}}; f^*) + n^{\frac{1}{2\gamma}} \right) n^{-1}.$$

此外, 注意到 (5.102) 中的上界随 l 单调递减。

定理 5.7: 考虑在假设 5.3 和假设 5.1 下的对角自适应特征模型 (5.13), 其中选择 $b_0 \asymp n^{-\frac{1}{2(D+2)}}$ 。令 $s > 0$ 为一个任意小的常数。记 $q = 2^{\frac{2(D+1)}{D+2}}$ 。那么, 可以选择 $L \asymp (-\frac{1}{2} + s) \log n$, 一个满足 $\delta_L \leq n^{-\frac{1}{2}+s}$ 的递减序列 $\delta_l = Cq^{-l}$ (对于 $l \leq L$), 以及满足 $t_l \lesssim \delta_l^{-l} \log n$ 的时间点 $t_0 = 0 < t_1 < \dots < t_L = t_* \lesssim n^{\frac{D+1}{D+2}}$, 使得以至少 $1 - Cn^{-2}$ 的概率, 有:

$$\mathcal{E}^*(n^{-1}; \Phi_{\theta(t), b(t)}, f^*) \leq \mathcal{R}(\delta_l; f^*) + \left(\mathcal{N}(n^{-\frac{1}{2}}; f^*) + n^{\frac{1+s}{(D+2)\gamma}} \right) n^{-1}, \quad \forall t \in [t_l, t_*], \forall l = 0, \dots, L. \quad (5.103)$$

特别地,

$$\mathcal{E}^*(n^{-1}; \Phi_{\theta(t_*)}, f^*) \leq \mathcal{R}(n^{-\frac{1}{2}+s}; f^*) + \left(\mathcal{N}(n^{-\frac{1}{2}}; f^*) + n^{\frac{1+s}{(D+2)\gamma}} \right) n^{-1}.$$

5.8.5.1 定理 5.6 的证明

回顾定理 5.1 的证明过程, 见第 5.6 节。记

$$S = S_1 \cup S_2 = \left\{ j \geq 1 : \left| \theta_j^* \right| \geq n^{-1/2} \sqrt{\ln n} \right\} \cup \left\{ j \geq 1 : \lambda_j \geq n^{-1/2} \right\}$$

以及如第 5.6 节中所定义的 $R = S^c$ 。令 $v_1 = Cn^{-1/2+s}$ 且 $L = \lfloor \log_2(B_\infty/v_1) \rfloor$, 其中 $s > 0$ 是一个任意小的常数。证明中的收缩动力学显示, 令 $\delta_i = 2^{-i} B_\infty$, 存在一个递增的时间点序列 t_i , 使得 $t_i \lesssim \delta_i^{-1} \log n$, 且对于 $i = 0, \dots, L$:

$$\|f_S^* - f_S(t)\|_\infty \leq \delta_{i+1}, \quad \forall t \geq t_{i+1}. \quad (5.104)$$

同时, 有 $\delta_{L+1} \lesssim n^{-1/2+s}$ 且 $t_* = t_L \lesssim \sqrt{n}$ 。因此, 有:

$$\theta_k(t_i)^2 \geq |f_k(t_i)| \geq |f_k^*| - \|f_S^* - f_S(t_i)\|_\infty \geq |f_k^*| - \delta_i \geq \delta_i, \quad \forall k, |f_k^*| \geq 2\delta_i. \quad (5.105)$$

另一方面, 噪声项被界定为:

$$|f_k(t)| \leq 2\lambda_k e^{C\sqrt{\ln n + \ln k}} \implies |\theta_k(t)|^2 \leq C\lambda_k e^{C\sqrt{\ln n + \ln k}} \leq n^{-1/2+s},$$

$$\forall k \in R, t \leq t_*,$$

其中 $s > 0$ 是一个任意小的常数。

现在考虑特征误差度量。有:

$$\begin{aligned} \mathcal{E}^*(n^{-1}; \lambda(t), f^*) &\leq \mathcal{E}(\delta_i, n^{-1}; \lambda(t)) \\ &= \# \{k \in N : \theta_k(t)^2 \geq \delta_i\} n^{-1} + \sum_{k \in N} (\theta_k^*)^2 \mathbf{1} \{ \theta_k(t)^2 < \delta_i \}. \end{aligned}$$

对于第一项, 利用对噪声项的控制以及 $\delta_i \geq Cn^{-\frac{1}{2}+s}$, 对于 $t \leq t_*$:

$$\# \{k \in N : \theta_k(t)^2 \geq \delta_i\} \leq |S| \leq \mathcal{N}(n^{-\frac{1}{2}}; f^*) + n^{\frac{1}{2\gamma}}.$$

对于第二项, 利用 (5.105) 得到:

$$\sum_{k \in N} (\theta_k^*)^2 \mathbf{1} \{ \theta_k(t)^2 < \delta_i \} \leq \sum_{k \in N} (\theta_k^*)^2 \mathbf{1} \{ |\theta_k^*| < 2\delta_i \} = \mathcal{R}(2\delta_i; f^*), \quad \forall t \in [t_i, t_*].$$

结合这两个估计就获得了对 $\mathcal{E}^*(n^{-1}; \lambda(t_i), f^*)$ 的上界。特别地, 利用 $\delta_L \lesssim n^{-1/2+s}$, 有:

$$\mathcal{E}^*(n^{-1}; \lambda(t_*), f^*) \leq \mathcal{R}(n^{-\frac{1}{2}+s}; f^*) + \left(\mathcal{N}(n^{-\frac{1}{2}+s}; f^*) + n^{\frac{1+s}{2\gamma}} \right) n^{-1}.$$

5.8.5.2 定理 5.7 的证明

证明类似于两层情况下的第 5.8.5.1 小节，但现在需要仿照定理 5.2 的证明过程，见第 5.7 节。记 $\tilde{\lambda}_j(t) = (\theta_j(t)b_j^D(t))^2$ 。在这种情况下，对于某个小的 $s' > 0$ ，记：

$$S = S_1 \cup S_2 = \left\{ j \geq 1 : |\theta_j^*| \geq n^{-1/2} \sqrt{\ln n} \right\} \cup \left\{ j \geq 1 : \lambda_j \geq n^{-\frac{1+s'}{D+2}} \right\}.$$

依然定义 $\nu_1 = Cn^{-\frac{1}{2}+s_1}$ ， $L = \lfloor \log_2(B_\infty/\nu_1) \rfloor$ 以及 $\delta_i = 2^{-i}B_\infty$ 。收缩动力学显示有一系列递增的时间点 t_i ，使得 (5.104) 成立。此外，有：

$$t_i \lesssim \sum_{j \leq i} \left[\delta_j^{-\frac{2D+2}{D+2}} + \delta_j^{-1} b_0^{-D} \log n \right] \lesssim \delta_i^{-\frac{2D+2}{D+2}} + \delta_i^{-1} b_0^{-D} \log n \lesssim \delta_i^{-1} b_0^{-D} \log n \asymp \delta_i^{-1} n^{\frac{D}{2(D+2)}} \log n,$$

其中最后一个不等式是因为 $\delta_i \gtrsim n^{-\frac{1}{2}+s_1}$ ，所以第二项占主导地位。关于多层情况，有 $t_L \lesssim n^{\frac{D+1}{D+2}}$ 且：

$$\tilde{\lambda}_k(t_i) \geq |f_k(t_i)|^{\frac{2(D+1)}{D+2}} \geq \delta_i^{\frac{2(D+1)}{D+2}}, \quad \forall k, |f_k^*| \geq 2\delta_i.$$

噪声项被界定为：

$$|\tilde{\lambda}_k(t)| \lesssim \lambda_k b_0^{2D} e^{E\sqrt{\ln n + \ln k}} \lesssim n^{-\frac{D+1+s'}{D+2}}.$$

因此，对特征误差度量的控制遵循与第 5.8.5.1 小节相同的论证。

5.9 补充内容

5.9.1 集中不等式

以下是次高斯 (sub-Gaussian) 随机变量的标准 Hoeffding 不等式，参见例如 Vershynin^[102] 的定理 2.6.2。对于一个随机变量 X ，如果存在常数 $K > 0$ 使得对于任意 $t \in \mathbb{R}$ 有

$$\mathbb{E} e^{tX} \leq e^{\frac{K^2 t^2}{2}},$$

则称 X 为 K -次高斯随机变量。

引理 5.23： 令 $X, X_1, \dots, X_n$ 为独立同分布的 K -次高斯随机变量。那么，对于每一个 $t \geq 0$ ，有

$$\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E} X \right| \geq t \right\} \leq 2e^{-\frac{cnt^2}{K^2}},$$

其中 $c > 0$ 是一个绝对常数。

该常数来自标准 Hoeffding 型不等式，其具体数值只会改变后续概率界中的通用常数，因此本文不进一步追踪。

以下引理给出了希尔伯特空间版本的 Hoeffding 不等式, 参见例如 Pinelis et al. ^[103]Theorem 3。

引理 5.24: 令 H 为一个可分希尔伯特空间。令 $X, X_1, \dots, X_n$ 为取值于 H 的独立同分布随机元, 且满足 $\|X_i\| \leq B$ 几乎处处成立。那么, 对于每一个 $t \geq 0$,

$$\mathbb{P} \left\{ \left\| \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}X \right\| \geq t \right\} \leq 2e^{-\frac{nt^2}{2B^2}}. \quad (5.106)$$

5.9.2 第 5.1 节中例子的分析

5.9.2.1 第 5.1.1 小节的分析

第 5.1.1 小节中的推导非常直观。级数展开式 (5.7) 可以写为:

$$[\mathcal{H}_k]^s = \left\{ f = \sum_{j \geq 1} f_j e_j \mid \sum_{j \geq 1} f_j^2 \lambda_j^{-s} < \infty \right\}.$$

回顾 $\lambda_{j,1} \asymp j^{-\gamma_1}$ 且 $\lambda_{j,2} \asymp j^{-\gamma_2}$ 。设 $f^* = \sum_{j \geq 1} f_j^* e_j$ 。那么, 对于 $\ell = 1, 2$, 有:

$$f^* \in [\mathcal{H}_{k_\ell}]^{s_\ell} \iff \sum_{j \geq 1} (f_j^*)^2 \lambda_{j,\ell}^{-s_\ell} < \infty \iff \sum_{j \geq 1} (f_j^*)^2 j^{\gamma_\ell s_\ell} < \infty.$$

因此可得:

$$f^* \in [\mathcal{H}_{k_1}]^{s_1} \iff f^* \in [\mathcal{H}_{k_2}]^{s_2} \quad \text{当且仅当} \quad \gamma_1 s_1 = \gamma_2 s_2.$$

5.9.2.2 第 5.1.2 小节的分析

这里证明第 5.1.2 小节中的结论。考虑环面 $\mathbb{T}^d = [-1, 1)^d$ 以及其上的均匀测度 μ (即 $\mu(\mathbb{T}^d) = 1$)。回顾多维傅里叶基由 $\phi_m(x) = e^{i\pi \langle m, x \rangle}$ 给出, 其中 $m \in \mathbb{Z}^d$ 。

Sobolev 空间 $H^s(\mathbb{T}^d)$ 通过傅里叶系数定义为:

$$H^s(\mathbb{T}^d) = \left\{ f \in L^2(\mathbb{T}^d) \mid \sum_{m \in \mathbb{Z}^d} |f_m|^2 (1 + \|m\|^2)^s < \infty \right\},$$

该空间配备如下内积 (以及由此导出的范数):

$$\langle f, g \rangle_{H^s(\mathbb{T}^d)} = \sum_{m \in \mathbb{Z}^d} f_m \overline{g_m} (1 + \|m\|^2)^s.$$

这里简要证明当 $s > d/2$ 时, $H^s(\mathbb{T}^d)$ 是一个再生核希尔伯特空间 (RKHS)。只需证明对于每个 $x \in \mathbb{T}^d$, 映射 $f \mapsto f(x)$ 是有界的^[13]。利用 ϕ_m 的有界性, 有:

$$\sum_{m \in \mathbb{Z}^d} |f_m \phi_m| \leq \sum_{m \in \mathbb{Z}^d} |f_m| \leq \left[\sum_{m \in \mathbb{Z}^d} |f_m|^2 (1 + \|m\|^2)^s \right]^{\frac{1}{2}} \left[\sum_{m \in \mathbb{Z}^d} (1 + \|m\|^2)^{-s} \right]^{\frac{1}{2}}.$$

进一步计算：

$$\sum_{\mathbf{m} \in \mathbb{Z}^d} (1 + \|\mathbf{m}\|^2)^{-s} \lesssim \int_{x \in \mathbb{R}^d} (1 + |x|^2)^{-s} dx \lesssim \int_0^\infty (1 + r^2)^{-s} r^{d-1} dr.$$

由于当 $s > d/2$ 时最后的积分是有限的，可以找到一个常数 C 使得：

$$\sum_{\mathbf{m} \in \mathbb{Z}^d} |f_{\mathbf{m}} \phi_{\mathbf{m}}| \leq C \|f\|_{H^s(\mathbb{T}^d)}.$$

因此，级数展开 $f(x) = \sum_{\mathbf{m} \in \mathbb{Z}^d} f_{\mathbf{m}} \phi_{\mathbf{m}}(x)$ 是绝对收敛且一致收敛的，从而 $|f(x)| \leq \sum_{\mathbf{m} \in \mathbb{Z}^d} |f_{\mathbf{m}} \phi_{\mathbf{m}}| \leq C \|f\|_{H^s(\mathbb{T}^d)}$ ，这表明对于每个 $x \in \mathbb{T}^d$ ， $f \mapsto f(x)$ 都是有界的。

此外，从 Mercer 分解 (3.8) 和幂级数展开 (5.7) 很容易看出， $H^s(\mathbb{T}^d)$ 的核由下式给出：

$$k(x, x') = \sum_{\mathbf{m} \in \mathbb{Z}^d} (1 + \|\mathbf{m}\|^2)^{-s} \phi_{\mathbf{m}}(x) \overline{\phi_{\mathbf{m}}(x')},$$

因此其特征值为 $\lambda_{\mathbf{m}} = (1 + \|\mathbf{m}\|^2)^{-s}$ 。

为了确定按降序重新排列后的特征值衰减率 $\lambda_{\mathbf{m}} = (1 + \|\mathbf{m}\|^2)^{-s}$ ，只需确定计数函数 $\#\{\mathbf{m} : \lambda_{\mathbf{m}} > \delta\}$ ：如果 $\#\{\mathbf{m} : \lambda_{\mathbf{m}} > \delta\} \asymp \delta^{-1/\beta}$ ，则特征值衰减率为 β （例如参见命题 3.8）。有：

$$\begin{aligned} \#\{\mathbf{m} : \lambda_{\mathbf{m}} > \delta\} &= \#\{\mathbf{m} : (1 + \|\mathbf{m}\|^2)^{-s} > \delta\} \\ &\asymp \text{Vol}(\{x \in \mathbb{R}^d : (1 + |x|^2)^{-s} > \delta\}) \\ &\asymp \text{Vol}(\{x \in \mathbb{R}^d : |x| < \delta^{-\frac{1}{2s}}\}) \asymp \delta^{-\frac{d}{2s}}. \end{aligned}$$

考虑具有低维结构的函数 $f^*(x) = g(x_1, \dots, x_{d_0})$ 。为了简便，记 $x_{\leq d_0} = (x_1, \dots, x_{d_0})$ 且 $x_{> d_0} = (x_{d_0+1}, \dots, x_d)$ 。那么， f^* 的傅里叶系数由下式给出：

$$\begin{aligned} f_{\mathbf{m}} &= \langle f^*, \phi_{\mathbf{m}} \rangle_{L^2(\mathbb{T}^d, d\mu)} \\ &= 2^{-d} \int_{\mathbb{T}^{d_0} \times \mathbb{T}^{d-d_0}} g(x_1, \dots, x_{d_0}) e^{i\pi \langle \mathbf{m}_{\leq d_0}, x_{\leq d_0} \rangle} \cdot e^{i\pi \langle \mathbf{m}_{> d_0}, x_{> d_0} \rangle} dx_{\leq d_0} dx_{> d_0} \\ &= 2^{-d_0} \int_{\mathbb{T}^{d_0}} g(x_1, \dots, x_{d_0}) e^{i\pi \langle \mathbf{m}_{\leq d_0}, x_{\leq d_0} \rangle} dx_{\leq d_0} \\ &\quad \cdot 2^{-(d-d_0)} \int_{\mathbb{T}^{d-d_0}} e^{i\pi \langle \mathbf{m}_{> d_0}, x_{> d_0} \rangle} dx_{> d_0} \\ &= g_{\mathbf{m}_{\leq d_0}} \cdot \mathbf{1}_{\{\mathbf{m}_{> d_0} = \mathbf{0}\}}, \end{aligned}$$

这证明了：

$$f_{\mathbf{m}} = \begin{cases} g_{\mathbf{m}_{\leq d_0}}, & \mathbf{m} = (\mathbf{m}_{\leq d_0}, \mathbf{0}), \mathbf{m}_{\leq d_0} \in \mathbb{Z}^{d_0}, \\ 0, & \text{否则.} \end{cases}$$

回顾 $g \in H^t(\mathbb{T}^{d_0})$, 因此:

$$\sum_{\mathbf{m} \leq d_0} \left| g_{\mathbf{m} \leq d_0} \right|^2 (1 + \|\mathbf{m} \leq d_0\|^2)^t < \infty.$$

为了确定 f^* 在 $[\mathcal{H}_{k_1}]^s$ 和 $[\mathcal{H}_{k_2}]^s$ 上的光滑性, 根据 (5.7) 计算:

$$\begin{aligned} \sum_{\mathbf{m} \in \mathbb{Z}^d} |f_{\mathbf{m}}|^2 [(1 + \|\mathbf{m}\|^2)^r]^s &= \sum_{\mathbf{m}=(\mathbf{m} \leq d_0, \mathbf{0}), \mathbf{m} \leq d_0 \in \mathbb{Z}^{d_0}} \left| g_{\mathbf{m} \leq d_0} \right|^2 \left[(1 + \|\mathbf{m} \leq d_0\|^2)^r \right]^s \\ &= \sum_{\mathbf{m} \leq d_0 \in \mathbb{Z}^{d_0}} \left| g_{\mathbf{m} \leq d_0} \right|^2 (1 + \|\mathbf{m} \leq d_0\|^2)^{rs}, \end{aligned}$$

因此当 $s = t/r$ 时, f^* 属于 $[\mathcal{H}_{k_1}]^s$ 和 $[\mathcal{H}_{k_2}]^s$.

5.9.2.3 第 5.1.3 小节的分析

回顾 $f^* = \sum_{j \geq 1} \theta_j^* e_j$, 且:

$$|\theta_{\ell(j)}^*| \asymp j^{-(p+1)/2} \quad \text{且} \quad \ell(j) \asymp j^q \quad \text{对于} \quad p > 0, q \geq 1,$$

其中 $\ell(j)$ 给出了 $|\theta_j^*|$ 的降序排列。为了计算 f^* 相对于 $\lambda_{j,1} \asymp j^{-\gamma}$ 的相对光滑度, 计算:

$$\sum_{j \geq 1} |\theta_j^*|^2 \lambda_j^{-s} = \sum_{j \geq 1} |\theta_{\ell(j)}^*|^2 \lambda_{\ell(j)}^{-s} \asymp \sum_{j \geq 1} j^{-(p+1)} (\ell(j))^{\gamma s} \asymp \sum_{j \geq 1} j^{-(p+1)} j^{q\gamma s} \asymp \sum_{j \geq 1} j^{-1-(p-q\gamma s)},$$

因此得到 $s < p/(q\gamma)$ (但可以任意接近), 相应的泛化误差速率为 $\frac{s\gamma}{s\gamma+1} < \frac{p}{p+q}$ 。相对于 $\lambda_{\ell(j),2} \asymp j^{-\gamma}$ 的泛化误差速率可以类似地计算。

5.9.3 杂项结果

命题 5.6: 令 $k > 0, p > 1$ 且 $M > 0$ 。考虑常微分方程 $\dot{x} \geq kx^p$, 初始条件为 $x(0) = x_0 > 0$, 有

$$x(t) \geq \left(x_0^{-(p-1)} - (p-1)kt \right)^{-\frac{1}{p-1}}, \quad \text{且} \quad \inf \{ t \geq 0 : x(t) \geq M \} \leq \left[(p-1)kx_0^{p-1} \right]^{-1}.$$

类似地, 对于方程 $\dot{x} \leq -kx^p$, 初始条件为 $x(0) = x_0 > 0$, 有

$$x(t) \leq \left(x_0^{-(p-1)} + (p-1)kt \right)^{-\frac{1}{p-1}}, \quad \text{且} \quad \inf \{ t \geq 0 : x(t) \leq M \} \leq \left[(p-1)kM^{p-1} \right]^{-1}.$$

引理 5.25: 假设 $(\theta_j)_{j \geq 1}$ 满足 $|\theta_{l(j)}| \asymp j^{-(p+1)/2}$ (对于某些 $p > 0$), 且在其他位置 $|\theta_j| = 0$, 其中 $l(j)$ 是一个指标序列。按照 (5.14) 中定义 $\Phi(\delta)$ 和 $\Psi(\delta)$, 有:

$$\Phi(\delta) \asymp \delta^{-\frac{2}{p+1}}, \quad \Psi(\delta) \asymp \delta^{\frac{2p}{p+1}}.$$

证明 首先, 从 $\Phi(\delta)$ 和 $\Psi(\delta)$ 的定义可以看出, 它们不依赖于指标的排列顺序, 也不受

θ_j 中零值的影响。因此,可以假设 $l(j) = j$ 。那么,假设 $c_1 j^{-(p+1)/2} \leq |\theta_j| \leq C_1 j^{-(p+1)/2}$, 有:

$$\Phi(\delta) = |\{j : |\theta_j| \geq \delta\}| \leq |\{j : C_1 j^{-(p+1)/2} \geq \delta\}| \leq (\delta/C_1)^{-\frac{2}{p+1}}.$$

此外,

$$\begin{aligned} \Psi(\delta) &= \sum_{j=1}^{\infty} |\theta_j|^2 \mathbf{1}\{|\theta_j| < \delta\} \\ &= \sum_{j > \Phi(\delta)} |\theta_j|^2 \leq C_1^2 \sum_{j > \Phi(\delta)} j^{-(p+1)} \\ &\leq C_1^2 C \Phi(\delta)^{-p} \leq C' \delta^{\frac{2p}{p+1}}, \end{aligned}$$

对于某些常数 $C' > 0$ 。这些量的下界可以类似地得到。 ■

第6章 方向参数化模型的自适应特征理论

本章研究具有可学习基函数的自适应特征模型，并将其应用于高斯单指标模型（Single-index models）和多指标模型（Multi-index models）。与前一章中的对角参数化不同，这里特征学习的核心不再是重新分配已有特征的权重，而是在训练过程中主动学习与目标函数相关的方向结构。因此，本章关注的问题是：梯度下降能否驱动特征映射逐步对齐潜在子空间，并由此提升模型的表达能力与泛化性能？

单指标和多指标模型在大量的现有文献中已得到广泛研究^[80,104-109]。早期工作^[110-111]利用了温和的数据分布假设下链接函数（link function）的可逆性或单调性等性质来实现学习；而 Dudeja et al.^[105]、Arous et al.^[109]和 Arous et al.^[112]为高斯数据上的（随机）梯度下降开发了调和框架，为单指标模型提供了定量保证。多指标模型的扩展^[113-114]涉及半参数学习和样本复杂度，通常关注特定的链接函数结构，如阶梯性质（staircase property）。然而，现有工作大多聚焦于特定估计程序、链接函数结构或样本复杂度分析，较少从“可学习特征映射”这一角度统一理解这些模型。本章将在自适应特征模型的框架下重新考察单指标与多指标问题，说明方向学习如何在训练过程中显式体现在特征表示的演化之中，并据此给出一个新的理论视角。

6.1 Hermite 多项式与高斯分布

本节汇总了 Hermite 多项式以及高斯分布的一些有用性质，这是后续研究单指标和多指标模型的基础。用 γ_d 表示 $\mathbb{R}^d$ 上的标准高斯测度，概率密度如下所示：

$$p(x) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{\|x\|^2}{2}}.$$

令 d 为维度，协变量 x 服从 d 维标准高斯测度 $\gamma_d = N(0, I_d)$ 。用 $\langle \cdot, \cdot \rangle_{\gamma_d}$ 表示 $L^2(\mathbb{R}^d, \gamma_d)$ 上的内积。

高斯指标模型的一个基本组成部分是 Hermite 多项式（Hermite polynomials），它们关于高斯测度是正交的。用 H_r 表示关于标准高斯测度 γ_1 归一化的 Hermite 多项式，使得

$$\langle H_r, H_s \rangle_{\gamma_1} = \mathbb{E}_{x \sim N(0,1)} H_r(x) H_s(x) = \delta_{rs}, \quad (6.1)$$

其中 δ_{rs} 是 Kronecker delta 函数。具体地，此处的 Hermite 多项式如下给出

$$H_r(x) = (-1)^r \frac{1}{\sqrt{r!}} e^{\frac{x^2}{2}} \frac{d^r}{dx^r} e^{-\frac{x^2}{2}}. \quad (6.2)$$

它们也满足如下的生成函数：

$$e^{xt - \frac{t^2}{2}} = \sum_{r=0}^{\infty} \frac{H_r(x)}{\sqrt{r!}} t^r. \quad (6.3)$$

众所周知，Hermite 多项式序列 $(H_r)_{r \geq 0}$ 构成了 $L^2(\mathbb{R}, \gamma_1)$ 的一组标准正交基。

注 6.1： 有多种不同的 Hermite 多项式定义方式。本文采用归一化（概率型）Hermite 多项式，其与物理学中常用的物理型 Hermite 多项式存在一个归一化因子差异。具体而言，物理型 Hermite 多项式 $\tilde{H}_r$ 与本文使用的归一化 Hermite 多项式 H_r 之间的关系为 $H_r(x) = c_r \tilde{H}_r(x/\sqrt{2})$ ，其中 c_r 是一个归一化常数。在经典文献中，未归一化的物理型 Hermite 多项式常用 $H_n(x)$ 表示，而概率型 Hermite 多项式常用 $\text{He}_n(x)$ 表示。但在本文中只会使用上述定义的归一化的概率型 Hermite 多项式。

对于高维情形，需要引入一些关于多重指标的记号。令 $\mathbf{m} = (m_1, \dots, m_d)$ 为一个多重指标。记其阶数为 $|\mathbf{m}| = m_1 + \dots + m_d$ 。令 $\mathbf{m}, \mathbf{n}$ 为 $\mathbb{N}^d$ 中的多重指标，约定多重指标的阶乘和组合数为

$$\mathbf{m}! := \prod_{i=1}^d m_i!, \quad \binom{\mathbf{m}}{\mathbf{n}} := \prod_{i=1}^d \binom{m_i}{n_i}.$$

此外，对于 $t \in \mathbb{R}^d$ ，定义多重指标的幂次为

$$t^{\mathbf{m}} := \prod_{i=1}^d t_i^{m_i}.$$

定义张量化（多重指标）Hermite 多项式为

$$H_{\mathbf{m}}(x) = \prod_{j=1}^d H_{m_j}(x_j), \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d. \quad (6.4)$$

这是一个总次数为 $|\mathbf{m}|$ 的多元多项式。多重指标 Hermite 多项式的生成函数由下式给出：

$$e^{\langle x, t \rangle - \frac{1}{2} \|t\|^2} = \sum_{\mathbf{m} \in \mathbb{N}^d} \frac{H_{\mathbf{m}}(x)}{\sqrt{\mathbf{m}!}} t^{\mathbf{m}}.$$

多重指标 Hermite 多项式 $\{H_{\mathbf{m}} : \mathbf{m} \in \mathbb{N}^d\}$ 构成了 $L^2(\mathbb{R}^d, \gamma_d)$ 的一组标准正交基。因此任何函数 $f \in L^2(\mathbb{R}^d, \gamma_d)$ 都可以展开为

$$f(x) = \sum_{\mathbf{m} \in \mathbb{N}^d} f_{\mathbf{m}} H_{\mathbf{m}}(x), \quad f_{\mathbf{m}} = \langle f, H_{\mathbf{m}} \rangle_{\gamma_d}. \quad (6.5)$$

在本章中，为了简化分析，始终考虑对应于序列模型 (4.8) 的序列损失 (4.9) 作为经验损失的近似 (4.2)，考虑对应的自适应特征动力学演化方程 (4.10)。具体地，利用标准正交基 $\{H_m : m \in \mathbb{N}^d\}$ ，考虑如 (4.8) 所示的高斯序列模型：

$$z_m = f_m^* + \varepsilon_m, \quad f_m^* = \langle f^*, H_m \rangle_{\gamma_d}, \quad \varepsilon_m \stackrel{i.i.d.}{\sim} N(0, 1/n), \quad m \in \mathbb{N}^d, \quad (6.6)$$

其中 f_m^* 代表目标函数的真实系数， ε_m 是噪声项。集合 $(z_m)_{m \in \mathbb{N}^d}$ 构成了观测数据。此外，如 (4.9) 所示，对于 $\mathbb{R}^d$ 上的候选函数 f ，序列损失定义为：

$$\bar{\mathcal{L}}_n(f) = \frac{1}{2} \sum_{m \in \mathbb{N}^d} (f_m - z_m)^2, \quad f_m = \langle f, H_m \rangle_{\gamma_d}. \quad (6.7)$$

与第 4.3 节中的讨论一致，上式在无限指标集上应理解为有限阶 Hermite 截断损失的极限或形式化等价代理；本章后续使用的梯度方程均对应于这一截断问题逐坐标求导后的极限形式。

6.2 单指标模型

首先考虑单指标模型，其真实函数由下式给出：

$$f^*(x) = g^*(\langle w_*, x \rangle), \quad (6.8)$$

其中单位向量 $w_* \in \mathbb{S}^{d-1}$ 是未知的方向， $g^* \in L^2(\gamma_1)$ 是未知的链接函数。令 $g^* = \sum_{r \geq 0} g_r^* H_r$ 为 g^* 关于 Hermite 多项式的展开。

由于本章的目标是学习未知方向 w_* ，本章考虑由下式给出的参数化特征映射：

$$\Phi_w(x) = \left(\lambda_r^{\frac{1}{2}} H_r(\langle w, x \rangle) \right)_{r \geq 0}, \quad w \in \mathbb{S}^{d-1}, \quad (6.9)$$

其中 w 是旨在学习真实方向的可训练向量， $(\lambda_r)_{r \geq 0}$ 是固定的可求和正权重序列。对应于以 Hermite 多项式作为特征函数且特征值呈指数级衰减的高斯核^[115]，对于某些固定的 $\gamma > 0$ ，取 $\lambda_r = e^{-\gamma r}$ 。这一取法主要用于保证权重可求和并使高阶 Hermite 分量快速衰减，从而简化级数和扰动项的控制；其他具有足够快衰减和相应正则性的权重参数化也可能导出类似的现象，但需要重新检查后续证明中的级数估计。令 $\beta \in \ell^2(\mathbb{N})$ 为泛函系数参数。参数化的回归函数由下式给出，对应 (4.1)：

$$f(x) = \langle \beta, \Phi_w(x) \rangle_{\ell^2(\mathbb{N})} = \sum_{r \geq 0} \beta_r \lambda_r^{\frac{1}{2}} H_r(\langle w, x \rangle).$$

自适应特征模型的训练过程由下式给出，对应 (4.3)：

$$\begin{cases} \dot{\beta}(t) = -\nabla_{\beta} \bar{\mathcal{L}}_n, & \beta(0) = \mathbf{0}, \\ \dot{w}(t) = -\nabla_w^{\mathbb{S}^{d-1}} \bar{\mathcal{L}}_n, & w(0) \sim \text{Unif}(\mathbb{S}^{d-1}), \end{cases} \quad (6.10)$$

其中 $\nabla_w^{\mathbb{S}^{d-1}}$ 表示球面 $\mathbb{S}^{d-1}$ 上的梯度, $\text{Unif}(\mathbb{S}^{d-1})$ 是球面上的均匀分布。具体而言, 假设 $\nabla_w F$ 是 $\mathbb{R}^d$ 上函数 F 的经典梯度, 则球面上的梯度由 $\nabla_w^{\mathbb{S}^{d-1}} F = P_w^\perp \nabla_w F$ 给出, 其中 $P_w^\perp x = x - \langle w, x \rangle w$ 是到 $\mathbb{S}^{d-1}$ 在 w 处切空间的正交投影。

关于单指标模型, 先前的文献^[109,112]观察到, g^* 展开式中第一个非零系数(被称为 g^* 的信息指数)极大地影响了单指标模型的样本复杂度。本章通过以下假设将其形式化。

假设 6.1: 函数 g^* 是固定的, 其信息指数为 $r_0 := \min\{r \geq 1 : g_r^* \neq 0\}$ 。

此外, 本章还引入以下关于 g^* 系数衰减的假设。

假设 6.2: 存在 $\alpha > 0$, 使得 g_r^* 的系数满足 $|g_r^*| \lesssim r^{-\frac{\alpha+1}{2}}$ 。

为了理解自适应特征模型的训练过程, 本章的第一个定理考虑了总体动力学, 即将序列损失函数 $\bar{\mathcal{L}}_n$ 替换为其总体版本 $\mathcal{L}(f) = \frac{1}{2} \int_{\mathbb{R}^d} (f(x) - f^*(x))^2 d\gamma_d(x)$ 。

定理 6.1 (单指标模型总体动力学): 考虑在假设 6.1 下 (6.10) 的总体动力学版本。那么, 相对于随机初始化, 以概率 1 成立:

$$\mathcal{E}(\delta, \epsilon^2; \Phi_{w(t)}, f^*) \text{ 关于 } t \text{ 单调递减}, \quad (6.11)$$

$\mathcal{E}^*(\epsilon^2; \Phi_{w(t)}, f^*)$ 亦然。此外, 在假设 6.2 下, 以至少 $0.99 - C e^{-cd}$ 的概率, 下式成立:

$$\mathcal{E}^*(\epsilon^2; \Phi_{w(0)}, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{w_*}, f^*) = \Theta(1), \quad (6.12)$$

且存在 $T_1 \lesssim \log d + d^{r_0-1}$ 使得

$$\mathcal{E}^*(\epsilon^2; \Phi_{w(T_1+s)}, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{w_*}, f^*) \lesssim e^{-Cs}, \quad \forall s \geq 0.$$

下一个结果表明, 自适应特征模型可以降低噪声观测下的特征误差度量。

定理 6.2: 考虑在假设 6.1 下 (6.10) 中定义的模型。进一步假设对于某些 $s > 0$, $n \gtrsim d^{2r_0+s}$ 。那么, 对于初始化和噪声的随机性, 以至少 $0.99 - C e^{-cd}$ 的概率, 存在时间 $T_0 = \Theta(1) \leq T_1 \leq T_2 \lesssim \log d + \log n + d^{r_0-1}$ 使得

$$\mathcal{E}(\delta, \epsilon^2; \Phi_{w(t)}, f^*) \text{ 在 } t \in [T_0, T_2] \text{ 内单调递减}, \quad (6.13)$$

且在假设 6.2 下下式成立:

$$\begin{aligned} \mathcal{E}^*(\epsilon^2; \Phi_{w(T_1)}, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{w_*}, f^*) &= \Theta(1), \\ \mathcal{E}^*(\epsilon^2; \Phi_{w(T_1+s)}, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{w_*}, f^*) &\lesssim e^{-Cs}, \quad \forall s \in [0, T_2 - T_1], \\ \mathcal{E}^*(\epsilon^2; \Phi_{w(T_2)}, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{w_*}, f^*) &\lesssim \left(\frac{d}{n}\right)^{\min(\alpha, 1)} \text{polylog}(n, d). \end{aligned} \quad (6.14)$$

定理 6.1 和定理 6.2 分析了自适应特征模型在单指标模型中的表现, 同时考虑了总体(无噪声)动力学和有噪声设置。这些结果阐明了自适应特征映射 $\Phi_{w(t)}$ 是

如何随时间学习的，从而逼近用于表示目标函数 f^* 的最优特征映射 Φ_{w_*} 。本章在下文中讨论其中的关键方面。

对齐度的衡量 单指标模型中存在可识别性问题，因为 w 和 w_* 之间的对齐程度不能直接用它们的距离来衡量，这是由于模型的固有对称性——即翻转 w_* 的符号并相应调整 g^* 不会改变函数 f^* 。通过在本文的框架下考虑超额特征误差 $\mathcal{E}^*(\epsilon^2; \Phi_w, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{w_*}, f^*)$ 作为对齐度量，可以自然地解决这个问题，因为通过直接评估 Φ_w 代表 f^* 的能力自然地避开了可识别性问题。尽管如此，本文指出本章的结果可以隐含文献中常用的其他对齐度量（如 $1 - |\langle w, w_* \rangle|$ ）。

改进特征误差度量 由于 $w(0)$ 的随机初始化，初始超额特征误差处于常数水平 $\Theta(1)$ 。随着训练的进行，自适应特征模型有效地降低了误差。在总体设置中，定理 6.1 确立了特征误差度量 $\mathcal{E}(\delta, \epsilon^2; \Phi_{w(t)}, f^*)$ 以及 $\mathcal{E}^*(\epsilon^2; \Phi_{w(t)}, f^*)$ 随训练时间 t 的增加而单调递减。此外，在特定时间 T_1 之后，超额误差呈指数级快速衰减，显示出特征映射 $\Phi_{w(t)}$ 向最优映射的改进。

在存在噪声的情况下，定理 6.2 也表明 $\mathcal{E}(\delta, \epsilon^2; \Phi_{w(t)}, f^*)$ 在初始阶段之后也表现出递减趋势。此外，在时间 T_2 之前也观察到了类似的指数衰减，此时超额误差接近速率 $(d/n)^{\min(\alpha, 1)}$ （在对数因子内）。如果链接函数 g^* 足够光滑，即 $\alpha \geq 1$ ，它达到了参数化速率 d/n （在对数因子内）；当 $\alpha < 1$ 时，该速率受限于链接函数有限的光滑性，导致收敛速度较慢。

对齐与逼近误差 由于 $\mathcal{E}^*(\epsilon^2; \Phi_{w_*}, f^*)$ 也由链接函数的光滑性决定，可以进一步得到最终总特征误差度量作为以下推论。

推论 6.1: 在与定理 6.2 相同的条件下，此外还成立：

$$\mathcal{E}^*(n^{-1}; \Phi_{w(T_2)}, f^*) \lesssim \left(\frac{d}{n}\right)^{\min(\alpha, 1)} \text{polylog}(n, d) + n^{-\frac{\alpha}{\alpha+1}}.$$

如推论 6.1 所示，最终特征误差由两项组成：第一项代表方向的对齐误差，而第二项捕获了链接函数的逼近误差。在结果中可以观察到一种有趣的相变现象。忽略对数因子，如果 $\alpha \geq 1$ ，当且仅当 $n \geq d^{1+\alpha}$ 时逼近误差占主导；而如果 $\alpha \leq 1$ ，当且仅当 $n \geq d^{1+\alpha^{-1}}$ 时逼近误差占主导。因此临界指数是 $1 + \max(\alpha, \alpha^{-1})$ 。这表明存在一个有趣的相变现象：当 α 很大或很小时，学习对齐至关重要；而当 α 适中时，主要误差源来自学习链接函数。

学习阶段 与第 5 章中基函数固定且特征误差度量通常平滑递减的训练动力学相比，方向自适应特征表现出更复杂的行为，可以分为三个阶段。在 $t \in [0, T_0]$ 的初

始化阶段, 模型通过学习 g^* 对应的系数, 在信息指数处识别链接函数的信号分量, 而方向 w 几乎保持不变。在第二阶段, 微弱但可识别的信号允许模型从头开始学习方向, 这反过来又进一步放大了信号。这一阶段耗时 $T_1 \lesssim \log d + d^{r_0-1}$, 因此较大的信息指数 r_0 会导致更长的训练时间。最后, 当方向基本学成后, 进入最终收敛阶段, 通过精细化方向, 特征误差度量呈指数级快速下降。这三个阶段演示了带有梯度下降的自适应特征如何能同时学习特征映射和链接函数。

信息指数 r_0 的影响 正如先前的文献^[109,112]所观察到的, 信息指数 r_0 在训练动力学和样本复杂度中都起着至关重要的作用。一方面, 它决定了自适应特征模型学习方向所需的时间, 如 $T_1 \lesssim \log d + d^{r_0-1}$ 。另一方面, 样本复杂度 $n \gtrsim d^{2r_0+s}$ 也取决于信息指数。直观地说, 信息指数决定了识别链接函数 g^* 信号分量的难度。在本章的结果中, 虽然对信息指数 r_0 的依赖性与以往专注于单指标模型的工作^[107,109]相比并非最优, 但这些结果已足以说明自适应特征方案的潜力, 进一步改进留待未来研究。

6.3 多指标模型

单指标模型的自适应特征结果可以扩展到多指标模型。定义 Stiefel 流形 $\text{St}(d, p) = \{W \in \mathbb{R}^{d \times p} : W^\top W = I_p\}$ 为具有正交规范列的 $d \times p$ 矩阵集合。多指标模型由下式给出:

$$f^*(x) = g^*(W_*^\top x), \quad W_* \in \text{St}(d, p^*), \quad (6.15)$$

其中 W_* 是未知方向, $g^* \in L^2(\gamma_{p^*})$ 是未知的低维链接函数。此外, 令 $g^* = \sum_{m \in \mathbb{N}^{p^*}} g_m^* H_m$ 为 g^* 关于 p^* 维 Hermite 多项式的展开。

对于多指标模型, 本章同样考虑由下式给出的参数化特征映射:

$$\Phi_W(x) = \left(\lambda_m^{\frac{1}{2}} H_m(W^\top x) \right)_{m \in \mathbb{N}^p}, \quad W \in \text{St}(d, p), \quad (6.16)$$

其中 W 是代表方向的可训练矩阵, $(\lambda_m)_{m \in \mathbb{N}^p}$ 是固定的可求和正权重序列。特别地, 取 $\lambda_m = e^{-\gamma|m|}$ (对于某些固定的 $\gamma > 0$), 这对应于单指标模型中特征映射的张量化版本, 也同样的为了保证权重可求和以及高阶分量的快速衰减。令 $\beta \in \ell^2(\mathbb{N}^p)$ 为泛函系数参数。那么, 预测器由下式给出:

$$f(x) = \langle \beta, \Phi_W(x) \rangle_{\ell^2(\mathbb{N}^p)} = \sum_{m \in \mathbb{N}^p} \beta_m \lambda_m^{\frac{1}{2}} H_m(W^\top x).$$

多指标模型与单指标模型有显著不同, 由于其更高维的方向分量, 其结构更加复杂。在单指标模型中, 方向 w 在正负号意义下是可识别的; 而多指标模型涉

及矩阵 $W \in \mathcal{S}(d, p)$, 它代表了由正交投影 $WW^\top$ 对应的子空间, 该子空间仅在正交变换下是唯一的。具体而言, 对于任何正交矩阵 $Q \in O(p)$, W 和 WQ 张成相同的子空间, 因此若相应地调整 g , 函数 $f = g(W^\top x)$ 保持不变。这种旋转歧义性为分析带来了额外的技术挑战。为了应对这种复杂性并专注于本质的统计性质, 引入以下关于函数 g^* 旋转不变性的假设。

假设 6.3: 假设 $p = p^*$ 是固定的, 且 $g^* \in L^2(\gamma_p)$ 是一个固定的旋转不变函数。

虽然假设 g^* 是旋转不变的, 但多指标模型的复杂性依然存在, 主要在于估计由 W_* 张成的子空间。因此, 即使在这一假设下, 单指标和多指标模型之间仍存在实质性差异。假设 6.3 允许通过关注子空间而非其具体取向来部分简化分析, 从而更有效地研究模型的核心行为。

对于梯度训练过程, 本章还将通过限制系数 β 来保持函数 g^* 的旋转不变性。引入代表旋转不变函数的系数子空间为

$$\mathcal{G}_\lambda(p) = \left\{ \beta \in \ell^2(\mathbb{N}^p) : f = \sum_{m \in \mathbb{N}^p} \lambda_m^{\frac{1}{2}} \beta_m H_m \text{ 是旋转不变的} \right\}.$$

用 $\nabla_\beta^{\mathcal{G}_\lambda(p)}$ 表示子空间 $\mathcal{G}_\lambda(p)$ 中的梯度, 用 $\nabla_W^{\text{St}(d, p)}$ 表示 Stiefel 流形上的梯度。考虑以下自适应特征模型

$$\begin{cases} \dot{\beta}(t) = -\nabla_\beta^{\mathcal{G}_\lambda(p)} \bar{\mathcal{L}}_n, & \beta(0) = \mathbf{0}, \\ \dot{W}(t) = -\nabla_W^{\text{St}(d, p)} \bar{\mathcal{L}}_n, & W(0) \sim \text{Unif}(\text{St}(d, p)), \end{cases} \quad (6.17)$$

其中初始化 $W(0) \sim \text{Unif}(\text{St}(d, p))$ 在 Stiefel 流形上均匀分布。

类似于单指标模型, 本章也在多指标模型中引入函数 g^* 的信息指数, 即 g^* 展开式中非零系数的最小阶数。

假设 6.4: g^* 的信息指数为 $m_0 := \min\{|m| : g_m^* \neq 0\}$ 。

此外, 本章对 g^* 系数的衰减做出以下假设, 其中衰减率中的项 p 确保了系数的平方可求和性。

假设 6.5: g^* 的系数满足 $|g_m^*| \lesssim |m|^{-\frac{\alpha+p}{2}}$ (对于某些 $\alpha > 0$)。

本节的第一个结果展示了总体动力学的收敛性。

定理 6.3 (总体动力学): 考虑在假设 6.3 和假设 6.4 下 (6.17) 的总体版本。那么, 对于随机初始化, 以概率 1 成立:

$$\mathcal{E}(\delta, \epsilon^2; \Phi_{W(t)}, f^*) \text{ 关于 } t \text{ 单调递减}, \quad (6.18)$$

$\mathcal{E}^*(\epsilon^2; \Phi_{W(t)}, f^*)$ 亦然。此外, 以至少 $0.99 - Ce^{-cd}$ 的概率, 下式成立:

$$\mathcal{E}^*(\epsilon^2; \Phi_{W(0)}, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{W_*}, f^*) = \Theta(1), \quad (6.19)$$

且在假设 6.5 下, 存在 $T_0 \lesssim \log d + d^{m_0-1}$ 使得

$$\mathcal{E}^*(\epsilon^2; \Phi_{W(T_0+s)}, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{W_*}, f^*) \lesssim e^{-Cs}, \quad \forall s \geq 0. \quad (6.20)$$

考虑序列损失, 本文给出以下结果。

定理 6.4: 考虑在假设 6.3、假设 6.4 和假设 6.5 下 (6.17) 中定义的模型。进一步假设对于某些 $s > 0$, $n \gtrsim d^{2m_0+1+s}$ 。那么, 对于初始化和噪声的随机性, 以至少 $0.99 - Ce^{-cd}$ 的概率, 存在时间 $T_1 \leq T_2 \lesssim \log d + \log n + d^{m_0-1}$ 使得

$$\begin{aligned} \mathcal{E}^*(\epsilon^2; \Phi_{W(0)}, f^*) &\geq \mathcal{E}^*(\epsilon^2; \Phi_{W(T_1)}, f^*) = \Theta(1), \\ \mathcal{E}^*(\epsilon^2; \Phi_{W(T_1+s)}, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{W_*}, f^*) &\lesssim e^{-Cs}, \quad \forall s \in [0, T_2 - T_1], \\ \mathcal{E}^*(\epsilon^2; \Phi_{W(T_2)}, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{W_*}, f^*) &\lesssim p \left(\frac{dp}{n} \right)^{\min(\alpha, 1)} \text{polylog}(n, d, p). \end{aligned}$$

本文在下文中讨论上述结果的一些关键方面。

改进特征误差度量 定理 6.3 和定理 6.4 表明, 多指标模型中的自适应特征方法具有与单指标模型相似的行为。特征误差度量表现出多个学习阶段, 初始阶段误差恒定, 后续阶段表现出指数衰减。最终的超额特征误差随 $p(dp/n)^{\min(\alpha, 1)}$ (在对数因子内) 缩放, 额外的 p 因子对应于方向的维度。然而, 注意到关于 d 的样本复杂度比单指标模型略大 1, 这是由于证明中的技术原因。总体而言, 在多指标模型下, 自适应特征模型也能够同时学习方向和链接函数, 这是自适应特征潜力的又一个说明性例子。

证明思路 以下简要讨论证明思路。详细证明因技术性较高且较为复杂, 推迟至第 6.5 节。挑战在于分析 W 的矩阵值动力学、它与泛函系数 β 的相互作用以及噪声项。首先, 本章引入角度矩阵 $P = W^\top W_*$ 并考虑奇异值分解 (SVD) $P = U \Sigma V^\top$ 。 W 和 W_* 之间的对齐程度可以通过 Σ 与单位矩阵的接近程度来衡量。通过关注 Σ , 本文将复杂的矩阵值动力学简化为逐分量的标量动力学。然而, 由于 SVD 的非唯一性, 这些逐分量动力学取决于正交矩阵 U, V 的选择, 从而导致无法统一控制的噪声项。为了解决这个问题, 本文引入了与 SVD 无关的对称量 (例如 $\text{Tr } \Sigma^2$)。一个特殊的量是对于某些 $K > 0$ 的 $\omega = -\log(e^{-K\Sigma^2})/K$, 它是最小平方奇异值的一个平滑近似。利用这个量, 本文应用多阶段分析来展示 Σ 的增加, 同时提供对噪声项的统一控制。最后, 特征误差度量也可以通过 Σ 来控制。本文相信, 本章的证明技术可以应用于噪声观测下的其他矩阵值模型, 这可能具有独立的研究价值。

与文献对比 本段将结果与最相关的文献^[104]进行比较, 该文献也考虑了多指标模型的梯度训练。主要区别之一是本章考虑了序列模型下的有噪声设置, 而 Bietti et al.^[104]仅考虑了总体动力学。另一个主要区别是本章使用同步梯度下降 (6.17) 学习泛函系数 β , 而在 Bietti et al.^[104]中, β 在每个时间步被直接设置为插值解。他们的 β 更新方式不适用于有噪声设置, 因为这会导致过拟合噪声。尽管采用了不同的训练方案, 本章结果中的训练时间复杂度 d^{m_0-1} 仍与 Bietti et al.^[104]中的时间复杂度一致。这显示了多指标模型的本质特征, 并表明该自适应特征模型能够有效学习方向, 同时防止过拟合噪声。

6.4 单指标模型的证明

在接下来的分析中, 引入 $\rho = \langle w, w_* \rangle$ 作为 w 与 w_* 之间夹角的余弦值。此外, 参数化还给出

$$f(x) = \langle \beta, \Phi_w(x) \rangle_{\ell^2(\mathbb{N})} = \sum_{r \geq 0} \beta_r \lambda_r^{\frac{1}{2}} H_r(\langle w, x \rangle) = g(\langle w, x \rangle),$$

其中函数 g 定义为

$$g(u) = \sum_{r \geq 0} g_r H_r(u), \quad g_r = \lambda_r^{\frac{1}{2}} \beta_r.$$

回顾特征值取为 $\lambda_r \asymp e^{-\gamma r}$ 。利用推论 6.5, 有

$$f_m = \langle f, H_m \rangle_{\gamma_d} = \left\langle \sum_{r \geq 0} g_r H_r(\langle w, \cdot \rangle), H_m \right\rangle_{\gamma_d} = \binom{r}{m}^{\frac{1}{2}} w^m g_r, \quad r = |m|. \quad (6.21)$$

关于球面上的投影, 引入投影算子 $P_w^\perp$, 对于 $w \in \mathbb{S}^{d-1}$, 其定义为 $P_w^\perp v = v - \langle v, w \rangle w$ 。

6.4.1 特征误差度量的基本性质

对于单指标模型, 特征误差度量可以被显式地计算。首先, 不难看出

$$\mathcal{E}(\delta, \epsilon^2; \Phi_w, f^*) - \mathcal{E}(\delta, \epsilon^2; \Phi_{w_*}, f^*) = \mathcal{E}_{\text{Proj}}(\Phi_w, f^*). \quad (6.22)$$

为了计算投影, 回顾

$$f^*(x) = g^*(\langle w_*, x \rangle) = \sum_{r \geq 0} g_r^* H_r(\langle w_*, x \rangle),$$

定义

$$f_{r,w}^* := \langle f^*, H_r(\langle w, \cdot \rangle) \rangle_{\gamma_d} = \left\langle \sum_{s \geq 0} g_s^* H_s(\langle w_*, x \rangle), H_r(\langle w, \cdot \rangle) \right\rangle_{\gamma_d} \quad (6.23)$$

$$= \langle w, w_* \rangle^r g_r^* = \rho^r g_r^*,$$

其中在第二个等号处应用了推论 6.5。因此，有

$$\mathcal{E}_{\text{Proj}}(\Phi_w, f^*) = \sum_{r \geq 0} (1 - \rho^{2r})(g_r^*)^2. \quad (6.24)$$

由于 $\rho \in [-1, 1]$, $\mathcal{E}_{\text{Proj}}(\Phi_w, f^*) = 0$ 当且仅当 $w = \pm w_*$ 。另一方面，计算得到

$$\begin{aligned} \mathcal{E}_{\text{Stat}}(\delta, \epsilon^2; \Phi_w, f^*) &= \#\{r \geq 0 : \lambda_r \geq \delta\} \epsilon^2 + \sum_{r \geq 0} (f_{r,w}^*)^2 \mathbf{1}\{\lambda_r < \delta\} \\ &= \#\{r \geq 0 : \lambda_r \geq \delta\} \epsilon^2 + \sum_{r \geq 0} \rho^{2r} (g_r^*)^2 \mathbf{1}\{\lambda_r < \delta\}. \end{aligned} \quad (6.25)$$

因此，

$$\begin{aligned} \mathcal{E}(\delta, \epsilon^2; \Phi_w, f^*) &= \mathcal{E}_{\text{Proj}}(\Phi_w, f^*) + \mathcal{E}_{\text{Stat}}(\delta, \epsilon^2; \Phi_w, f^*) \\ &= \sum_{r \geq 0} (1 - \rho^{2r})(g_r^*)^2 + \#\{r \geq 0 : \lambda_r \geq \delta\} \epsilon^2 + \sum_{r \geq 0} \rho^{2r} (g_r^*)^2 \mathbf{1}\{\lambda_r < \delta\} \\ &= \sum_{r \geq 0} [\epsilon^2 + (1 - \rho^{2r})(g_r^*)^2] \mathbf{1}\{\lambda_r \geq \delta\} + \sum_{r \geq 0} (g_r^*)^2 \mathbf{1}\{\lambda_r < \delta\}. \end{aligned} \quad (6.26)$$

命题 6.1： 在假设 6.2 下，有

$$\mathcal{E}(\delta, \epsilon^2; \Phi_w, f^*) - \mathcal{E}(\delta, \epsilon^2; \Phi_{w_*}, f^*) = \mathcal{E}_{\text{Proj}}(\Phi_w, f^*) \lesssim \begin{cases} 1 - \rho, & \alpha > 1, \\ (1 - \rho) \log(1 - \rho)^{-1}, & \alpha = 1, \\ (1 - \rho)^\alpha, & \alpha \in (0, 1). \end{cases} \quad (6.27)$$

证明 不失一般性，考虑 $\rho > 0$ 。利用初等不等式，有 $1 - \rho^{2r} \leq 2r(1 - \rho)$ ，因此

$$\mathcal{E}_{\text{Proj}}(\Phi_w, f^*) = \sum_{r \geq 0} (1 - \rho^{2r})(g_r^*)^2 \leq \sum_{r \geq 0} \min(1, 2r(1 - \rho))(g_r^*)^2 = 2(1 - \rho) \sum_{r \leq L} r(g_r^*)^2 + \sum_{r > L} (g_r^*)^2,$$

其中 $L = \frac{1}{2(1 - \rho)}$ 。由于 $g_r^* \lesssim r^{-\frac{\alpha+1}{2}}$ ，有 $\sum_{r > L} (g_r^*)^2 \lesssim L^{-\alpha}$ 。同时，有

$$\sum_{r \leq L} r(g_r^*)^2 \lesssim \sum_{r \leq L} r^{-\alpha} \lesssim \begin{cases} 1, & \alpha > 1, \\ \log L, & \alpha = 1, \\ L^{1-\alpha}, & \alpha \in (0, 1). \end{cases}$$

结合以上不等式就可以得出结论。 ■

6.4.1.1 初始化

以下命题展示了 ρ 的初始化情况。

命题 6.2: 令 $w \sim \text{Unif}(\mathbb{S}^{d-1})$ 且 $w_* \in \mathbb{S}^{d-1}$ 固定。那么, 存在一个依赖于 d 的绝对常数 $c > 0$ 使得

$$\mathbb{P} \left\{ |\langle w, w_* \rangle| \geq \frac{t}{\sqrt{d}} \right\} \geq 1 - ct, \quad \forall t > 0. \quad (6.28)$$

此外, 还有

$$\mathbb{P} \left\{ |\langle w, w_* \rangle| \leq \frac{1}{2} \right\} \geq 1 - 2e^{-cd}, \quad \forall d \geq 1. \quad (6.29)$$

证明 利用 $\langle w, w_* \rangle$ 的显式密度函数, (6.28) 的证明非常直接。参见例如 Bietti et al.^[107] 中的引理 B.7。对于 (6.29), 可以使用球面均匀分布的次高斯集中不等式^[102]。 ■

6.4.2 总体动力学

在本小节中, 考虑单指标模型下自适应核模型的总体动力学。用 $\mathcal{L} = \frac{1}{2} \mathbb{E} (f(x) - f^*(x))^2$ 表示总体损失。考虑以下方程, 它是 (6.10) 的总体版本:

$$\begin{cases} \dot{\beta}_r = -\nabla_{\beta_r} \mathcal{L}, & \beta_r(0) = 0, \quad r \geq 0, \\ \dot{w} = -\nabla_w^{\mathbb{S}^{d-1}} \mathcal{L}, & w(0) \sim \text{Unif}(\mathbb{S}^{d-1}). \end{cases} \quad (6.30)$$

利用 (6.105), 总体损失可以计算为

$$\begin{aligned} \mathcal{L} &= \frac{1}{2} \mathbb{E} [f^*(x) - f(x)]^2 = \frac{1}{2} \mathbb{E} [g^*(\langle w_*, x \rangle) - g(\langle w, x \rangle)]^2 \\ &= \frac{1}{2} \mathbb{E} \left[\sum_{r \geq 0} (g_r^* H_r(\langle w_*, x \rangle) - g_r H_r(\langle w, x \rangle)) \right]^2 \\ &= \frac{1}{2} \sum_{r \geq 0} [(g_r^*)^2 + (g_r)^2 - 2g_r^* g_r \rho^r]. \end{aligned}$$

因此, 有

$$\nabla_{\beta_r} \mathcal{L} = -\lambda_r^{\frac{1}{2}} (\rho^r g_r^* - g_r), \quad (6.31)$$

所以 $\dot{\beta}_r = -\lambda_r^{\frac{1}{2}} (\rho^r g_r^* - g_r)$ 。同时, 有

$$\nabla_w \mathcal{L} = -\sum_{r \geq 1} r g_r^* g_r \rho^{r-1} w_*.$$

取球面上的投影，有

$$-\nabla_w^{\mathbb{S}^{d-1}} \mathcal{L} = -P_w^\perp \nabla_w \mathcal{L} = \left(\sum_{r \geq 1} r g_r^* g_r \rho^{r-1} \right) P_w^\perp w_*. \quad (6.32)$$

进一步计算 ρ 的动力学。有

$$\dot{\rho} = \langle \dot{w}, w_* \rangle = \left(\sum_{r \geq 1} r g_r^* g_r \rho^{r-1} \right) \langle P_w^\perp w_*, w_* \rangle = \sum_{r \geq 1} r g_r^* g_r \rho^{r-1} (1 - \rho^2), \quad (6.33)$$

其中注意到

$$\langle P_w^\perp w_*, w_* \rangle = \langle w_* - \langle w, w_* \rangle w, w_* \rangle = 1 - \rho^2. \quad (6.34)$$

在下文中汇总由 (6.30) 导出的 $g_r = \lambda_r^{\frac{1}{2}} \beta_r$ 和 ρ 的诱导动力学：

$$\begin{cases} \dot{g}_r = \lambda_r (\rho^r g_r^* - g_r), & r \geq 0, \\ \dot{\rho} = \sum_{r \geq 1} r g_r^* g_r \rho^{r-1} (1 - \rho^2). \end{cases} \quad (6.35)$$

以下命题展示了总体动力学的基本性质。

命题 6.3: 考虑总体动力学 (6.30)。假设 $\rho(0) \neq 0$ 。那么，

$$g_r^* g_r(t) \geq 0, \quad \forall r \geq 0, \quad \forall t \geq 0. \quad (6.36)$$

此外，对于所有 $t \geq 0$ ，若 $\rho(0) > 0$ 则 $\dot{\rho}(t) \geq 0$ ；若 $\rho(0) < 0$ 则 $\dot{\rho}(t) \leq 0$ 。

证明 这是动力学方程 (6.35) 的直接结果。这里只给出证明思路，严格证明可通过标准的 ODE 连续性论证得到。 g_r 的动力学表明 $g_r(t)$ 与 $\rho^r g_r^*$ 具有相同的符号。因此， $\dot{\rho}$ 中的每一项 $\rho^{r-1} g_r^* g_r(t) = \rho^{-1} (\rho^r g_r^*) g_r(t)$ 都与 ρ 具有相同的符号。因此， $\dot{\rho}$ 与 ρ 同号，结论得证。 ■

根据命题 6.3，在随后的分析中，可以不失一般性地假设 $g_r^* \geq 0$ 并考虑 $\rho(0) > 0$ 。

命题 6.4: 考虑总体动力学 (6.30)。令假设 6.1 成立。假设 $\rho(0) = \rho_0 \neq 0$ 。那么，对于所有 $t \geq T^{\text{app}}$ ，

$$|\rho(t)| \geq \frac{1}{2}, \quad |g_{r_0}(t)| \geq 2^{-(r_0+1)} |g_{r_0}^*|, \quad (6.37)$$

其中

$$T^{\text{app}} \lesssim \log \rho_0^{-1} + \rho_0^{-2(r_0-1)}. \quad (6.38)$$

证明 从 (6.37) 中的第一个条件开始。利用命题 6.3，只需考虑 $\rho_0 > 0$ 并界定 $\rho(t) \geq \frac{1}{2}$ 的首个时刻。此外，由于 $\dot{\rho}$ 中的每一项都是非负的，且当 $\rho < 1/2$ 时 $1 - \rho^2 \geq 3/4$ ，有

$$\dot{\rho} \geq \frac{3r_0}{4} g_{r_0}^* \rho^{r_0-1} g_{r_0} = c r_0 g_{r_0}^* \rho^{r_0-1} g_{r_0}. \quad (6.39)$$

同时, 回顾

$$\dot{g}_{r_0} = \lambda_{r_0}(\rho^{r_0} g_{r_0}^* - g_{r_0}).$$

取 $L := 1 + \lceil \log_2 \rho_0^{-1} \rceil$ 。对于 $k < L$ 定义 $\rho_k = 2^k \rho_0$, 且 $\rho_L = \frac{1}{2}$ 。还引入以下时间点:

$$T_0^\rho = 0, \quad T_k^g = \inf \left\{ t \geq T_k^\rho : g_{r_0}(t) \geq \frac{1}{2} \rho_k^{r_0} g_{r_0}^* \right\}, \quad T_k^\rho = \inf \left\{ t \geq T_{k-1}^g : \rho(t) \geq \rho_k \right\}. \quad (6.40)$$

现在, 为了界定 T_k^g , 有

$$\dot{g}_{r_0} = \lambda_{r_0}(\rho^{r_0} g_{r_0}^* - g_{r_0}) \geq \lambda_{r_0} \left(\rho_k^{r_0} g_{r_0}^* - g_{r_0} \right) \geq \frac{1}{2} \lambda_{r_0} \rho_k^{r_0} g_{r_0}^*, \quad t \in [T_k^\rho, T_k^g],$$

所以

$$T_k^g - T_k^\rho \leq \frac{\frac{1}{2} \rho_k^{r_0} g_{r_0}^*}{\frac{1}{2} \lambda_{r_0} \rho_k^{r_0} g_{r_0}^*} = \frac{1}{\lambda_{r_0}}.$$

另一方面, 对于 T_k^ρ , 有

$$\dot{\rho} \geq \frac{r_0}{2} g_{r_0}^* \rho^{r_0-1} g_{r_0} \geq \frac{r_0}{2} g_{r_0}^* \cdot \rho_{k-1}^{r_0-1} \cdot \frac{1}{2} \rho_k^{r_0} g_{r_0}^* = \frac{r_0}{2^{r_0+1}} (g_{r_0}^*)^2 \rho_k^{2r_0-1}, \quad t \in [T_{k-1}^g, T_k^\rho],$$

所以

$$T_k^\rho - T_{k-1}^g \leq \frac{\rho_k - \rho_{k-1}}{\frac{r_0}{2^{r_0+1}} (g_{r_0}^*)^2 \rho_k^{2r_0-1}} = \rho_0^{-2(r_0-1)} 2^{r_0} 2^{-2(r_0-1)k} r_0^{-1} (g_{r_0}^*)^{-2}.$$

因此,

$$\begin{aligned} T_L^\rho &\leq \sum_{k \leq L} [(T_k^\rho - T_{k-1}^g) + (T_{k-1}^g - T_{k-1}^\rho)] \\ &\leq \sum_{k \leq L} \left[\lambda_{r_0}^{-1} + \rho_0^{-2(r_0-1)} 2^{r_0} 2^{-2(r_0-1)k} r_0^{-1} (g_{r_0}^*)^{-2} \right] \\ &= L \lambda_{r_0}^{-1} + \rho_0^{-2(r_0-1)} 2^{r_0} r_0^{-1} (g_{r_0}^*)^{-2} \sum_{k \leq L} 2^{-2(r_0-1)k} \\ &\leq L \lambda_{r_0}^{-1} + \rho_0^{-2(r_0-1)} 2^{r_0} r_0^{-1} (g_{r_0}^*)^{-2} (1 - 2^{-2(r_0-1)})^{-1} \\ &\leq L \lambda_{r_0}^{-1} + \rho_0^{-2(r_0-1)} 2^{r_0+1} r_0^{-1} (g_{r_0}^*)^{-2} \\ &\lesssim \lambda_{r_0}^{-1} \log \rho_0^{-1} + (g_{r_0}^*)^{-2} \rho_0^{-2(r_0-1)}, \end{aligned}$$

且 $T_L^g \leq T_L^\rho + \lambda_{r_0}^{-1}$ 。 ■

命题 6.5: 考虑总体动力学 (6.30)。令假设 6.1 成立。假设 (6.37) 在某个 t_0 时刻成立。那么,

$$1 - |\rho(t_0 + t)| \leq \frac{1}{2} e^{-r_0 2^{-2r_0} (g_{r_0}^*)^2 t}. \quad (6.41)$$

证明 不失一般性, 假设 $\rho(t_0) > 0$ 且 $g_{r_0}^* > 0$. 由于单调性, 有

$$\dot{\rho} \geq r_0 g_{r_0}^* g_{r_0} \rho^{r_0-1} (1 - \rho^2) \geq r_0 g_{r_0}^* g_{r_0} \rho^{r_0-1} (1 - \rho) \geq r_0 2^{-2r_0} (g_{r_0}^*)^2 (1 - \rho),$$

故结论成立. ■

定理 6.1 的证明 特征误差度量的单调性由 (6.26) 以及命题 6.3 中 $|\rho|$ 的单调性得出. 对于初始化, 根据命题 6.2, 以高概率成立:

$$\frac{C}{\sqrt{d}} \leq |\rho(0)| \leq \frac{1}{2}.$$

因此, 有

$$\mathcal{E}_{\text{Proj}}(\Phi_{w(0)}, f^*) = \sum_{r \geq 0} (1 - \rho(0)^{2r}) (g_r^*)^2 \gtrsim \sum_{r \geq 1} (g_r^*)^2 \gtrsim 1.$$

另一方面, 由于 $\rho(0) \geq C/\sqrt{d}$, 命题 6.4 和命题 6.5 表明对于某个 $t_0 \asymp \log d + d^{r_0-1}$, 有

$$1 - |\rho(t_0 + t)| \leq \frac{1}{2} e^{-Ct}.$$

因此, 通过应用命题 6.1 并调整常数, 结论得证. ■

6.4.3 序列模型

在本小节考虑 (6.10) 中的自适应特征动力学. 利用 (6.43) 和 (6.44) 中动力学关于负的 g_r^* 和负的 $\rho(0)$ 的对称性, 在随后的部分中, 本文将不失一般性地假设 $g_r^* \geq 0$ 且 $\rho(0) > 0$. 此外, 本文假设假设 6.1 成立而不再明确提及.

6.4.3.1 计算动力学

首先计算噪声序列模型下的动力学. 本文将结合总体情况下的计算和 (6.110) 来简化计算过程.

β_r 项 回顾 (6.21), 有

$$\nabla_{\beta_r} f_m = \lambda_r^{\frac{1}{2}} \binom{r}{m}^{\frac{1}{2}} w^m \delta_{r,|m|}.$$

定义

$$e_r = e_r(w) = \sum_{|m|=r} \binom{r}{m}^{\frac{1}{2}} w^m \epsilon_m. \quad (6.42)$$

那么, 结合 (6.31), (6.110) 表明

$$-\nabla_{\beta_r} \hat{\mathcal{L}} = \lambda_r^{\frac{1}{2}} (\rho^r g_r^* - g_r + e_r),$$

由此得出

$$\dot{g}_r = \lambda_r^{\frac{1}{2}} \dot{\theta}_r = \lambda_r (\rho^r g_r^* - g_r + e_r). \quad (6.43)$$

w 项 对于 w 项, 有

$$\nabla_w^{\mathbb{S}^{d-1}} f_m = g_r \binom{r}{m}^{\frac{1}{2}} \nabla_w^{\mathbb{S}^{d-1}} (w^m).$$

因此, 利用 (6.32) 和 (6.110), 有

$$\dot{w} = \left(\sum_{r \geq 1} r g_r^* g_r \rho^{r-1} \right) P_w^\perp w_* + E, \quad E = \sum_{r \geq 0} g_r \sum_{|m|=r} \binom{r}{m}^{\frac{1}{2}} \nabla_w^{\mathbb{S}^{d-1}} (w^m) \varepsilon_m.$$

此外,

$$\begin{aligned} \dot{\rho} &= \langle \dot{w}, w_* \rangle = \left[\sum_{r \geq 1} r g_r^* g_r \rho^{r-1} \right] \langle P_w^\perp w_*, w_* \rangle + \langle E, w_* \rangle \\ &= \sum_{r \geq 1} r g_r^* g_r \rho^{r-1} (1 - \rho^2) + \tau, \quad \tau = \langle E, w_* \rangle. \end{aligned} \quad (6.44)$$

6.4.3.2 界定扰动项

此处将通过计算扰动项 e_r 和 τ 的协方差并使用高斯过程的一致上界, 分别控制 (6.43) 和 (6.44) 中的这两项。注意到 e_r 和 τ 取决于参数 w 和 β , 因此有必要在参数空间上对它们给出一致的上界。

命题 6.6: 令 e_r 定义如 (6.42)。那么, $\text{Cov}(e_r(u), e_r(v)) = \langle u, v \rangle^r / n$ 。因此, 以至少 $1 - 4e^{-d}$ 的概率, 有

$$\sup_{w \in \mathbb{S}^{d-1}} |e_r(w)| \lesssim \sqrt{\frac{d \log r}{n}}, \quad \forall r \geq 0. \quad (6.45)$$

证明 首先计算协方差函数。利用二项式定理, 有

$$\text{Cov}(e_r(u), e_r(v)) = \frac{1}{n} \sum_{|m|=r} \binom{r}{m} u^m v^m = \frac{1}{n} \langle u, v \rangle^r.$$

对于高概率上界, 应用引理 6.4 并注意到 u^m 关于 u 是 Lipschitz 连续的, 其 Lipschitz 常数为 $|m|$, 其中对 r 使用了联合概率界, 使得不等式对所有 r 同时成立。 ■

命题 6.7: 令 τ 定义如 (6.44)。那么, 以至少 $1 - Ce^{-d}$ 的概率,

$$|\tau(w)| \lesssim \sqrt{\frac{d}{n}} \sum_{r \geq 0} r^{\frac{1}{2}} (\log r)^{\frac{1}{2}} |g_r|, \quad \forall w \in \mathbb{S}^{d-1}. \quad (6.46)$$

此外, 对于任何固定的 $v \in (0, 1)$, 以至少 $1 - Ce^{-d}$ 的概率,

$$|\tau(w)| \lesssim \sqrt{1 - \rho^2} \sqrt{\frac{d \log v^{-1}}{n}} \sum_{r \geq 0} r^{\frac{1}{2}} (\log r)^{\frac{1}{2}} |g_r|, \quad \forall w \in \mathbb{S}^{d-1} \setminus B_v(w_*). \quad (6.47)$$

证明 首先, 可以等价地表达

$$\begin{aligned} \tau &= (w_*)^\top \sum_{r \geq 0} g_r \sum_{|m|=r} \binom{r}{m}^{\frac{1}{2}} P_w^\perp \nabla_w (w^m) \epsilon_m \\ &= \sum_{r \geq 0} g_r \tau_r, \end{aligned}$$

其中

$$\tau_r = \tau_r(w) = (P_w^\perp w_*)^\top \sum_{|m|=r} \binom{r}{m}^{\frac{1}{2}} \nabla_w^{\mathbb{S}^{d-1}} (w^m) \epsilon_m.$$

为了计算 τ_r 的协方差, 令 $a = P_u^\perp w_*$ 且 $b = P_v^\perp w_*$ 。那么,

$$\begin{aligned} \kappa(u, v) &:= n \text{Cov}(\tau_r(u), \tau_r(v)) = \sum_{i,j} a_i b_j \nabla_{u_i} \nabla_{v_j} \langle u, v \rangle^r \\ &= r(r-1) \langle u, v \rangle^{r-2} \langle P_u^\perp w_*, v \rangle \langle P_v^\perp w_*, u \rangle + r \langle u, v \rangle^{r-1} \langle P_u^\perp w_*, P_v^\perp w_* \rangle. \end{aligned}$$

引入

$$\rho_u = \langle u, w_* \rangle, \quad \rho_v = \langle v, w_* \rangle, \quad q = \langle u, v \rangle,$$

并注意到

$$P_u^\perp w_* = w_* - \rho_u u, \quad P_v^\perp w_* = w_* - \rho_v v,$$

有

$$\kappa(u, v) = r(r-1)q^{r-2}(\rho_v - q\rho_u)(\rho_u - q\rho_v) + rq^{r-1}(1 + q\rho_u\rho_v - \rho_u^2 - \rho_v^2).$$

当 $u = v$ 时, 由于 $\rho_v = \rho_u$ 且 $q = 1$, 第一项消失, 因此

$$\kappa(u, u) = r(1 - \rho^2).$$

现在, 界定 $\kappa(u, v)$ 的导数。这里将频繁使用以下事实:

$$\|\nabla(fg)\| = \|f\nabla g + g\nabla f\| \leq |f| \|\nabla g\| + |g| \|\nabla f\|.$$

结合各分量的控制, 有

$$\nabla_u \kappa_r(u, v) \leq C(r-1)(r-2),$$

其中 C 是一个绝对常数。因此, 可以对 $\sqrt{n}\tau_r$ 应用引理 6.4, 得到以至少 $1 - Cr^{-2}e^{-d}$

的概率:

$$\sup_{w \in \mathbb{S}^{d-1}} |\tau_r(w)| \lesssim \sqrt{d(\log r)/n}.$$

求和后, 有

$$\sup_{w \in \mathbb{S}^{d-1}} |\tau(w)| \lesssim n^{-\frac{1}{2}} \sum_{r \geq 0} \sqrt{rd \log r} |g_r| = \sqrt{d/n} \sum_{r \geq 0} r^{\frac{1}{2}} (\log r)^{\frac{1}{2}} |g_r|.$$

此外, 引入缩放版本

$$\bar{\tau}_r(w) = \sqrt{n} [r(1 - \rho_w^2)]^{-\frac{1}{2}} \tau_r(w),$$

使得 $\text{Var}(\bar{\tau}_r(w)) = 1$ 。由此, 可以对 $\bar{\tau}_r$ 应用引理 6.4 来获得高概率上界。以至少 $1 - Cr^{-2}e^{-d}$ 的概率, 有

$$\sup_{w \in \mathbb{S}^{d-1} \setminus B_\delta(w_*)} |\bar{\tau}_r(w)| \lesssim \sqrt{d(\log r + \log \delta^{-1})}.$$

回到 τ_r 并求和就得出第二个上界。 ■

命题 6.8: 对于任何 $r \geq 0$, 有

$$|g_r(t)| \leq \left(\sup_{s \leq t} |\rho(s)|^r |g_r^*| + |e_r| \right) \min(1, \lambda_r t). \quad (6.48)$$

此外, 若 $g_r^* \geq 0$ 且对于 $s \in [0, t]$ 满足 $\rho(s) \geq 0$, 那么

$$g_r(t) \geq -|e_r| \min(1, \lambda_r t). \quad (6.49)$$

证明 第一部分由 (6.43) 中的动力学和比较定理得出。对于第二部分, 只需注意到

$$\lambda_r(\rho^r g_r^* - g_r + e_r) \geq \lambda_r(-|e_r| - g_r)$$

并应用比较定理。 ■

推论 6.2: 令 (6.45) 成立且 $n \gtrsim d$ 。那么, 有

$$|g_r(t)| \lesssim \left(|g_r^*| + \sqrt{\frac{d \log r}{n}} \right) \min(1, \lambda_r t). \quad (6.50)$$

因此, 命题 6.7 中的求和可以界定为

$$\sum_{r \geq 0} r^{\frac{1}{2}} (\log r)^{\frac{1}{2}} |g_r| \lesssim 1 + (\log^+ t)^2. \quad (6.51)$$

证明 只需证明第二部分, 其中可以应用命题 6.31。 ■

6.4.3.3 初始化附近的训练动力学

本小节将考虑初始化附近的训练动力学。

命题 6.9: 在假设 6.1 下, 假设对于某些 $s > 0$, $n \gtrsim d^{2r_0+s}$ 。令 (6.45) 和 (6.46) 成

立。令 $\delta > 0$ 固定。那么，以至少 $1 - \delta$ 的概率，当 n, d 足够大时，有

$$\inf \{t \geq 0 : |\rho(t)| \geq 1/2\} = T^{\text{app}} \lesssim \log d + d^{r_0-1}, \quad (6.52)$$

其中 $\lesssim$ 符号中的常数可以取决于 δ 。此外，在经过常数时间后，当 $t \leq T^{\text{app}}$ 时， $|\rho(t)|$ 是单调递增的。

证明 首先，根据命题 6.2，以至少 $1 - \delta$ 的概率，有 $|\rho(0)| \geq cd^{-1/2}$ （对于某些 $c = c(\delta) > 0$ ）。同时，由动力学的对称性，可以不失一般性地考虑 $\rho(0) > 0$ 。此外，只需要关注 $\rho(t) \leq 1/2$ ，且不再明确提及。取 $c_0 = c/2$ ，此处断言在所关注的时间范围内将有 $\rho(t) \geq \rho_0 := c_0 d^{-1/2}$ ，将在稍后证明这一断言。

回顾分量 r_0 的动力学：

$$\dot{g}_{r_0} = \lambda_{r_0} \left(\rho^{r_0} g_{r_0}^* - g_{r_0} + e_{r_0} \right).$$

因此，当 $\rho(t) \geq c_0 d^{-1/2}$ 时，由于 $n = \Omega(d^{r_0+1+s})$ ，当 n 足够大时有

$$\rho^{r_0} g_{r_0}^*/2 \geq C c_0^{\frac{r_0}{2}} d^{-\frac{r_0}{2}} \geq C \sqrt{\frac{d}{n}} \geq |e_{r_0}|,$$

这意味着

$$\dot{g}_{r_0} \geq \lambda_{r_0} (\rho_0^{r_0} g_{r_0}^*/2 - g_{r_0}). \quad (6.53)$$

因此， $g_{r_0}(t)$ 是单调递增的，且有

$$T_0^g := \inf \left\{ t \geq 0 : g_{r_0}(t) \geq \frac{1}{4} \rho_0^{r_0} g_{r_0}^* \right\} \leq \frac{1}{\lambda_{r_0}} \lesssim 1.$$

现在证明 $t \leq T_0^g$ 时的断言。引入

$$S_0 = \left\{ r \geq 1 : (c_0 d^{-1/2})^r |g_r^*| \geq C \sqrt{d(\log r)/n} \geq |e_r| \right\}$$

并写成

$$\dot{\rho} = \sum_{r \in S_0} r \rho^{r-1} g_r^* g_r (1 - \rho^2) + \sum_{r \notin S_0} r \rho^{r-1} g_r^* g_r (1 - \rho^2) + \tau = P_0 + P_1 + \tau.$$

对于每个 $r \in S_0$ ，(6.43) 表明 $g_r \geq 0$ ，因此 $P_0 \geq 0$ 。另一方面，对于 $r \notin S_0$ ，命题 6.8 给出

$$g_r(t) \geq -C \min(1, \lambda_r t) |e_r| \geq -C \min(1, \lambda_r t) \sqrt{d(\log r)/n}.$$

因此，利用 $\lambda_r \lesssim e^{-\gamma r}$ 和命题 6.31，有 $P_1 \geq -C \sqrt{d/n}$ 。此外，(6.46) 和 (6.51) 给出 $|\tau(t)| \lesssim \sqrt{d/n}(1 + (\log^+ t)^2)$ 。因此，在 $t \leq T_0^g$ 时 $\dot{\rho}(t) \geq -C \sqrt{d/n}$ ，由此得出

$$\rho(T_0^g) \geq \rho(0) - C \sqrt{d/n} T_0^g \geq c_0 d^{-1/2} = \rho_0,$$

因为 $n \gtrsim d^{2+s}$ 。

当 $t \geq T_0^g$ 时, 有 $P_0 \gtrsim \rho_0^{2r_0-1}$ 。将其与 $P_1, |\tau|$ 的估计结合并利用 $n \gtrsim d^{2r_0+s}$, 只要 t 关于 d, n 是多项式级的, 有

$$\rho_0^{2r_0-1} \gtrsim \sqrt{d/n}(1 + (\log^+ t)^2) \gtrsim -P_1 + |\tau(t)|,$$

从而

$$\dot{\rho}(t) \geq c\rho^{r_0-1}g_{r_0}^*g_{r_0}. \quad (6.54)$$

利用 (6.53) 和 (6.54), 可以遵循与命题 6.4 证明中相同的论证, 并推导出 $T_L^\rho \lesssim \log d + d^{r_0-1}$ 。 ■

6.4.3.4 收敛附近的训练动力学

命题 6.10: 在假设 6.1 下, 假设对于某些 $s > 0$, $n \gtrsim d^{1+s}$ 。令 (6.45)、(6.46) 和 (6.47) 在 $v = 1/n$ 时成立。假设 (6.52) 在某些时间 $t_0 \lesssim \text{poly}(d)$ 成立。那么, 存在某个 $t_1 \leq t_0 + C$ 使得

$$|\rho(t_1 + s)| \text{ 是单调递增的且 } 1 - |\rho(t_1 + s)| \lesssim e^{-cs},$$

前提是

$$1 - |\rho(t_1 + s)| \gtrsim \frac{d \text{polylog}(n, d)}{n}.$$

证明 首先证明对于所有 $t \geq t_1 = t_0 + C$, 有 $g_{r_0}(t) \geq cg_{r_0}^* \geq c$, 同时断言对于 $t \in [t_0, t_1]$, $\rho(t) \geq 1/4$ 。回顾 $g_{r_0}^*$ 的动力学, 由于 $\rho(t) \geq 1/4$ 且 e_{r_0} 由 (6.45) 界定, 有

$$\dot{g}_{r_0} \geq \lambda_{r_0} (cg_{r_0}^* - g_{r_0}),$$

因此 g_{r_0} 是单调递增的, 且对于 $t_1 \leq t_0 + \lambda_{r_0}^{-1}$, 有 $g_{r_0}(t_1) \geq \frac{c}{2}g_{r_0}^*$ 。现在通过下界界定 $\rho(t)$ 来证明该断言, 这与命题 6.9 证明中相应部分的论证类似。

现在证明直到 $\rho(t) = 1 - d \text{polylog}(n, d)/n$ 的收敛性。在 $t \geq t_1$ 之后, 利用 $v = 1/n$ 时的 (6.47) 和 (6.51) 得到

$$\dot{\rho} \geq c(1 - \rho^2) - (1 - \rho^2)^{\frac{1}{2}} \sqrt{\frac{d \log n}{n}} (1 + (\log^+ t)^2).$$

因此, 只要 t 关于 n, d 是多项式级的且

$$\rho(t) \leq 1 - C \frac{d \text{polylog}(n, d)}{n},$$

有

$$\dot{\rho} \geq c(1 - \rho^2) \geq c(1 - \rho) \implies \rho(t_1 + s) \geq 1 - \frac{3}{4}e^{-cs}.$$

因此, 只需额外花费 $\log n$ 的时间, $\rho(t)$ 即可增加到 $1 - d \text{polylog}(n, d)/n$ 。 ■

6.4.3.5 定理 6.2 的证明

首先, 应用 $v = 1/n$ 时的命题 6.6 和命题 6.7, 使得这些估计以至少 $1 - Ce^{-d}$ 的概率成立。然后, 应用命题 6.9 进行初始化 (取 $T_1 = T^{\text{app}}$), 并应用命题 6.10 进行收敛。 $\mathcal{E}(\delta, \epsilon^2; \Phi_w, f^*)$ 的单调性由 (6.26) 以及在这两个命题中推导出的 ρ 的单调性得出。这里, 注意到如果 $\rho(t)$ 进入了命题 6.9 中的单调递增阶段, 它将持续增加直到命题 6.10 中的收敛。最后, $\mathcal{E}^*(\epsilon^2; \Phi_{w(T_1+s)}, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{w_*}, f^*)$ 和 $\mathcal{E}^*(\epsilon^2; \Phi_{w(T_2)}, f^*) - \mathcal{E}^*(\epsilon^2; \Phi_{w_*}, f^*)$ 的上界来自 $1 - |\rho(t)|$ 的上界和命题 6.1。

6.5 多指标模型的证明

先回顾多指标模型:

$$f^*(x) = g^*(W_*^\top x), \quad W_* \in \text{St}(d, p^*),$$

以及参数化方式

$$\begin{aligned} \Phi_W(x) &= \left(\lambda_m^{\frac{1}{2}} H_m(W^\top x) \right)_{m \in \mathbb{N}^p}, \quad W \in \text{St}(d, p), \\ f(x) &= \langle \beta, \Phi_W(x) \rangle_{\ell^2(\mathbb{N}^p)} = \sum_{m \in \mathbb{N}^p} \beta_m \lambda_m^{\frac{1}{2}} H_m(W^\top x) = g(W^\top x), \\ g(u) &= \sum_{m \in \mathbb{N}^p} \beta_m \lambda_m^{\frac{1}{2}} H_m(u) \in L^2(\gamma_p), \end{aligned}$$

关于权重序列 λ_m , 回顾其取值为 $\lambda_m = \mu_{|m|} \asymp e^{-\gamma|m|}$, 其中 $\gamma > 0$ 为固定常数。

利用将在第 6.5.1.2 小节中引入的辅助算子, 可以将 f^* 写作 $P_{W_*} g^*$, 将 f 写作 $P_W g$ 。此外, 在假设 6.3 下, 本文将在第 6.5.1.6 小节中展示, 可以将系数表示为

$$g_{2r} = v_r h_r, \quad g_{2r}^* = v_r h_r^*, \quad r = |r|, r \in \mathbb{N}^p,$$

而其他系数均为零, 其中 v_r 是引理 6.3 中定义的一组系数。因此, 可以假设信息指数 (information index) $m_0 = 2r_0$, 其中 $r_0 > 0$ 。

此外, 记 $P = W^\top W_*$ 。考虑奇异值分解 (SVD) $P = U \Sigma V^\top$ 。在此汇总定义一些辅助量:

$$\phi_r = \sum_{|r|=r} v_r^2 \sigma^{2r}, \quad \phi_0 = 1, \quad \phi_1 = \frac{1}{p} \sum_{i=1}^p \sigma_i^2 \quad (6.55)$$

$$\omega = -\frac{1}{K} \log(\text{Tr } e^{-K \Sigma^2}) = -\frac{1}{K} \log\left(\sum_{i=1}^p e^{-K \sigma_i^2}\right) \leq \min(\sigma_1^2, \dots, \sigma_p^2),$$

其中 K 是稍后确定的常数。关于 ϕ_r 的定义请参考 (6.72)。

为了方便起见, 对于兼容的矩阵 (或向量) A, B , 使用 $\langle A, B \rangle_{\gamma_p} = \int A(x)B(x) d\gamma_p(x)$ 。同时, 将 $P_W, P_W^\top, \mathcal{A}_M$ (定义见第 6.5.1.2 小节) 通过逐元素应用的方式扩展到向量值函数。

6.5.1 预备知识

首先引入一些符号。对于矩阵 $A \in \mathbb{R}^{p \times p}$, 用 $\text{Diag}(A)$ 表示由 A 的对角线元素构成的对角矩阵, $\text{Sym}(A) = (A + A^\top)/2$ 表示 A 的对称部分。注意到 $\text{Diag}(A) = \text{Diag}(A^\top)$, 且对于对角矩阵 Λ , 有 $\text{Diag}(A\Lambda) = \text{Diag}(A)\Lambda = \Lambda \text{Diag}(A)$ 。

6.5.1.1 Stiefel 流形

Stiefel 流形 $S(d, p)$ 是所有列正交的实 $(d \times p)$ 矩阵的集合:

$$S(d, p) = \{X \in \mathbb{R}^{d \times p} \mid X^\top X = I_p\}.$$

众所周知, $S(d, p)$ 是一个黎曼流形, 其度量由 $\mathbb{R}^{d \times p}$ 上的欧几里得度量诱导。对于每个 $X \in S(d, p)$, 切空间 $T_X S(d, p)$ 给定为

$$T_X S(d, p) = \{Z \in \mathbb{R}^{d \times p} \mid X^\top Z + Z^\top X = 0\}.$$

此外, 可以利用欧几里得梯度计算定义在 $S(d, p)$ 上的函数的黎曼梯度。设 $F : S(d, p) \rightarrow \mathbb{R}$ 为光滑函数。记 $\nabla F(X)$ 为 F 在 X 处的普通 (欧几里得) 梯度, $\nabla^S F(X)$ 为 F 在 $S(d, p)$ 上的黎曼梯度。那么, $\nabla^S F(X)$ 即为 $\nabla F(X)$ 在切空间 $T_X S(d, p)$ 上的正交投影。该投影的一个表达式为

$$\nabla^S F(X) = \Pi_{T_X S}(\nabla F(X)), \quad \Pi_{T_W S}(Z) = Z - W \text{Sym}(W^\top Z), \quad (6.56)$$

其中 $\text{Sym}(A) = (A + A^\top)/2$ 表示矩阵的对称部分。

6.5.1.2 辅助算子

引入一些在多指标模型分析中有用的辅助算子^[104]。对于矩阵 $W \in S(d, p)$, 定义算子 $P_W : L^2(\gamma_p) \rightarrow L^2(\gamma_d)$ 为

$$f = P_W g, \quad f(x) = g(W^\top x). \quad (6.57)$$

由于当 $x \sim N(0, I_d)$ 时, $W^\top x \sim N(0, I_p)$, 因此 P_W 是等距算子, 即

$$\|P_W g\|_{\gamma_d}^2 = \mathbb{E}_{x \sim \gamma_d} g(W^\top x)^2 = \mathbb{E} g(y)^2 = \|g\|_{\gamma_p}^2.$$

因此, 可以定义其伴随算子 $P_W^\top : L^2(\gamma_d) \rightarrow L^2(\gamma_p)$:

$$\langle P_W g, h \rangle_{\gamma_d} = \langle g, P_W^\top h \rangle_{\gamma_p}.$$

由于 P_W 是等距的, $P_W^\top$ 是到空间 $L_{\Phi_W}^2 = \{f = g(W^\top x), g \in L^2(\gamma_p)\}$ 上的正交投影。此外, $P_W^\top P_W = I_{L^2(\gamma_p)}$, 且 $P_W P_W^\top$ 是 $L^2(\gamma_d)$ 中到空间 $L_{\Phi_W}^2$ 上的正交投影。

命题 6.11: 伴随算子 $P_W^\top : L^2(\gamma_d) \rightarrow L^2(\gamma_p)$ 的显式表达为

$$(P_W^\top h)(y) = \mathbb{E}h(Wy + \xi), \quad \xi \sim N(0, I_d - WW^\top). \quad (6.58)$$

证明 引入 $W_\perp$ 使得 $\tilde{W} = (W, W_\perp)$ 为正交矩阵。令 $\tilde{y} = (y, y')^\top = \tilde{W}^\top x$, 注意到

$$\tilde{W}y = (W, W_\perp)(y, y')^\top = Wy + W_\perp y',$$

因此

$$\begin{aligned} \langle P_W g, h \rangle_{\gamma_d} &= \int g(W^\top x)h(x) d\gamma_d(x) = \int g(y)h(\tilde{W}y) d\gamma_d(y) = \int g(y)h(Wy + W_\perp y') d\gamma_d(y) \\ &= \int \int g(y)h(Wy + W_\perp y') d\gamma_p(y) d\gamma_{d-p}(y') \\ &= \int g(y) \left[\int h(Wy + W_\perp y') d\gamma_{d-p}(y') \right] d\gamma_p(y). \end{aligned}$$

由此得出 $P_W^\top : L^2(\gamma_d) \rightarrow L^2(\gamma_p)$ 的如下公式:

$$(P_W^\top h)(y) = \int h(Wy + W_\perp y') d\gamma_{d-p}(y'),$$

其中 $W_\perp$ 是使得 $\tilde{W} = (W, W_\perp)$ 为正交矩阵的矩阵。或者, 它可以写作

$$(P_W^\top h)(y) = \mathbb{E}h(Wy + \xi), \quad \xi \sim N(0, W_\perp W_\perp^\top).$$

由于 $W_\perp W_\perp^\top$ 是到 W 列空间的正交补空间上的正交投影, 因此

$$W_\perp W_\perp^\top = I_d - WW^\top.$$

最后得出

$$(P_W^\top h)(y) = \mathbb{E}h(Wy + \xi), \quad \xi \sim N(0, I_d - WW^\top). \quad (6.59)$$

■

此外, 对于满足 $\|M\| \leq 1$ 的矩阵 $M \in \mathbb{R}^{p_1 \times p_2}$, 定义

$$\mathcal{A}_M : L^2(\gamma_{p_2}) \rightarrow L^2(\gamma_{p_1}), \quad \mathcal{A}_M h(y) = \mathbb{E}h(M^\top y + \xi), \quad \xi \sim N(0, I_{p_2} - M^\top M). \quad (6.60)$$

以下命题展示了算子 $\mathcal{A}_M$ 的一些基本性质及其与算子 P_W 的关系。

命题 6.12: 具有以下性质:

- (i) 设 $W_1 \in S(d, p_1)$ 且 $W_2 \in S(d, p_2)$ 。那么 $P_{W_1}^\top P_{W_2} = \mathcal{A}_{W_1^\top W_2}$ 。
- (ii) $\mathcal{A}_{M_1} \mathcal{A}_{M_2} = \mathcal{A}_{M_1 M_2}$, $\mathcal{A}_M^\top = \mathcal{A}_{M^\top}$ 。
- (iii) 若 Q 为正交矩阵, 则 $P_Q^\top = P_{Q^\top}$ 且 $\mathcal{A}_Q = P_Q$ 。

证明 为了证明 (i), 记 $\mathcal{A}_{W_1, W_2} = P_{W_1}^\top P_{W_2}$ 并计算:

$$\begin{aligned}\mathcal{A}_{W_1, W_2} h(y) &= \mathbb{E}(P_{W_2} h)(W_1 y + \xi) = \mathbb{E}h(W_2^\top (W_1 y + \xi)) \\ &= \mathbb{E}h(W_2^\top W_1 y + W_2^\top \xi),\end{aligned}$$

其中 $\xi \sim N(0, I_d - W_1 W_1^\top)$ 。现在记 $M = W_1^\top W_2 \in \mathbb{R}^{p_1 \times p_2}$, 有

$$W_2^\top \xi \sim N(0, W_2^\top (I_d - W_1 W_1^\top) W_2) = N(0, I_{p_2} - M^\top M).$$

因此, 有

$$\mathcal{A}_{W_1, W_2} h(y) = \mathcal{A}_M h(y) = \mathbb{E}h(M^\top y + \xi), \quad \xi \sim N(0, I_{p_2} - M^\top M).$$

对于 (ii), 注意到

$$\begin{aligned}\mathcal{A}_{M_1 M_2} h(y) &= \mathbb{E}h((M_1 M_2)^\top y + \xi), \\ \xi &\sim N(0, I_p - (M_1 M_2)^\top (M_1 M_2)) = N(0, I_p - M_2^\top M_1^\top M_1 M_2).\end{aligned}$$

现在可以取

$$\xi = M_2^\top \xi_1 + \xi_2, \quad \xi_1 \sim N(0, I_m - M_1^\top M_1), \quad \xi_2 \sim N(0, I_p - M_2^\top M_2),$$

可以验证 $M_2^\top \xi_1 + \xi_2$ 的方差为

$$M_2^\top (I_m - M_1^\top M_1) M_2 + I_p - M_2^\top M_2 = I_p - M_2^\top M_1^\top M_1 M_2.$$

因此有

$$\begin{aligned}\mathcal{A}_{M_1 M_2} h(y) &= \mathbb{E}h((M_1 M_2)^\top y + \xi) = \mathbb{E}h(M_2^\top M_1^\top y + M_2^\top \xi_1 + \xi_2) \\ &= \mathbb{E}h(M_2^\top (M_1^\top y + \xi_1) + \xi_2) \\ &= \mathcal{A}_{M_1} \mathcal{A}_{M_2} h(y).\end{aligned}$$

此外, 对于 $\mathcal{A}_M$ 的伴随算子:

$$\langle \mathcal{A}_M h, g \rangle_{\gamma_{p_1}} = \mathbb{E}h(M^\top y + \xi_2)g(y), \quad \xi_2 \sim N(0, I_{p_2} - M^\top M), \quad y \sim N(0, I_{p_1}).$$

令 $z = M^\top y + \xi_2 \sim N(0, I_{p_2})$, 有

$$\xi_1 = y - Mz = (I_{p_1} - MM^\top)y + M\xi_2 \sim N(0, I_{p_1} - MM^\top)$$

与 z 独立。因此,

$$\langle \mathcal{A}_M h, g \rangle_{\gamma_{p_1}} = \mathbb{E}h(M^\top y + \xi_2)g(y) = \mathbb{E}h(z)g(Mz + \xi_1) = \langle h, \mathcal{A}_{M^\top} g \rangle_{\gamma_{p_2}}.$$

结论 (iii) 直接由命题 6.11 和 $\mathcal{A}$ 的定义得出。■

命题 6.13: 设 Σ 为对角矩阵, 其对角元为 $\sigma = (\sigma_1, \dots, \sigma_p)$, 满足 $\|\Sigma\| \leq 1$ 。那么

$$\mathcal{A}_\Sigma H_m = \sigma^m H_m.$$

证明 根据 $\mathcal{A}_\Sigma$ 的定义, 有

$$\mathcal{A}_\Sigma H_m(y) = \mathbb{E}_\xi H_m(\Sigma y + \xi) = \prod_{j=1}^p \mathbb{E}_{\xi_j} H_{m_j}(\sigma_j y_j + \xi_j), \quad \xi \sim N(0, I_p - \Sigma^2).$$

对每个 j 使用引理 6.8 即可得出结果。 ■

首先, 有

$$\nabla_x(P_W g)(x) = \nabla_x g(W^\top x) = W(\nabla_u g)(W^\top x) = W P_W(\nabla_u g).$$

引理 6.1: 设 $W \in S(d, p)$, $g \in L^2(\gamma_p)$ 且 $h \in L^2(\gamma_d)$ 。那么

$$\nabla_x(P_W g)(x) = P_W(W \nabla g), \quad \nabla_x(P_W^\top h)(x) = W^\top P_W^\top \nabla h, \quad (6.61)$$

$$\nabla_W P_W g = x P_W(\nabla g)^\top, \quad (6.62)$$

$$\nabla_W^S \langle P_W g, h \rangle_{\gamma_d} = -\Pi_{T_W S} \int \nabla h(x) (P_W \nabla g)^\top d\gamma_d(x). \quad (6.63)$$

证明 为了证明 (6.61), 计算:

$$\nabla_x(P_W g)(x) = \nabla_x g(W^\top x) = W(\nabla g)(W^\top x) = W P_W(\nabla g) = P_W(W \nabla g).$$

同时, 令 $\xi \sim N(0, I_d - W W^\top)$, 有

$$\nabla_x(P_W^\top h)(x) = \nabla_x \mathbb{E} h(W x + \xi) = \mathbb{E} W^\top (\nabla h)(W x + \xi) = W^\top P_W^\top \nabla h.$$

对于 (6.63), 首先计算欧几里得梯度:

$$\begin{aligned} \nabla_W \langle P_W g, h \rangle_{\gamma_d} &= \nabla_W \int g(W^\top x) h(x) d\gamma_d(x) = \int \nabla_W g(W^\top x) h(x) d\gamma_d(x) \\ &= \int x (P_W(\nabla g)^\top) h(x) d\gamma_d(x). \end{aligned}$$

利用 Stein 恒等式, 有

$$\begin{aligned} \nabla_W \langle P_W g, h \rangle_{\gamma_d} &= - \int \nabla_x [(P_W(\nabla g)^\top) h(x)] d\gamma_d(x) \\ &= - \int [h(x) \nabla_x (P_W(\nabla g)^\top) + (\nabla h(x)) P_W(\nabla g)^\top] d\gamma_d(x). \end{aligned}$$

对于第一部分, 有

$$\int h(x) \nabla_x (P_W(\nabla g)^\top) d\gamma_d(x) = \int h(x) W P_W(\nabla^2 g) d\gamma_d(x) = W \int P_W(\nabla^2 g) h(x) d\gamma_d(x).$$

由于 $\int P_W(\nabla^2 g) h(x) d\gamma_d(x)$ 是一个对称矩阵, 其黎曼投影为零:

$$\Pi_{T_W S} \int h(x) \nabla_x (P_W(\nabla g)^\top) d\gamma_d(x) = 0.$$

因此, 最终结果中只保留第二部分。 ■

引理 6.2: 对于满足 $\|M\| \leq 1$ 的 $M \in \mathbb{R}^{p_1 \times p_2}$ 以及 $f, f_i \in L^2(\gamma_{p_1}), i = 1, 2$, 有

$$\nabla_x(\mathcal{A}_M f)(x) = \mathcal{A}_M(M \nabla f). \quad (6.64)$$

证明 (6.64) 的证明类似于引理 6.1 中 (6.61) 的证明, 在此回忆 $\mathcal{A}_M f(x) = \mathbb{E} f(M^\top x + \xi)$, 其中 $\xi \sim N(0, I_{p_2} - M^\top M)$. ■

6.5.1.3 导数与奇异值分解

给定矩阵流 $X(t)$, 对其进行奇异值分解 (SVD) $X(t) = U(t)\Sigma(t)V(t)^\top$, 其中 $U(t), V(t)$ 为正交矩阵, $\Sigma(t)$ 为对角矩阵。那么有

$$\dot{X} = \dot{U}\Sigma V^\top + U\dot{\Sigma}V^\top + U\Sigma\dot{V}^\top.$$

由于 U, V 是正交的, 有 $\dot{U}^\top U + U^\top \dot{U} = 0, \dot{V}^\top V + V^\top \dot{V} = 0$ 。为了展示 Σ 的动力学, 有

$$\dot{\Sigma} = U^\top \dot{X} V - (U^\top \dot{U} \Sigma + \Sigma \dot{V}^\top V) = \text{Diag}(U^\top \dot{X} V) - \text{Diag}(U^\top \dot{U} \Sigma + \Sigma \dot{V}^\top V).$$

由于 $U^\top \dot{U}$ 和 $\dot{V}^\top V$ 是反对称的且 Σ 是对角的, 有

$$\text{Diag}(U^\top \dot{U} \Sigma) = \text{Diag}(\Sigma \dot{V}^\top V) = 0.$$

因此,

$$\dot{\Sigma} = \text{Diag}(U^\top \dot{X} V). \quad (6.65)$$

相应地, 对于方向导数, 同样有

$$D_H \Sigma = \text{Diag}(U^\top (D_H X) V), \quad D_H \sigma_j = u_j^\top (D_H X) v_j, \quad (6.66)$$

其中 u_j, v_j 分别是 U, V 的第 j 列。

6.5.1.4 矩阵微积分

设 M 为对称矩阵, f 为一个函数。可以通过谱分解 $M = Q\Lambda Q^\top$ 来定义 $f(M)$, 即 $f(M) = Qf(\Lambda)Q^\top$, 其中 Λ 是由 M 的特征值构成的对角矩阵, 且 $f(\Lambda)$ 是逐元素应用的。设 X 为一般矩阵, 同样可以定义 $f(X^\top X), f(XX^\top)$ 。假设 $X = U\Sigma V^\top$ 是 X 的奇异值分解。那么容易看出:

$$\begin{aligned} f(X^\top X) &= V f(\Sigma^2) V^\top, & f(XX^\top) &= U f(\Sigma^2) U^\top, \\ X f(X^\top X) &= f(XX^\top) X = U f(\Sigma^2) \Sigma V^\top, & & \\ X^\top f(XX^\top) &= f(X^\top X) X^\top = V \Sigma f(\Sigma^2) U^\top. & & \end{aligned} \quad (6.67)$$

6.5.1.5 初始化

以下命题展示了 Stiefel 流形上随机初始化的奇异值行为^[116]。该内容改编自 Bietti et al.^[104] 中的引理 3.14。

命题 6.14: 设 $W \sim \text{Unif}(S(d, p))$ 且 $W_* \in S(d, p)$ 为固定值。那么对于任意 $\delta > 0$, 存在取决于 δ, p 的常数 $c_1, c_2 > 0$, 使得

$$\mathbb{P} \left\{ \frac{c_1}{\sqrt{d}} \leq \sigma_p \leq \sigma_1 \leq \frac{c_2}{\sqrt{d}} \right\} \geq 1 - \delta. \quad (6.68)$$

6.5.1.6 旋转不变函数

考虑旋转不变函数 $g \in L^2(\gamma_p)$, 其展开式给定为 $g = \sum_{\mathbf{m} \in \mathbb{N}^p} g_{\mathbf{m}} H_{\mathbf{m}}$ 。将给出 g 的梯度的更显式形式。

引理 6.3: 设 $\mathcal{G}(p)$ 为 $L^2(\gamma_p)$ 中旋转不变函数的子空间。那么有

$$\mathcal{G}(p) = \left\{ g = \sum_{r \geq 0} h_r \sum_{|\mathbf{r}|=r} v_r H_{2\mathbf{r}} : \sum_{r \geq 0} h_r^2 < \infty \right\}, \quad (6.69)$$

其中系数 v_r 定义为

$$v_r := C_r^{-\frac{1}{2}} \binom{2\mathbf{r}}{\mathbf{r}}^{1/2} = C_r^{-\frac{1}{2}} \frac{\sqrt{(2\mathbf{r})!}}{(\mathbf{r})!}, \quad C_r = 4^r \frac{(p/2)_r}{r!}, \quad r = |\mathbf{r}|, \quad (6.70)$$

并满足 $\sum_{|\mathbf{r}|=r} v_r^2 = 1$ 。此处 $(a)_r = a(a+1) \cdots (a+r-1)$ 表示升阶乘 (rising factorial)。

证明 首先, 由于 g 是旋转不变的, 则 g 在每个变量上必须是偶函数, 因此如果 $\mathbf{m}$ 包含奇数分量, 则必有 $g_{\mathbf{m}} = 0$ 。剩下的只需考虑 $\mathbf{m} = 2\mathbf{r}$ 的情形。固定 $\mathbf{r}$ 并令 $r = |\mathbf{r}|$ 。对于任何正交矩阵 Q , 计算 $P_Q g$ 的 Hermite 系数。利用引理 6.7, 有

$$\begin{aligned} \langle P_Q g, H_{2\mathbf{r}} \rangle_{\gamma_d} &= \left\langle \sum_{\mathbf{s}} g_{2\mathbf{s}} H_{2\mathbf{s}}, P_Q^\top H_{2\mathbf{r}} \right\rangle_{\gamma_d} \\ &= \sum_{\mathbf{s}: |\mathbf{s}|=r} g_{2\mathbf{s}} \langle H_{2\mathbf{s}}, P_Q^\top H_{2\mathbf{r}} \rangle \\ &= \sum_{\mathbf{s}: |\mathbf{s}|=r} g_{2\mathbf{s}} \sqrt{(2\mathbf{s})! (2\mathbf{r})!} [\alpha^{2\mathbf{r}} \beta^{2\mathbf{s}}] e^{\alpha^\top Q^\top \beta} \\ &= \sqrt{2\mathbf{r}!} [\alpha^{2\mathbf{r}}] \sum_{\mathbf{s}: |\mathbf{s}|=r} g_{2\mathbf{s}} \sqrt{(2\mathbf{s})!} [\beta^{2\mathbf{s}}] e^{(Q\alpha)^\top \beta} \\ &= \sqrt{2\mathbf{r}!} [\alpha^{2\mathbf{r}}] \sum_{\mathbf{s}: |\mathbf{s}|=r} (Q\alpha)^{2\mathbf{s}} ((2\mathbf{s})!)^{-\frac{1}{2}} g_{2\mathbf{s}}. \end{aligned}$$

一方面, 假设 g 的形式为 $g = \sum_{r \geq 0} \sum_{|\mathbf{r}|=r} v_r h_r H_{2\mathbf{r}}$ 。记 $v = Q\alpha$ 。代入 $g_{2\mathbf{s}}$ 的表

达式可得:

$$\begin{aligned}
 \langle P_Q g, H_{2r} \rangle_{\gamma_d} &= C_{|r|}^{-\frac{1}{2}} \sqrt{(2r)!} [\alpha^{2r}] \sum_{s: |s|=r} \frac{1}{s!} h_r v^{2s} \\
 &= C_{|r|}^{-\frac{1}{2}} \sqrt{(2r)!} h_r [\alpha^{2r}] \sum_{s: |s|=r} \frac{1}{s!} \prod_i (v_i^2)^{s_i} \\
 &= C_{|r|}^{-\frac{1}{2}} \sqrt{(2r)!} h_r [\alpha^{2r}] \frac{1}{r!} \left(\sum_i v_i^2 \right)^r \\
 &= C_{|r|}^{-\frac{1}{2}} \sqrt{(2r)!} h_r [\alpha^{2r}] \frac{1}{r!} \|Q\alpha\|^{2r} \\
 &= C_{|r|}^{-\frac{1}{2}} \sqrt{(2r)!} h_r [\alpha^{2r}] \frac{1}{r!} \|\alpha\|^{2r} \\
 &= C_{|r|}^{-\frac{1}{2}} \frac{\sqrt{(2r)!}}{(r)!} h_r,
 \end{aligned}$$

因此 $P_Q g$ 和 g 具有相同的系数, 表明 $P_Q g = g$, 故 $g \in \mathcal{G}(p)$ 。

另一方面, 假设对于所有正交矩阵 Q 都有 $P_Q g = g$ 。定义多项式

$$p(\alpha) = \sum_{|s|=r} g_{2s} \frac{\alpha^{2s}}{\sqrt{(2s)!}},$$

这是一个次数为 $2r$ 的偶次齐次多项式。那么, 对于任何正交矩阵 Q , 有

$$g_{2r} = \langle P_Q g, H_{2r} \rangle_{\gamma_p} = \sqrt{(2r)!} [\alpha^{2r}] p(Q\alpha), \quad \text{对于 } r \in \mathbb{N}^p,$$

这表明 $p(Q^\top \alpha)$ 的 α^{2r} 系数与 $p(\alpha)$ 相同。由于集合 $\{\alpha^{2s} : |s| = r\}$ 张成了 $2r$ 次偶次齐次多项式空间, 该等式意味着 $p(Q^\top \alpha)$ 对所有 $Q \in O(p)$ 均相同。已知次数为 $2r$ 的此类旋转不变多项式是 $\|\alpha\|^{2r}$ 的倍数。因此, 利用多项式定理, 得到

$$p(\alpha) = c_r \|\alpha\|^{2r} = c_r \sum_{|s|=r} \frac{r!}{s!} \alpha^{2s}.$$

通过对比系数可得

$$g_{2s} \frac{1}{\sqrt{(2s)!}} = c_r \frac{r!}{s!} \implies g_{2s} = c_r r! \frac{\sqrt{(2s)!}}{s!}.$$

因此, g 必然具有 (6.69) 中的形式。

最后, 归一化常数 C_r 是由命题 6.15 在 $\lambda = \mathbf{1}$ 时计算得出的。 ■

命题 6.15: 设 $p, r \in \mathbb{N}$ 且 $\lambda \in \mathbb{R}_+^p$ 。那么

$$f(\lambda) = \sum_{|r|=r} \binom{2r}{r} \lambda^r = [z^r] \prod_{j=1}^p (1 - 4\lambda_j z)^{-\frac{1}{2}}. \quad (6.71)$$

特别地, 若 $\lambda = \mathbf{1}$, 有

$$f(\mathbf{1}) = [z^r](1 - 4z)^{-p/2} = 4^r \frac{(p/2)_r}{r!}.$$

证明 回顾恒等式

$$(1 - 4\alpha)^{-1/2} = \sum_{n=0}^{\infty} \binom{2n}{n} \alpha^n.$$

令 $\alpha = \lambda_j z_j$, 有

$$\begin{aligned} \prod_{j=1}^p (1 - 4\lambda_j z_j)^{-1/2} &= \prod_{j=1}^p \sum_{r_j=0}^{\infty} \binom{2r_j}{r_j} \lambda_j^{r_j} z_j^{r_j} \\ &= \sum_{n_1, \dots, n_p} \prod_{j=1}^p \binom{2r_j}{r_j} \lambda_j^{r_j} z_j^{r_j} \\ &= \sum_{\mathbf{r}} \frac{(2\mathbf{r})!}{(\mathbf{r}!)^2} \lambda^{\mathbf{r}} z^{|\mathbf{r}|} \\ &= \sum_{r \geq 0} \left[\sum_{|\mathbf{r}|=r} \frac{(2\mathbf{r})!}{(\mathbf{r}!)^2} \lambda^{\mathbf{r}} \right] z^r. \end{aligned}$$

通过对比系数即可证明该命题。 ■

设 $P \in \mathbb{R}^{p \times p}$ 满足 $\|P\| \leq 1$ 。设 $P = U \Sigma V^T$ 为其 SVD, σ 为 Σ 的对角线。利用命题 6.15, 可以引入函数

$$\phi_r(P) := \sum_{|\mathbf{r}|=r} v_{\mathbf{r}}^2 \sigma^{2\mathbf{r}} = C_r^{-1} [z^r] \det(I - 4P^T P z)^{-\frac{1}{2}}, \quad C_r = 4^r \frac{(p/2)_r}{r!}, \quad (6.72)$$

其中定义 $\phi_0 = 1$ 。该函数定义良好, 且仅取决于 P 的奇异值。为了说明这一点, 利用命题 6.15 得到

$$\begin{aligned} \sum_{|\mathbf{r}|=r} v_{\mathbf{r}}^2 \sigma^{2\mathbf{r}} &= C_r^{-1} \sum_{|\mathbf{r}|=r} \binom{2\mathbf{r}}{\mathbf{r}} \sigma^{2\mathbf{r}} \\ &= C_r^{-1} [z^r] \prod_{j=1}^p (1 - 4\sigma_j^2 z)^{-\frac{1}{2}} = [z^r] \det(I - 4\Sigma^2 z)^{-\frac{1}{2}} \\ &= C_r^{-1} [z^r] \det(I - 4P^T P z)^{-\frac{1}{2}}. \end{aligned}$$

此外, 注意 $\phi_1 = \frac{1}{p} \sum_{j=1}^p \sigma_j^2$ 。

推论 6.3: 考虑旋转不变函数空间上的损失 $L(g)$ 。设 $g = \sum_{\mathbf{m}} g_{\mathbf{m}} H_{\mathbf{m}}$ 。记 $\nabla_{g_{\mathbf{m}}}$ 为关于系数 $g_{\mathbf{m}}$ 的经典梯度, $\nabla_{g_{\mathbf{m}}}^{\mathcal{G}(p)}$ 为关于 $\mathcal{G}(p)$ 的梯度。那么, 有

$$\nabla_{g_{2r}}^{\mathcal{G}(p)} L(g) = v_{\mathbf{r}} \nabla_{h_{\mathbf{r}}} L(g), \quad \nabla_{h_{\mathbf{r}}} L(g) = \sum_{|\mathbf{r}|=r} v_{\mathbf{r}} \nabla_{g_{2r}} L(g), \quad r = |\mathbf{r}|. \quad (6.73)$$

系数还具有以下性质。

命题 6.16: 设系数 v_r 如引理 6.3 中所定义。那么, 有

$$\sum_{|r|=r} r_k v_r^2 = \frac{r}{p}, \quad \forall k = 1, \dots, p. \quad (6.74)$$

此外, 定义

$$A_{ij} = \begin{cases} \sum_{|r|=r} v_r^2 (2r_i)(2r_j) v_r^2, & i \neq j \\ \sum_{|r|=r} v_r^2 (2r_i)(2r_i - 1) v_r^2, & i = j \end{cases}$$

有

$$\sum_{i=1}^p A_{ki} = \frac{2r(2r-1)}{p}, \quad \forall k = 1, \dots, p. \quad (6.75)$$

证明 令 $A_j = \sum_{|r|=r} r_j v_r^2$ 。由对称性可知 $A_1 = \dots = A_p$, 因此

$$A_j = \frac{1}{p} \sum_{j=1}^p A_j = \frac{1}{p} \sum_{|r|=r} v_r^2 (r_1 + \dots + r_p) = \frac{1}{p} \sum_{|r|=r} v_r^2 r = \frac{r}{p} \sum_{|r|=r} v_r^2 = \frac{r}{p}.$$

对于关于 A_{ij} 的陈述, 有

$$\begin{aligned} \sum_{i=1}^p A_{ki} &= \sum_{|r|=r} (2r_k) \left(\sum_{i=1}^p (2r_i) - 1 \right) v_r^2 \\ &= \sum_{i=1}^p A_{ki} = \sum_{|r|=r} (2r_k) (2r - 1) v_r^2 = (2r - 1) \sum_{|r|=r} 2r_k v_r^2 \\ &= \frac{2r(2r-1)}{p}. \end{aligned}$$

■

命题 6.17: 设 $g \in \mathcal{G}(p)$ 为旋转不变函数。那么对于任何正交矩阵 Q , 有

$$\nabla g = Q P_Q \nabla g = P_Q Q \nabla g, \quad \nabla^2 g = Q (P_Q \nabla^2 g) Q^\top.$$

证明 由于 g 是旋转不变的, 有

$$\nabla_x g(x) = \nabla_x (P_Q g)(x) = P_Q (Q \nabla g) = Q P_Q (\nabla g).$$

取二阶导数:

$$\nabla_x^2 g(x) = \nabla_x^\top [Q P_Q (\nabla g)] = Q \nabla_x^\top (P_Q (\nabla g)) = Q (P_Q \nabla^2 g) Q^\top.$$

■

推论 6.4: 设 $\bar{H}_r = \sum_{|r|=r} v_r H_{2r}$ 为 H_r 在 $\mathcal{G}(p)$ 上的投影。那么有

$$\int \nabla \bar{H}_r (\nabla \bar{H}_r)^\top d\gamma_p = \frac{2r}{p} I_p.$$

$$\int (\nabla^2 \bar{H}_r)^2 d\gamma_p = c_r I_p, \quad c_r = \frac{2r(2r-1)}{p} \leq 4r^2.$$

证明 利用命题 6.17, 有

$$\begin{aligned} A &= \int \nabla \bar{H}_r (\nabla \bar{H}_r)^\top d\gamma_p = \int Q(P_Q \nabla \bar{H}_r)(P_Q \nabla \bar{H}_r)^\top Q^\top d\gamma_p \\ &= Q \int (P_Q \nabla \bar{H}_r)(P_Q \nabla \bar{H}_r)^\top d\gamma_p Q^\top = Q \int \nabla \bar{H}_r (\nabla \bar{H}_r)^\top d\gamma_p Q^\top \\ &= Q A Q^\top, \end{aligned}$$

因此 A 是单位矩阵的标量倍。

利用 Hermite 多项式的导数, 进一步计算:

$$\nabla_{x_1} \bar{H}_r = \sum_{|r|=r} v_r \sqrt{2r_1} H_{2r-2e_1},$$

因此

$$A_{11} = \int (\nabla_{x_1} \bar{H}_r)^2 d\gamma_p = \sum_{|r|=r} v_r^2 2r_1 = \frac{2r}{p}.$$

■

6.5.2 特征误差度量

回顾子空间

$$L_{\Phi_W}^2 = \overline{\text{span}} \{H_m(W^\top x), m \in \mathbb{N}^p\} = \{f = g(W^\top x), g \in L^2(\gamma_p)\}.$$

由于 $P_W^\top$ 是到 $L_{\Phi_W}^2$ 的投影, f^* 在 $L_{\Phi_W}^2$ 上的正交投影给定为

$$\begin{aligned} f_W^* &:= (P_W P_W^\top) f^* = P_W P_W^\top P_{W_*} g^* = P_W \mathcal{A}_P g^* \\ &= P_W \mathcal{A}_U \mathcal{A}_\Sigma \mathcal{A}_{V^\top} g^* = P_W P_U \mathcal{A}_\Sigma g^*, \end{aligned}$$

此处使用了奇异值分解 $P = U \Sigma V^\top$ 、命题 6.12 以及 g^* 的旋转不变性。因此, 利用命题 6.13, 有

$$\|f_W^*\|_{\gamma_d}^2 = \|P_W P_U \mathcal{A}_\Sigma g^*\|_{\gamma_d}^2 = \|\mathcal{A}_\Sigma g^*\|_{\gamma_p}^2 = \sum_{m \in \mathbb{N}^p} \sigma^{2m} (g_m^*)^2.$$

那么投影误差为

$$\begin{aligned} \mathcal{E}_{\text{Proj}}(\Phi_W, f^*) &= \|f^* - f_W^*\|_{\gamma_d}^2 = \|f^*\|_{\gamma_d}^2 - \|f_W^*\|_{\gamma_d}^2 = \|g^*\|_{\gamma_p}^2 - \|f_W^*\|_{\gamma_d}^2 \\ &= \sum_{m \in \mathbb{N}^p} (1 - \sigma^{2m}) (g_m^*)^2. \end{aligned}$$

为了计算统计误差，回顾其定义，引入

$$g_{W,m}^* := \langle f^*, H_m(W^\top x) \rangle_{\gamma_d} = \langle f^*, P_W H_m \rangle_{\gamma_d} = \langle g_W^*, H_m \rangle_{\gamma_p},$$

其中

$$g_W^* := P_W^\top f^* = P_W^\top P_{W_*} g^* = \mathcal{A}_P g^* = \mathcal{A}_U \mathcal{A}_\Sigma \mathcal{A}_V^\top g^* = P_U \mathcal{A}_\Sigma g^*.$$

由于 $\lambda_m = \mu_{|m|}$ ，有

$$\mathcal{E}_B = \sum_{m \in \mathbb{N}^p} \mathbf{1}\{\lambda_m < \delta\} (g_{W,m}^*)^2 = \sum_{r \geq 0} \mathbf{1}\{\mu_r < \delta\} \sum_{|m|=r} (g_{W,m}^*)^2.$$

此外，考虑子空间 $\mathcal{H}_r = \text{span}\{H_m, |m| = r\}$ ，并令 $P_{\mathcal{H}_r}$ 为到 $\mathcal{H}_r$ 的正交投影。由于根据推论 6.6， $\mathcal{H}_r$ 和 $\mathcal{H}_r^\perp$ 在 P_U 下都是不变的，有

$$\begin{aligned} \sum_{|m|=r} (g_{W,m}^*)^2 &= \|P_{\mathcal{H}_r} g_W^*\|_{\gamma_p}^2 = \|P_{\mathcal{H}_r} P_U \mathcal{A}_\Sigma g^*\|_{\gamma_p}^2 = \|P_U P_{\mathcal{H}_r} \mathcal{A}_\Sigma g^*\|_{\gamma_p}^2 \\ &= \|P_{\mathcal{H}_r} \mathcal{A}_\Sigma g^*\|_{\gamma_p}^2 = \sum_{|m|=r} \sigma^{2m} (g_m^*)^2. \end{aligned}$$

因此，

$$\mathcal{E}_B = \sum_{r \geq 0} \mathbf{1}\{\mu_r < \delta\} \sum_{|m|=r} \sigma^{2m} (g_m^*)^2.$$

另一方面，

$$\begin{aligned} \mathcal{E}_V &= \#\{m \in \mathbb{N}^p : \lambda_m \geq \delta\} \epsilon^2 = \epsilon^2 \sum_{r \geq 0} \mathbf{1}\{\mu_r \geq \delta\} \sum_{|m|=r} 1 \\ &= \epsilon^2 \sum_{r \geq 0} \mathbf{1}\{\mu_r \geq \delta\} \binom{p+r-1}{r}. \end{aligned}$$

合并这两项就得到以下命题。

命题 6.18： 在假设 6.3 下，有

$$\mathcal{E}_{\text{Proj}}(\Phi_W, f^*) = \sum_{m \in \mathbb{N}^p} (1 - \sigma^{2m}) (g_m^*)^2, \quad (6.76)$$

$$\mathcal{E}_B = \sum_{r \geq 0} \mathbf{1}\{\mu_r < \delta\} \sum_{|m|=r} \sigma^{2m} (g_m^*)^2, \quad \mathcal{E}_V = \epsilon^2 \sum_{r \geq 0} \mathbf{1}\{\mu_r \geq \delta\} \binom{p+r-1}{r}. \quad (6.77)$$

命题 6.19： 在假设 6.5 下，有

$$\mathcal{E}(\delta, \epsilon^2; \Phi_W, f^*) - \mathcal{E}(\delta, \epsilon^2; \Phi_{W_*}, f^*) = \mathcal{E}_{\text{Proj}}(\Phi_W, f^*) \lesssim \begin{cases} 1 - \rho, & \alpha > 1, \\ (1 - \rho) \log(1 - \rho)^{-1}, & \alpha = 1, \\ (1 - \rho)^\alpha, & \alpha \in (0, 1). \end{cases} \quad (6.78)$$

其中 $\rho = \min_j \sigma_j^2$ 。

证明 该证明类似于命题 6.1，但现在处理的是多指标模型的情况。根据 (6.76)，取 $L = c(1 - \rho)^{-1}$ ，有

$$\begin{aligned} \mathcal{E}_{\text{Proj}}(\Phi_W, f^*) &= \sum_{\mathbf{m} \in \mathbb{N}^p} (1 - \sigma^{2\mathbf{m}})(g_{\mathbf{m}}^*)^2 \leq \sum_{\mathbf{m} \in \mathbb{N}^p} (1 - \rho^{|\mathbf{m}|})(g_{\mathbf{m}}^*)^2 \leq \sum_{\mathbf{m} \in \mathbb{N}^p} \min(1, |\mathbf{m}|(1 - \rho))(g_{\mathbf{m}}^*)^2 \\ &= \sum_{|\mathbf{m}| \leq L} 2|\mathbf{m}|(1 - \rho)(g_{\mathbf{m}}^*)^2 + \sum_{|\mathbf{m}| > L} (g_{\mathbf{m}}^*)^2 \\ &= I_1 + I_2. \end{aligned}$$

将使用以下事实：

$$\#\{\mathbf{m} \in \mathbb{N}^p : |\mathbf{m}| = r\} = \binom{p+r-1}{r} \lesssim r^{p-1}.$$

另外，回顾 $|g_{\mathbf{m}}^*| \lesssim |\mathbf{m}|^{-\frac{\alpha+p}{2}}$ 。

第一项的界为：

$$\begin{aligned} I_1 &\lesssim 2(1 - \rho) \sum_{|\mathbf{m}| \leq L} |\mathbf{m}|^{-(\alpha+p-1)} = 2(1 - \rho) \sum_{r \leq L} \sum_{|\mathbf{m}|=r} |\mathbf{m}|^{-(\alpha+p-1)} \\ &\lesssim 2(1 - \rho) \sum_{r \leq L} r^{p-1} r^{-(\alpha+p-1)} = 2(1 - \rho) \sum_{r \leq L} r^{-\alpha}. \end{aligned}$$

因此，

$$I_1 \lesssim \begin{cases} 1 - \rho, & \alpha > 1, \\ (1 - \rho) \log(1 - \rho)^{-1}, & \alpha = 1, \\ (1 - \rho)^\alpha, & \alpha \in (0, 1). \end{cases}$$

对于第二项，类似地有

$$I_2 \lesssim \sum_{|\mathbf{m}| > L} |\mathbf{m}|^{-(\alpha+p)} \lesssim \sum_{r > L} r^{p-1} r^{-(\alpha+p)} = \sum_{r > L} r^{-(\alpha+1)} \lesssim L^{-\alpha}.$$

合并这些界限，命题得证。 ■

6.5.3 总体动力学

本节引入多指标模型的总体动力学。令 $\mathcal{L} = \frac{1}{2} \|f - f^*\|_{L^2(\gamma_d)}^2$ 表示总体损失 (population loss)。遵循 (6.17)，考虑如下动力学方程：

$$\begin{cases} \dot{\beta}(t) = -\nabla_{\beta}^{\mathcal{G}(p)} \mathcal{L}, & \beta(0) = \mathbf{0}, \\ \dot{W}(t) = -\nabla_W^{\text{St}(d,p)} \mathcal{L}, & W(0) \sim \text{Unif}(\text{St}(d, p)). \end{cases} \quad (6.79)$$

6.5.3.1 梯度流的计算

利用辅助算子并注意到 $f = P_W g$ 且 $f^* = P_{W^*} g^*$, 可以将总体损失写作:

$$\begin{aligned}\mathcal{L} &= \frac{1}{2} \|f - f^*\|_{L^2(\gamma_d)}^2 = \frac{1}{2} \|f\|_{L^2(\gamma_d)}^2 + \frac{1}{2} \|f^*\|_{L^2(\gamma_d)}^2 - \langle f, f^* \rangle_{L^2(\gamma_d)} \\ &= \frac{1}{2} \|g\|_{\gamma_p}^2 + \frac{1}{2} \|g^*\|_{\gamma_{p^*}}^2 - \left\langle P_W g, P_{W^*} g^* \right\rangle_{\gamma_d} = \frac{1}{2} \|g\|_{\gamma_p}^2 + \frac{1}{2} \|g^*\|_{\gamma_{p^*}}^2 - \left\langle g, \mathcal{A}_{W^\top W^*} g^* \right\rangle_{\gamma_p},\end{aligned}$$

其中利用了 P_W 是等距算子的性质。此外, 根据命题 6.12 以及 g 和 g^* 的旋转不变性, 有:

$$\begin{aligned}\left\langle g, \mathcal{A}_{W^\top W^*} g^* \right\rangle_{\gamma_p} &= \langle g, \mathcal{A}_P g^* \rangle_{\gamma_p} = \langle g, \mathcal{A}_U \mathcal{A}_\Sigma \mathcal{A}_{V^\top} g^* \rangle_{\gamma_p} \\ &= \langle \mathcal{A}_{U^\top} g, \mathcal{A}_\Sigma \mathcal{A}_{V^\top} g^* \rangle_{\gamma_p} = \langle P_{U^\top} g, \mathcal{A}_\Sigma P_{V^\top} g^* \rangle_{\gamma_p} \\ &= \langle g, \mathcal{A}_\Sigma g^* \rangle_{\gamma_p}.\end{aligned}$$

因此得到:

$$\mathcal{L} = \frac{1}{2} \|g\|_{\gamma_p}^2 + \frac{1}{2} \|g^*\|_{\gamma_{p^*}}^2 - \langle g, \mathcal{A}_\Sigma g^* \rangle_{\gamma_p}. \quad (6.80)$$

h_r 的动力学 利用 (6.80), 有:

$$\nabla_{g_m} \mathcal{L} = \nabla_{g_m} \frac{1}{2} \|g\|_{\gamma_p}^2 - \nabla_{g_m} \langle g, \mathcal{A}_\Sigma g^* \rangle_{\gamma_p} = g_m - (\mathcal{A}_\Sigma g^*)_m,$$

根据推论 6.3, 对于 $m = 2r, |r| = r$, 有:

$$\nabla_{g_m}^{\mathcal{G}(p)} \mathcal{L} = \nu_r G_r, \quad G_r = \sum_{s: |s|=r} \nu_s [g_{2s} - (\mathcal{A}_\Sigma g^*)_{2s}] = \sum_{s: |s|=r} \nu_s [g_{2s} - (\mathcal{A}_\Sigma g^*)_{2s}].$$

因此有:

$$\dot{\beta}_m = -\nabla_{g_m}^{\mathcal{G}(p)} \mathcal{L} \cdot \frac{\partial g_m}{\partial \beta_m} = -\lambda_m^{\frac{1}{2}} \nu_r G_r,$$

从而有:

$$\dot{g}_m = \lambda_m^{\frac{1}{2}} \dot{\beta}_m = -\lambda_m \nu_r G_r, \quad \dot{h}_r = -\mu_{2r} G_r,$$

此处回顾 $\lambda_m = \mu_{|m|} = \mu_{2r}$ 。进一步计算 G_r , 由命题 6.13 可得:

$$(\mathcal{A}_\Sigma g^*)_{2s} = \sigma^{2s} g_{2s}^* = \sigma^{2s} \nu_s h_r^*.$$

因此:

$$\begin{aligned}G_r &= \sum_{s: |s|=r} \nu_s [\nu_s h_r - \sigma^{2s} \nu_s h_r^*] = \sum_{s: |s|=r} \nu_s^2 h_r - \sum_{s: |s|=r} \nu_s^2 \sigma^{2s} h_r^* \\ &= h_r - \sum_{s: |s|=r} \nu_s^2 \sigma^{2s} h_r^*.\end{aligned}$$

引入辅助量:

$$\phi_r := \sum_{r: |r|=r} v_r^2 \sigma^{2r}. \quad (6.81)$$

最终有:

$$\dot{h}_r = \mu_{2r} [\phi_r h_r^* - h_r].$$

W 的动力学 回顾 $\mathcal{L} = \frac{1}{2} \|g\|_{\gamma_p}^2 + \frac{1}{2} \|g^*\|_{\gamma_{p^*}}^2 - \langle P_W g, f^* \rangle_{\gamma_d}$. 利用引理 6.1, 有:

$$\begin{aligned} \dot{W} &= -\nabla_W^S \mathcal{L} = \nabla_W^S \langle P_W g, f^* \rangle_{\gamma_d} = \Pi_{T_W S} \int (\nabla f^*)(x) (P_W \nabla g)^\top d\gamma_d(x) \\ &= \Pi_{T_W S} \int W_*(P_{W_*}(\nabla g^*)) (P_W \nabla g)^\top d\gamma_d(x) \\ &= \Pi_{T_W S} W_* B^\top, \end{aligned}$$

其中

$$B = \int P_{W_*} (P_W \nabla g) (P_{W_*}(\nabla g^*))^\top d\gamma_d(x).$$

因此:

$$\begin{aligned} \dot{P} &= \dot{W}^\top W_* = (\Pi_{T_W S} W_* B^\top)^\top W_* = [W_* B^\top - W \text{Sym}(W^\top W_* B^\top)]^\top W_* \\ &= B - \text{Sym}(P B^\top) P. \end{aligned}$$

Σ 的动力学 为了计算 Σ 的动力学, 利用 (6.65) 得到:

$$\begin{aligned} \dot{\Sigma} &= \text{Diag}(U^\top \dot{P} V) = \text{Diag}(U^\top (B - \text{Sym}(P B^\top) P) V) \\ &= \text{Diag}(U^\top B V) - \text{Diag}(U^\top \text{Sym}(P B^\top) P V) \\ &= \text{Diag}(U^\top B V) - \frac{1}{2} \text{Diag}(U^\top U \Sigma V^\top B^\top U \Sigma V^\top V) - \frac{1}{2} \text{Diag}(U^\top B P^\top P V) \\ &= \text{Diag}(U^\top B V) - \frac{1}{2} \text{Diag}(\Sigma V^\top B^\top U \Sigma) - \frac{1}{2} \text{Diag}(U^\top B V \Sigma^2). \end{aligned}$$

定义 $\tilde{B} = U^\top B V \in \mathbb{R}^{p \times p}$. 则:

$$\begin{aligned} \dot{\Sigma} &= \text{Diag}(\tilde{B}) - \frac{1}{2} \text{Diag}(\Sigma \tilde{B}^\top \Sigma) - \frac{1}{2} \text{Diag}(\tilde{B} \Sigma^2) \\ &= \text{Diag}(\tilde{B}) - \frac{1}{2} \Sigma \text{Diag}(\tilde{B}^\top) \Sigma - \frac{1}{2} \Sigma^2 \text{Diag}(\tilde{B}) \\ &= \text{Diag}(\tilde{B}) - \Sigma^2 \text{Diag}(\tilde{B}) \\ &= (I_p - \Sigma^2) \text{Diag}(\tilde{B}). \end{aligned}$$

接下来计算 $\text{Diag}(\tilde{B})$. 利用 SVD 和命题 6.12, 可将 B 写作:

$$B = \left\langle P_W \nabla g, P_{W_*} \nabla g^* \right\rangle_{\gamma_d} = \left\langle \nabla g, P_W^\top P_{W_*}(\nabla g^*)^\top \right\rangle_{\gamma_p} = \left\langle \nabla g, \mathcal{A}_P(\nabla g^*)^\top \right\rangle_{\gamma_p}$$

$$\begin{aligned}
 &= \langle \nabla g, \mathcal{A}_U \mathcal{A}_\Sigma \mathcal{A}_{V^\top} (\nabla g^*)^\top \rangle_{\gamma_p} \\
 &= \langle \mathcal{A}_{U^\top} \nabla g, \mathcal{A}_\Sigma \mathcal{A}_{V^\top} (\nabla g^*)^\top \rangle_{\gamma_p}.
 \end{aligned}$$

由此得：

$$\begin{aligned}
 \tilde{B} &= U^\top B V = U^\top \langle \mathcal{A}_{U^\top} \nabla g, \mathcal{A}_\Sigma \mathcal{A}_{V^\top} (\nabla g^*)^\top \rangle_{\gamma_p} V = \langle U^\top \mathcal{A}_{U^\top} \nabla g, \mathcal{A}_\Sigma \mathcal{A}_{V^\top} (V^\top \nabla g^*)^\top \rangle_{\gamma_p} \\
 &\stackrel{(a)}{=} \langle \nabla(\mathcal{A}_{U^\top} g), \mathcal{A}_\Sigma \nabla(\mathcal{A}_{V^\top} g^*)^\top \rangle_{\gamma_p} \\
 &\stackrel{(b)}{=} \langle \nabla(P_{U^\top} g), \mathcal{A}_\Sigma (\nabla(P_{V^\top} g^*))^\top \rangle_{\gamma_p} \\
 &\stackrel{(c)}{=} \langle \nabla g, \mathcal{A}_\Sigma (\nabla g^*)^\top \rangle_{\gamma_p},
 \end{aligned}$$

其中：步骤 (a) 利用了引理 6.2 中的 (6.64)；步骤 (b) 利用了命题 6.12；步骤 (c) 利用了 g 和 g^* 的旋转不变性。最后，有：

$$\begin{aligned}
 \tilde{B}_{ii} &= \left\langle \nabla_{x_i} g, \mathcal{A}_\Sigma \nabla_{x_i} g^* \right\rangle_{\gamma_p} = \left\langle \sum \sqrt{m_i + 1} g_{m+e_i} H_m, \mathcal{A}_\Sigma \sum \sqrt{n_i + 1} g_{n+e_i}^* H_n \right\rangle_{\gamma_p} \\
 &= \sum \sqrt{(n_i + 1)(m_i + 1)} g_{m+e_i} g_{n+e_i}^* \langle H_m, \mathcal{A}_\Sigma H_n \rangle_{\gamma_p} \\
 &= \sum \sqrt{(n_i + 1)(m_i + 1)} g_{m+e_i} g_{n+e_i}^* \langle H_m, \sigma^n H_n \rangle_{\gamma_p} \\
 &= \sum_m (m_i + 1) \sigma^m g_{m+e_i} g_{m+e_i}^* \\
 &= \sum_m m_i \sigma^{m-e_i} g_m g_m^* \\
 &= \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} 2r_i \sigma^{2r-e_i} v_r^2,
 \end{aligned}$$

从而得到：

$$\dot{\sigma}_i = (1 - \sigma_i^2) \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} 2r_i \sigma^{2r-e_i} v_r^2,$$

以及：

$$\frac{d}{dt} \sigma_i^2 = 2\sigma_i \dot{\sigma}_i = 4(1 - \sigma_i^2) \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} r_i v_r^2 \sigma^{2r}.$$

ϕ_r 和 ω 的动力学 对于 ϕ_r ，有：

$$\frac{d}{dt} \phi_r = \sum_{|r|=r} v_r^2 \frac{d}{dt} \sigma^{2r} = \sum_{|r|=r} v_r^2 \sum_{i=1}^p \sigma^{2(r-e_i)} 2r_i \frac{d}{dt} \sigma_i^2.$$

对于 ω ，回顾定义：

$$\omega := -\frac{1}{K} \log \left(\sum_{i=1}^p e^{-K\sigma_i^2} \right).$$

利用命题 6.32 得到:

$$\dot{\omega} = \sum_{i=1}^p \frac{\partial \omega}{\partial \sigma_i^2} \frac{d}{dt} \sigma_i^2 = \sum_{i=1}^p \frac{e^{-K\sigma_i^2}}{\sum_{j=1}^p e^{-K\sigma_j^2}} \cdot (1 - \sigma_i^2) \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} r_i v_r^2 \sigma^{2r}.$$

汇总以上结果, 有以下命题:

命题 6.20: 考虑总体动力学 (6.79). 其演化遵循以下动力学方程:

$$\begin{aligned} \dot{h}_r &= \mu_{2r} (\phi_r h_r^* - h_r) \\ \dot{W} &= \Pi_{T_W S} W_* B^\top, \quad B = \left\langle P_W \nabla g, (P_{W_*} \nabla g^*)^\top \right\rangle_{\gamma_d} \in \mathbb{R}^{p \times p} \\ \dot{P} &= B - \text{Sym}(P B^\top) P \\ \frac{d}{dt} \sigma_i^2 &= 4(1 - \sigma_i^2) \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} r_i \sigma^{2r} v_r^2 \\ \dot{\phi}_r &= \sum_{|r|=r} v_r^2 \sum_{i=1}^p \sigma^{2(r-e_i)} 2r_i \frac{d}{dt} \sigma_i^2, \quad \dot{\phi}_1 = \frac{1}{p} \sum_{i=1}^p \frac{d}{dt} \sigma_i^2 \\ \dot{\omega} &= \frac{4}{\sum_{j=1}^p e^{-K\sigma_j^2}} \sum_{i=1}^p e^{-K\sigma_i^2} \frac{d}{dt} \sigma_i^2. \end{aligned} \tag{6.82}$$

6.5.3.2 动力学分析

命题 6.21: 考虑总体动力学 (6.79). 假设 $\rho(0) \neq 0$. 那么, 对于所有的 $r \geq 0$ 和 $t \geq 0$, 有:

$$h_r^* h_r(t) \geq 0, \quad \dot{\phi}_r(t) \geq 0, \quad \dot{\omega} \geq 0.$$

此外, 有如下界限:

$$\begin{aligned} \dot{\omega} &\geq C(1 - \phi_1) \sum_{r \geq 1} r \omega^r h_r h_r^* \\ \dot{\phi}_1 &\geq C(1 - \phi_1) \sum_{r \geq 1} r \omega^r h_r h_r^* \\ \dot{h}_r &\geq \mu_{2r} (\omega^r h_r^* - h_r), \quad (\text{假设 } h_r^* > 0). \end{aligned} \tag{6.83}$$

证明 不失一般性地, 可以假设 h_r^* 为正. 为了证明第一个陈述, 观察到如果 $\phi_r(t) \geq 0$, 则 h_r 将是非负的. 随后, ϕ_r 的动力学表明 $\dot{\phi}_r \geq 0$, 这反过来保证了 h_r 的非负性. 可以通过常微分方程理论中的标准反证法进行严格证明.

对于第二个陈述, 首先有:

$$\frac{d}{dt} \sigma_i^2 = 4(1 - \sigma_i^2) \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} r_i v_r^2 \sigma^{2r} \geq 4(1 - \sigma_i^2) \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} r_i v_r^2 \omega^{|r|}$$

$$\begin{aligned}
 &\stackrel{(a)}{=} 4(1 - \sigma_i^2) \sum_{r \geq 1} h_r h_r^* \omega^r \frac{r}{p} = \frac{4}{p} (1 - \sigma_i^2) \sum_{r \geq 1} r \omega^r h_r h_r^* \\
 &=: (1 - \sigma_i^2) A, \quad A = \frac{4}{p} \sum_{r \geq 1} r \omega^r h_r h_r^*,
 \end{aligned}$$

其中在步骤 (a) 中应用了 (6.74)。将其代入 ϕ_r 的动力学方程中, 有:

$$\begin{aligned}
 \dot{\phi}_r &= \sum_{|r|=r} v_r^2 \sum_{i=1}^p \sigma_i^{2(r-e_i)} 2r_i \frac{d}{dt} \sigma_i^2 \geq \sum_{|r|=r} v_r^2 \sum_{i=1}^p \omega^{|r|-1} 2r_i (1 - \sigma_i^2) A \\
 &= A \omega^{r-1} \sum_{i=1}^p (1 - \sigma_i^2) \sum_{|r|=r} v_r^2 2r_i = A \omega^{r-1} \sum_{i=1}^p (1 - \sigma_i^2) \frac{2r}{p} \\
 &= 2r A (1 - \phi_1) \omega^{r-1}.
 \end{aligned}$$

特别地, 有:

$$\phi_1 \geq C(1 - \phi_1) \sum_{r \geq 1} r \omega^r h_r h_r^*.$$

对于 ω , 有:

$$\begin{aligned}
 \dot{\omega} &= \frac{1}{\sum_{j=1}^p e^{-K\sigma_j^2}} \sum_{i=1}^p e^{-K\sigma_i^2} \frac{d}{dt} \sigma_i^2 \geq \frac{1}{\sum_{j=1}^p e^{-K\sigma_j^2}} \sum_{i=1}^p e^{-K\sigma_i^2} (1 - \sigma_i^2) A \\
 &\geq (1 - \phi_1) A,
 \end{aligned}$$

其中在最后一个不等式中应用了命题 6.33。 ■

命题 6.22: 考虑总体动力学 (6.79)。令假设 6.4 成立且 $r_0 = m_0/2$ 。假设 $\rho_0 = \min_{1 \leq j \leq p} \sigma_j(0) \neq 0$ 。那么, 通过取 $K \geq 2\rho_0^{-2} \log p$, 有对于所有的 $t \geq T^{\text{app}}$:

$$\min_{1 \leq j \leq p} |\sigma_j(t)| \geq \frac{1}{2}, \quad |h_{r_0}(t)| \geq 2^{-(m_0+1)} |h_{r_0}^*|, \quad (6.84)$$

其中

$$T^{\text{app}} \lesssim \log \rho_0^{-1} + \rho_0^{-2(m_0-1)}. \quad (6.85)$$

证明 可以专注于满足 $\phi_1(t) \leq 1 - \frac{1}{2p}$ 的 t : 如果 $\phi_1(t) \geq 1 - \frac{1}{2p}$, 由于 $p\phi_1(t) \leq p-1+\sigma_j^2$, 已有 $\sigma_j^2 \geq 1/2$ 。利用命题 6.32 中关于 ω 的性质, 只要取 $K \geq 2\rho_0^{-2} \log p$, 就有:

$$\omega(0) \geq \rho_0^2 - \frac{1}{K} \log p \geq \frac{1}{2} \rho_0^2.$$

现在, 可以利用 ω 的动力学得到:

$$\dot{\omega} \geq c p^{-2} \sum_{r \geq 1} r \omega^r h_r h_r^* \geq c \omega^{r_0} h_{r_0} h_{r_0}^*,$$

同时有：

$$\dot{h}_{r_0} \geq c \left(\omega^{r_0} h_{r_0}^* - h_{r_0} \right).$$

因此，可以遵循与命题 6.4 相同的分析思路。定义：

$$T_0^\rho = 0, \quad T_k^g = \inf \left\{ t \geq T_k^\rho : h_{r_0}(t) \geq \frac{1}{2} \rho_k^{2r_0} h_{r_0}^* \right\}, \quad T_k^\rho = \inf \left\{ t \geq T_{k-1}^g : \omega(t) \geq \rho_k^2 \right\}, \quad (6.86)$$

其中 $\rho_k = 2^k \rho_0$ 且 $k \leq L$, $L := 1 + \lceil \log_2 \rho_0^{-1} \rceil$ 。当 $t \in [T_k^\rho, T_k^g]$ 时，有：

$$\dot{h}_{r_0} \geq \frac{1}{2} c \rho_k^{2r_0} h_{r_0}^* = c \rho_k^{2r_0} h_{r_0}^*.$$

故

$$T_k^g - T_k^\rho \leq C.$$

另一方面，当 $t \in [T_{k-1}^g, T_k^\rho]$ 时，有：

$$\dot{\omega} \geq c \rho_k^{2r_0} \cdot \rho_k^{2r_0} h_{r_0}^* \cdot h_{r_0}^* = c \rho_k^{4r_0} (h_{r_0}^*)^2.$$

因此，

$$T_k^\rho - T_{k-1}^g \leq \frac{\rho_k^2 - \rho_{k-1}^2}{\rho_k^{4r_0} (h_{r_0}^*)^2} = C \rho_0^{-2(2r_0-1)} 2^{-2(2r_0-1)k}.$$

综上所述，

$$T_L^\rho \leq \sum_{k \leq L} [(T_k^\rho - T_{k-1}^g) + (T_{k-1}^g - T_{k-1}^\rho)] \lesssim \log \rho_0^{-1} + \rho_0^{-2(2r_0-1)}.$$

由于 $m_0 = 2r_0$ ，上式就给出了所需的界限。 ■

命题 6.23： 考虑总体动力学 (6.79)。令假设 6.4 成立且 $r_0 = m_0/2$ 。假设 (6.84) 对于某个 t_0 成立。那么，

$$1 - \phi_1(t_0 + t) \leq \frac{1}{2} e^{-ct}. \quad (6.87)$$

证明 当条件 (6.84) 成立时，有：

$$\dot{\phi}_1 \geq C(1 - \phi_1) \sum_{r \geq 1} r \omega^r h_r h_r^* \geq C(1 - \phi_1) r_0 \omega^{r_0} h_{r_0} h_{r_0}^* \geq c(1 - \phi_1). \quad \blacksquare$$

定理定理 6.3 的证明 特征误差度量的单调性遵循 (6.76) 以及命题 6.21 中 σ_j 的单调性。对于初始化，根据命题 6.14，有：

$$\frac{c_1}{\sqrt{d}} \leq \sigma_j(0) \leq \frac{c_2}{\sqrt{d}}, \quad \forall j = 1, \dots, p.$$

以高概率成立。因此，有：

$$\mathcal{E}_{\text{Proj}}(\Phi_{W(0)}, f^*) = \sum_{\mathbf{m} \in \mathbb{N}^p} (1 - \sigma^{2\mathbf{m}})(g_{\mathbf{m}}^*)^2 \gtrsim \sum_{\mathbf{m} \in \mathbb{N}^p} (g_{\mathbf{m}}^*)^2 \gtrsim 1.$$

另一方面，由于 $\rho_0 = \min_j \sigma_j(0) \geq c_1/\sqrt{d}$ ，命题 6.22 和命题 6.23 表明对于某些 $t_0 \asymp \log d + d^{r_0-1}$ ，有：

$$1 - |\sigma_j(t_0 + t)| \leq \frac{1}{2}e^{-Ct}, \quad \forall j = 1, \dots, p.$$

因此，通过应用命题 6.19 并调整常数，即可得出结论。 ■

6.5.4 序列模型的分析

6.5.4.1 动力学计算

首先计算自适应核模型的动力学。利用 (6.110) 对第 6.5.3.1 小节中的计算进行修正以完成计算。

h_r 的动力学 首先，利用 (6.110) 和推论 6.3，有：

$$-\nabla_{g_{\mathbf{m}}}^{\mathcal{G}(p)} \bar{\mathcal{L}}_n = -\nabla_{g_{\mathbf{m}}}^{\mathcal{G}(p)} \mathcal{L} + \sum_{\mathbf{n} \in \mathbb{N}^d} \varepsilon_n \nabla_{g_{\mathbf{m}}}^{\mathcal{G}(p)} f_n = -\nabla_{g_{\mathbf{m}}}^{\mathcal{G}(p)} \mathcal{L} + \sum_{\mathbf{n} \in \mathbb{N}^d} \varepsilon_n \nu_r \nabla_{h_r} f_n,$$

其中 $\mathbf{m} = 2\mathbf{r}$ ， $r = |\mathbf{r}|$ ，且

$$\nabla_{h_r} f_n = \sum_{|\mathbf{r}|=r} \nu_r \nabla_{g_{2\mathbf{r}}} f_n, \quad f_n := \langle f, H_n \rangle_{\gamma_d} = \langle P_W g, H_n \rangle_{\gamma_d}.$$

因此，有：

$$\dot{\beta}_{\mathbf{m}} = -\nabla_{g_{\mathbf{m}}}^{\mathcal{G}(p)} \bar{\mathcal{L}}_n \cdot \frac{\partial g_{\mathbf{m}}}{\partial \beta_{\mathbf{m}}} = \lambda_{\mathbf{m}}^{\frac{1}{2}} \left[-\nabla_{g_{\mathbf{m}}}^{\mathcal{G}(p)} \mathcal{L} + \nu_r \sum_{\mathbf{n} \in \mathbb{N}^d} \varepsilon_n \nabla_{h_r} f_n \right],$$

进而得出：

$$\dot{h}_r = \mu_{2r} (-G_r + e_r), \quad e_r = \sum_{\mathbf{n} \in \mathbb{N}^d} \varepsilon_n \nabla_{h_r} f_n.$$

进一步计算 e_r 。首先，有：

$$\begin{aligned} \nabla_{h_r} f_n &= \nabla_{h_r} \langle P_W g, H_n \rangle_{\gamma_d} = \nabla_{h_r} \langle g, P_W^\top H_n \rangle_{\gamma_p} \\ &= \nabla_{h_r} \left\langle \sum_{s \geq 0} h_s \sum_{|\mathbf{s}|=s} \nu_s H_{2s}, P_W^\top H_n \right\rangle_{\gamma_p} \\ &= \nabla_{h_r} \sum_{s \geq 0} h_s \sum_{|\mathbf{s}|=s} \nu_s \langle H_{2s}, P_W^\top H_n \rangle_{\gamma_p} \end{aligned}$$

$$= \sum_{|r|=r} \nu_r \langle H_{2r}, P_W^\top H_n \rangle_{\gamma_p}.$$

将此代入 e_r , 有:

$$\begin{aligned} e_r &= \sum_{n \in \mathbb{N}^d} \varepsilon_n \sum_{|r|=r} \nu_r \langle H_{2r}, P_W^\top H_n \rangle_{\gamma_d} = \sum_{n \in \mathbb{N}^d} \varepsilon_n \left\langle P_W \left(\sum_{|r|=r} \nu_r H_{2r} \right), H_n \right\rangle_{\gamma_d} \\ &= \sum_{n \in \mathbb{N}^d} \varepsilon_n \langle P_W \bar{H}_r, H_n \rangle_{\gamma_d}, \end{aligned}$$

其中 $\bar{H}_r = \sum_{|r|=r} \nu_r H_{2r}$.

Σ 的动力学 利用 (6.110), 有:

$$\dot{W} = -\nabla_W^S \bar{\mathcal{L}}_n = -\nabla_W^S \mathcal{L} + \sum_{n \in \mathbb{N}^d} \varepsilon_n \nabla_W^S f_n,$$

因此, 遵循总体情形的计算, 有 $\dot{P} = \dot{W}^\top W_*$ 且

$$\dot{\Sigma} = \text{Diag}(U^\top \dot{P} V) = (I_p - \Sigma^2) \text{Diag}(\tilde{B}) + \text{Diag}(U^\top \left(\sum_{n \in \mathbb{N}^d} \varepsilon_n \nabla_W^S f_n \right)^\top W_* V).$$

最后一项是需要分析的误差项。记:

$$\Delta := \sum_{n \in \mathbb{N}^d} \varepsilon_n U^\top (\nabla_W^S f_n)^\top W_* V.$$

首先, 回顾 f_n 的定义和 (6.56), 有:

$$\nabla_W^S f_n = \nabla_W^S \langle P_W g, H_n \rangle_{\gamma_d} = \Pi_{T_W S} \nabla_W \langle P_W g, H_n \rangle_{\gamma_d} = \Pi_{T_W S} Z_n,$$

其中有记 $Z_n = \nabla_W \langle P_W g, H_n \rangle_{\gamma_d} \in \mathbb{R}^{d \times p}$ 。

利用 (6.56) 展开 $\Pi_{T_W S} Z_n$, 有:

$$\begin{aligned} \Delta &= \sum_n \varepsilon_n U^\top [Z_n - W \text{Sym}(W^\top Z_n)]^\top W_* V \\ &= \sum_n \varepsilon_n U^\top \left[Z_n^\top - \frac{1}{2}(W^\top Z_n + Z_n^\top W) W^\top \right] W_* V. \end{aligned}$$

现在, 记

$$W_* = WQ + E, \quad Q := UV^\top, \quad E := W_* - WQ,$$

则有

$$W^\top E = W^\top W_* - Q = U \Sigma V^\top - UV^\top = U(\Sigma - I_p) V^\top.$$

于是有

$$\begin{aligned}
 \Delta &= \sum_n \varepsilon_n U^\top \left[Z_n^\top - \frac{1}{2}(W^\top Z_n + Z_n^\top W)W^\top \right] (WU + EV) \\
 &= \sum_n \varepsilon_n U^\top \left[Z_n^\top W - \frac{1}{2}(W^\top Z_n + Z_n^\top W)W^\top W \right] U \\
 &\quad + \sum_n \varepsilon_n U^\top \left[Z_n^\top EV - \frac{1}{2}(W^\top Z_n + Z_n^\top W)W^\top EV \right] \\
 &= \sum_n \varepsilon_n U^\top \left[\frac{1}{2}Z_n^\top W - \frac{1}{2}W^\top Z_n \right] U \\
 &\quad + \sum_n \varepsilon_n U^\top \left[Z_n^\top EV - \frac{1}{2}(W^\top Z_n + Z_n^\top W)U(\Sigma - I_p) \right] \\
 &= \frac{1}{2} \sum_n \varepsilon_n U^\top \left[\frac{1}{2}Z_n^\top W - \frac{1}{2}W^\top Z_n \right] U \\
 &\quad + \frac{1}{2} \sum_n \varepsilon_n U^\top [W^\top Z_n + Z_n^\top W] U(I - \Sigma) \\
 &\quad + \sum_n \varepsilon_n U^\top Z_n^\top EV \\
 &= \sum_n \varepsilon_n \left(\Delta_n^{(0)} + \Delta_n^{(1)} + \Delta_n^{(2)} \right),
 \end{aligned}$$

其中

$$\begin{aligned}
 \Delta_n^{(0)} &= \frac{1}{2} U^\top \left[\frac{1}{2}Z_n^\top W - \frac{1}{2}W^\top Z_n \right] U, \\
 \Delta_n^{(1)} &= \frac{1}{2} U^\top [W^\top Z_n + Z_n^\top W] U(I - \Sigma), \\
 \Delta_n^{(2)} &= U^\top Z_n^\top EV.
 \end{aligned}$$

现在, 进一步令:

$$\tilde{W} = WU, \quad \tilde{W}_* = W_*V = WU + EV, \quad \tilde{E} = \tilde{W}_* - \tilde{W} = W_*V - WU, \quad \tilde{Z}_n := Z_nU. \quad (6.88)$$

对于 $\Delta_n^{(0)}$, 很容易发现:

$$\left(\Delta_n^{(0)} \right)^\top = -\Delta_n^{(0)},$$

因此 $\Delta_n^{(0)}$ 的对角线为零:

$$\text{Diag}(\Delta_n^{(0)}) = 0.$$

对于 $\Delta_n^{(1)}$, 由于 $\frac{1}{2}U^\top [W^\top Z_n + Z_n^\top W] U = \text{Sym}(U^\top W^\top Z_n U)$, 有:

$$\text{Diag}(\Delta_n^{(1)}) = \frac{1}{2} \text{Diag}(U^\top (W^\top Z_n + Z_n^\top W) U)(I - \Sigma)$$

$$\begin{aligned}
 &= \text{Diag}(U^\top W^\top Z_n U)(I - \Sigma) \\
 &= \text{Diag}((I - \Sigma)\tilde{W}^\top \tilde{Z}_n) \\
 &= \text{Diag}((\tilde{W}(I - \Sigma))^\top \tilde{Z}_n).
 \end{aligned}$$

最后一项 $\Delta_n^{(2)}$ 可以写为:

$$\begin{aligned}
 \text{Diag}(\Delta_n^{(2)}) &= \text{Diag}(U^\top Z_n^\top E V) = \text{Diag}((Z_n U)^\top (W_* - W Q) V) \\
 &= \text{Diag}((Z_n U)^\top (W_* V - W U)) \\
 &= \text{Diag}((\tilde{W}_* - \tilde{W})^\top \tilde{Z}_n).
 \end{aligned}$$

因此,

$$\begin{aligned}
 \text{Diag}(\Delta_n) &= \text{Diag}(\Delta_n^{(0)} + \Delta_n^{(1)} + \Delta_n^{(2)}) \\
 &= \text{Diag}((\tilde{W}(I - \Sigma))^\top \tilde{Z}_n) + \text{Diag}((\tilde{W}_* - \tilde{W})^\top \tilde{Z}_n) \\
 &= \text{Diag}[(\tilde{W}(I - \Sigma) + \tilde{W}_* - \tilde{W})^\top \tilde{Z}_n] \\
 &= \text{Diag}((\tilde{W}_* - \tilde{W} \Sigma)^\top \tilde{Z}_n).
 \end{aligned}$$

代回 W, W_* , 有:

$$\begin{aligned}
 \tilde{W}_* - \tilde{W} \Sigma &= W_* V - W U \Sigma = (W_* - W U \Sigma V^\top) V = (W_* - W P) V \\
 &= (W_* - W W^\top W_*) V \\
 &= P_W^\perp W_* V,
 \end{aligned}$$

其中 $P_W^\perp = I - W W^\top$ 是正交于 W 的投影。因此,

$$\text{Diag}(\Delta_n) = \text{Diag}((\tilde{W}_* - \tilde{W} \Sigma)^\top \tilde{Z}_n) = \text{Diag}(V^\top (P_W^\perp W_*)^\top Z_n U).$$

总而言之, 有:

$$\dot{\sigma}_i = (1 - \sigma_i^2) \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} 2r_i \sigma^{2r-e_i} v_r^2 + \sum_{n \in \mathbb{N}^d} \varepsilon_n \text{Diag}(V^\top (P_W^\perp W_*)^\top Z_n U)_{ii},$$

并且有:

$$\begin{aligned}
 \frac{d}{dt} \sigma_i^2 &= 4(1 - \sigma_i^2) \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} r_i \sigma^{2r} v_r^2 + \sum_{n \in \mathbb{N}^d} \varepsilon_n \text{Diag}(\Sigma V^\top (P_W^\perp W_*)^\top Z_n U)_{ii} \\
 &= (1 - \sigma_i^2) A_i + \sum_{n \in \mathbb{N}^d} \varepsilon_n \text{Diag}(\Sigma V^\top (P_W^\perp W_*)^\top Z_n U)_{ii},
 \end{aligned}$$

其中 $A_i = 4 \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} r_i v_r^2 \sigma^{2r}$ 。

ϕ_1 和 ω 的动力学 利用链式法则，可以计算 ϕ_1 和 ω 的动力学。只需关注噪声项。引入 $M = P^\top P$ 。对于 ϕ_1 ，噪声项写为：

$$\xi = \sum_{i=1}^p \sum_{n \in \mathbb{N}^d} \varepsilon_n \text{Diag}(\Sigma V^\top (P_W^\perp W_*)^\top Z_n U)_{ii} = \sum_{n \in \mathbb{N}^d} \varepsilon_n \text{Tr} [\Sigma V^\top (P_W^\perp W_*)^\top Z_n U].$$

而对于 ω ，利用链式法则及命题 6.32，有：

$$\begin{aligned} \zeta &= \frac{1}{\sum_{j=1}^p e^{-K\sigma_j^2}} \sum_{i=1}^p e^{-K\sigma_i^2} \sum_{n \in \mathbb{N}^d} \varepsilon_n \text{Diag}(\Sigma V^\top (P_W^\perp W_*)^\top Z_n U)_{ii} \\ &= \frac{1}{\sum_{j=1}^p e^{-K\sigma_j^2}} \sum_{i=1}^p \sum_{n \in \mathbb{N}^d} \varepsilon_n \text{Diag}(e^{-K\sigma_i^2} \Sigma V^\top (P_W^\perp W_*)^\top Z_n U)_{ii} \\ &= \frac{1}{\sum_{j=1}^p e^{-K\sigma_j^2}} \sum_{n \in \mathbb{N}^d} \varepsilon_n \text{Tr} [e^{-K\Sigma^2} \Sigma V^\top (P_W^\perp W_*)^\top Z_n U], \end{aligned}$$

其中 $e^{-K\Sigma^2}$ 表示矩阵函数。此外，分母可以写为：

$$\sum_{j=1}^p e^{-K\sigma_j^2} = \text{Tr} e^{-K\Sigma^2} = \text{Tr} e^{-KM}.$$

其与 P 的 SVD 分解无关。

因此，考虑一般的函数 φ 和形式如下的噪声项：

$$\chi = \sum_{n \in \mathbb{N}^d} \varepsilon_n \text{Tr} [\varphi(\Sigma^2) \Sigma V^\top (P_W^\perp W_*)^\top Z_n U] = \sum_{n \in \mathbb{N}^d} \varepsilon_n T_n.$$

那么，可以将 T_n 表示为：

$$\begin{aligned} T_n &= \text{Tr} [\varphi(\Sigma^2) \Sigma V^\top (P_W^\perp W_*)^\top Z_n U] \\ &\stackrel{(a)}{=} \text{Tr} [U \varphi(\Sigma^2) \Sigma V^\top (P_W^\perp W_*)^\top \langle \nabla_W (P_W g), H_n \rangle] \\ &\stackrel{(b)}{=} \text{Tr} [U \Sigma \varphi(\Sigma^2) V^\top (P_W^\perp W_*)^\top \langle x(P_W \nabla g)^\top, H_n \rangle] \\ &\stackrel{(c)}{=} \langle \text{Tr} [P \varphi(M) (P_W^\perp W_*)^\top x(P_W \nabla g)^\top], H_n \rangle \\ &\stackrel{(d)}{=} \langle (P_W \nabla g)^\top W^T W_* \varphi(M) (P_W^\perp W_*)^\top x, H_n \rangle \\ &\stackrel{(e)}{=} \langle (\nabla_x (P_W g))^\top W_* \varphi(M) (P_W^\perp W_*)^\top x, H_n \rangle \\ &= \langle (\nabla_x (P_W g))^\top W_* \varphi(M) W_*^\top P_W^\perp x, H_n \rangle, \end{aligned}$$

此处 (a, d) 中使用了迹的交换性，在 (b, e) 中使用了梯度公式引理 6.1，在 (c) 中使用了矩阵微积分 (6.67)。在此注意到，最终结果与奇异值分解无关。

总结 在以下命题中总结这些结果：

命题 6.24: 考虑假设 6.3 下的动力学 (6.17)。则有:

$$\begin{aligned}\dot{h}_r &= \mu_{2r} [\phi_r h_r^* - h_r + e_r], \\ \dot{\omega} &= \frac{1}{\sum_{j=1}^p e^{-K\sigma_j^2}} \sum_{i=1}^p e^{-K\sigma_i^2} (1 - \sigma_i^2) A_i + \zeta \\ \dot{\phi}_1 &= \sum_{i=1}^p (1 - \sigma_i^2) A_i + \xi,\end{aligned}\tag{6.89}$$

其中 $A_i = 4 \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} r_i v_r^2 \sigma^{2r}$, 且噪声项给定为:

$$\begin{aligned}e_r &= e_r(W) = \sum_{n \in \mathbb{N}^d} \varepsilon_n \langle P_W \bar{H}_r, H_n \rangle_{\gamma_d}, \\ \zeta &= \zeta(W, g) = \frac{1}{\text{Tr} e^{-KM}} \sum_{n \in \mathbb{N}^d} \varepsilon_n \langle (\nabla_x (P_W g))^{\top} W_* e^{-KM} W_*^{\top} P_W^{\perp} x, H_n \rangle, \\ \xi &= \xi(W, g) = \sum_{n \in \mathbb{N}^d} \varepsilon_n \langle (\nabla_x (P_W g))^{\top} W_* W_*^{\top} P_W^{\perp} x, H_n \rangle,\end{aligned}$$

其中 $M = P^{\top} P$ 。

6.5.4.2 扰动项的界

现在应用引理 6.4 来估计命题 6.24 中的扰动项。

命题 6.25: 设 e_r 定义如命题 6.24。那么,

$$\kappa(W, W') := \text{Cov}(e_r(W), e_r(W')) = \frac{1}{n} \sum_{|r|=r} v_r^2 \lambda^{2r},$$

其中 λ 为 $W^{\top} W'$ 的奇异值。此外, $\kappa(W, W')$ 关于 $\mathbb{R}^{d \times p}$ 中的欧几里得范数是 Lipschitz 连续的, 其 Lipschitz 常数为 $2r$ 。因此, 至少以概率 $1 - 4e^{-d}$, 有

$$\sup_{W \in \mathcal{S}(d, p)} |e_r(W)| \lesssim \sqrt{\frac{dp \log r}{n}}, \quad \forall r \geq 0.\tag{6.90}$$

证明 根据引理 6.3, 回顾 $\bar{H}_r$ 是一个 $2r$ 次的旋转不变多项式。可以计算 $e_r(W)$ 的协方差函数:

$$\begin{aligned}\kappa(W, W') &= \text{Cov}(e_r(W), e_r(W')) = \frac{1}{n} \sum_{n \in \mathbb{N}^d} \langle P_W \bar{H}_r, H_n \rangle_{\gamma_d} \langle P_{W'} \bar{H}_r, H_n \rangle_{\gamma_d} \\ &= \frac{1}{n} \langle P_W \bar{H}_r, P_{W'} \bar{H}_r \rangle_{\gamma_d} = \frac{1}{n} \langle \bar{H}_r, P_W^{\top} P_{W'} \bar{H}_r \rangle_{\gamma_p} \\ &= \frac{1}{n} \langle \bar{H}_r, \mathcal{A}_{W^{\top} W'} \bar{H}_r \rangle_{\gamma_p}.\end{aligned}$$

考虑 SVD 分解 $W^{\top} W' = Q_1 \Lambda Q_2^{\top}$, 令 λ 为 Λ 的对角线元素。那么,

$$\kappa(W, W') = \frac{1}{n} \langle \bar{H}_r, \mathcal{A}_{Q_1 \Lambda Q_2^{\top}} \bar{H}_r \rangle_{\gamma_p} = \frac{1}{n} \langle \bar{H}_r, \mathcal{A}_{Q_1} \mathcal{A}_{\Lambda} \mathcal{A}_{Q_2^{\top}} \bar{H}_r \rangle_{\gamma_p}$$

$$\begin{aligned}
 &= \frac{1}{n} \langle \bar{H}_r, \mathcal{A}_\Lambda \bar{H}_r \rangle_{\gamma_p} \\
 &= \frac{1}{n} \sum_{|r|=r} v_r^2 \lambda^{2r}.
 \end{aligned}$$

现在将 κ 视为定义在 $\{W \in \mathbb{R}^{d \times p} : \|W\| \leq 1\}$ 上的二元函数。可以计算 κ 关于 W 的导数。取切方向 $H \in \mathbb{R}^{d \times p}$, 利用 (6.66) 得到

$$D_H \Lambda = \text{Diag}(U^\top D_H(W^\top W') V) = \text{Diag}(U^\top H^\top W' V), \quad D_H \lambda_j = u_j^\top H^\top W' v_j,$$

因此

$$|D_H \lambda_j| \leq \|H\| \|W'\| \leq \|H\|.$$

接着, 利用

$$D_H \lambda^{2r} = \sum_{j=1}^p 2r_j \lambda^{2r-e_j} D_H \lambda_j,$$

有

$$|D_H \lambda^{2r}| \leq 2 \sum_{j=1}^p r_j \lambda^{2r-e_j} |D_H \lambda_j| \leq 2 \sum_{j=1}^p r_j \|H\| = 2|r| \|H\|.$$

最后,

$$\begin{aligned}
 |D_H k(W, W')| &= \left| \sum_{|r|=r} v_r^2 D_H \lambda^{2r} \right| \leq \sum_{|r|=r} v_r^2 |D_H \lambda^{2r}| \\
 &\leq 2 \sum_{|r|=r} v_r^2 |r| \|H\| \leq 2r \|H\| \sum_{|r|=r} v_r^2 \\
 &= 2r \|H\| \leq 2r \|H\|_2.
 \end{aligned}$$

因此有

$$\|Dk(W, W')\|_2 \leq 2r,$$

这意味着 $k(W, W')$ 关于欧几里得范数是 Lipschitz 的。应用引理 6.4 就得到了所需的界。 ■

在估计 ξ 和 ζ 的界之前, 先进行一些初步计算。考虑形式如下的噪声项:

$$\chi = \sum_{n \in \mathbb{N}^d} \varepsilon_n \langle (\nabla_x (P_W \bar{H}_r))^\top W_* \varphi(M) W_*^\top P_W^\perp x, H_n \rangle.$$

可以将 g 分解为

$$g = \sum_{r \geq 0} \sum_{m=2r, |r|=r} g_m H_m = \sum_{r \geq 0} h_r \bar{H}_r, \quad \bar{H}_r = \sum_{|r|=r} v_r H_{2r}.$$

那么, 有

$$\chi = \sum_{r \geq 0} h_r \chi_r, \quad \chi_r = \sum_{n \in \mathbb{N}^d} \varepsilon_n \langle (\nabla_x (P_W \bar{H}_r))^{\top} W_* \varphi(M) W_*^{\top} P_W^{\perp} x, H_n \rangle,$$

其中 χ_r 与 g 无关。

取 W' 并记相应的量为带撇符号 (例如 P', M')。计算 $\chi_r(W)$ 与 $\chi_r(W')$ 的协方差:

$$\begin{aligned} k_{\chi,r}(W, W') &= \text{Cov}(\chi_r(W), \chi_r(W')) \\ &= \langle (\nabla_x (P_W \bar{H}_r))^{\top} W_* \varphi(M) W_*^{\top} P_W^{\perp} x, (\nabla_x (P_{W'} \bar{H}_r))^{\top} W_* \varphi(M') W_*^{\top} P_{W'}^{\perp} x \rangle_{\gamma_d}. \end{aligned}$$

此处利用了 $(H_n)_{n \in \mathbb{N}^d}$ 是一组正交基的事实。注意到

$$P_W^{\perp} \nabla_x (P_W \bar{H}_r) = P_W^{\perp} W P_W (\nabla \bar{H}_r) = 0,$$

可以对 $f = P_W \bar{H}_r$ 和 $A = W_* \varphi(M) W_*^{\top} P_W^{\perp}$ 使用引理 6.6 得到

$$k_{\chi,r}(W, W') = k_{\chi,r}^{(1)}(W, W') + k_{\chi,r}^{(2)}(W, W'),$$

其中

$$\begin{aligned} k_{\chi,r}^{(1)}(W, W') &= \mathbb{E} (\nabla_x (P_W \bar{H}_r))^{\top} W_* \varphi(M) W_*^{\top} P_W^{\perp} (W_* \varphi(M') W_*^{\top} P_{W'}^{\perp})^{\top} \nabla_x (P_{W'} \bar{H}_r) \\ &= \mathbb{E} (P_W \nabla_x \bar{H}_r)^{\top} W^{\top} W_* \varphi(M) W_*^{\top} P_W^{\perp} P_{W'}^{\perp} W_* \varphi(M') W_*^{\top} W' (P_{W'} \nabla_x \bar{H}_r) \\ &= \mathbb{E} (P_W \nabla \bar{H}_r)^{\top} P \varphi(M) W_*^{\top} P_W^{\perp} P_{W'}^{\perp} W_* \varphi(M') (P')^{\top} (P_{W'} \nabla \bar{H}_r), \end{aligned}$$

以及

$$\begin{aligned} k_{\chi,r}^{(2)}(W, W') &= \mathbb{E} \text{Tr} \nabla_x^2 (P_W \bar{H}_r) (W_* \varphi(M) W_*^{\top} P_W^{\perp}) \nabla_x^2 (P_{W'} \bar{H}_r) (W_* \varphi(M') W_*^{\top} P_{W'}^{\perp}) \\ &= \mathbb{E} \text{Tr} W (P_W \nabla^2 \bar{H}_r) W^{\top} W_* \varphi(M) W_*^{\top} P_W^{\perp} W' (P_{W'} \nabla^2 \bar{H}_r) (W')^{\top} W_* \varphi(M') W_*^{\top} P_{W'}^{\perp}. \end{aligned}$$

可以进一步计算方差。由于当 $W = W'$ 时, $P_W^{\perp} W' = 0$, 因此

$$\begin{aligned} k_{\chi,r}(W, W) &= k_{\chi,r}^{(1)}(W, W) = \mathbb{E} (P_W \nabla \bar{H}_r)^{\top} P \varphi(M) W_*^{\top} P_W^{\perp} W_* \varphi(M) P^{\top} (P_W \nabla \bar{H}_r) \\ &= \mathbb{E} (P_W \nabla \bar{H}_r)^{\top} P \varphi(M) (I - M) \varphi(M) P^{\top} (P_W \nabla \bar{H}_r) \\ &\stackrel{(a)}{=} \text{Tr} P \varphi(M) (I - M) \varphi(M) P [\mathbb{E} (\nabla \bar{H}_r) (\nabla \bar{H}_r)^{\top}] \\ &= \frac{2r}{p} \text{Tr} P \varphi(M) (I - M) \varphi(M) P^{\top} \\ &= \frac{2r}{p} \text{Tr} \varphi(M)^2 M (I - M). \end{aligned}$$

命题 6.26: 至少以概率 $1 - 4e^{-d}$, 对于任何满足 $\phi_1(W) \leq 1 - \nu$ 的 $W \in S(d, p)$ 以

及任何 g , 有

$$|\xi(W, g)| \lesssim (1 - \phi_1)^{\frac{1}{2}} \sqrt{\frac{pd}{n}} \sqrt{\log(p\nu^{-1})} \sum_{r \geq 1} (r \log r)^{\frac{1}{2}} |h_r|. \quad (6.91)$$

证明 在之前的计算中取 $\varphi \equiv 1$, 有

$$k_{\xi, r}(W, W) = \frac{2r}{p} \text{Tr}(I - \Sigma^2) \Sigma^2 \leq \frac{2r}{p} \text{Tr}(I - \Sigma^2) = 2r(1 - \phi_1).$$

为了估计 $k_{\xi, r}(W, W')$ 导数的界, 可以迭代地使用 $\text{Tr } A \leq p \|A\|$ 和导数的链式法则。量 $P = W^\top W_*$ 和 $P_W^\top = I - WW^\top$ 都是关于 W 的多项式。此外, 回顾 $\nabla_W(P_W h) = x P_W (\nabla g)^\top$ 。结合 Hermite 多项式的导数公式以及 $\|W\| \leq 1, P_W(\nabla^2 \bar{H}_r)$ 中每个项的导数都被一个最高次数为 $2r + 1$ 且系数规模为 r 的多项式所控制, 因此其 $L^2(\gamma_d)$ 范数也由 r 的多项式控制。由此可以推导出 $\|\nabla_W k_{\xi, r}(W, W')\|_2$ 被 p, r 的某个多项式控制。此外, 由于

$$\phi_1 = \sum_{i=1}^p \sigma_i^2 = \text{Tr } P^\top P,$$

ϕ_1 的导数也由 p, r 的多项式控制。令 $\bar{\xi} = \frac{1}{2r}(1 - \phi_1)^{-\frac{1}{2}} \xi$ 以及缩放后的协方差

$$\bar{k}_{\xi, r}(W, W') := \frac{1}{2r} [(1 - \phi_1)(1 - \phi_1')]^{-\frac{1}{2}} k_{\xi, r}(W, W').$$

那么, 只要 $1 - \phi_1 \geq \nu$, $\bar{k}_{\xi, r}(W, W')$ 的导数就由 p, r, ν^{-1} 的某个多项式控制。最后, 可以应用引理 6.4 结合联合概率界得到

$$|\bar{\xi}_r| \lesssim \sqrt{\log(rp\nu^{-1})} + C\sqrt{pd}, \quad \forall r \geq 1$$

以至少 $1 - 4e^{-d}$ 的概率成立。回到 ξ_r 并对 r 求和即可得到所需的界。 ■

命题 6.27: 至少以概率 $1 - 4e^{-d}$, 对于任何 $W \in S(d, p)$ 和任何 g , 有

$$|\zeta(W, g)| \lesssim \sqrt{\frac{pd}{n}} \sqrt{\log(pK)} \sum_{r \geq 1} (r \log r)^{\frac{1}{2}} |h_r|. \quad (6.92)$$

证明 在之前的计算中取 $\varphi(t) = e^{-Kt}$, 有

$$k_{\zeta, r}(W, W') = k_{\zeta, r}^{(1)}(W, W') + k_{\zeta, r}^{(2)}(W, W'),$$

其中

$$\begin{aligned} k_{\zeta, r}^{(1)}(W, W') &= q \mathbb{E} (P_W \nabla \bar{H}_r)^\top P e^{-KM} W_*^\top P_W^\perp P_{W'}^\perp W_* e^{-KM'} (P')^\top (P_{W'} \nabla \bar{H}_r) \\ k_{\zeta, r}^{(2)}(W, W') &= q \mathbb{E} \text{Tr } W (P_W \nabla^2 \bar{H}_r) W^\top W_* e^{-KM} W_*^\top P_W^\perp W' (P_{W'} \nabla^2 \bar{H}_r) (W')^\top W_* e^{-KM'} W_*^\top P_{W'}^\perp, \\ q &= \left[(\text{Tr } e^{-KM}) (\text{Tr } e^{-KM'}) \right]^{-1}. \end{aligned}$$

对于 $W = W'$ 的情况, 有

$$\begin{aligned} k_{\zeta,r}(W, W') &= \frac{2r}{p} (\text{Tr} e^{-KM})^{-2} \text{Tr} e^{-2KM} M(I - M) \\ &= \frac{2r}{p} (\text{Tr} e^{-K\Sigma^2})^{-2} \text{Tr} e^{-2K\Sigma^2} \Sigma^2(I - \Sigma^2) \\ &\leq \frac{2r}{p} (\text{Tr} e^{-K\Sigma^2})^{-2} \text{Tr} e^{-2K\Sigma^2} \\ &\leq 2r. \end{aligned}$$

对于导数, 可以使用与前一个证明相同的参数。唯一的区别是必须额外考虑 $f(M) = (\text{Tr} e^{-KM})^{-1} e^{-KM}$ 的矩阵导数。有

$$\frac{d}{dt} f(M) = (\text{Tr} e^{-KM})^{-1} \frac{d}{dt} e^{-KM} - (\text{Tr} e^{-KM})^{-2} \left(\frac{d}{dt} \text{Tr} e^{-KM} \right) e^{-KM} = I + II.$$

第二项很容易估计。利用矩阵导数和迹的性质, 有

$$\frac{d}{dt} \text{Tr} e^{-KM} = -K \text{Tr} e^{-KM} \dot{M},$$

因此

$$\begin{aligned} \left\| (\text{Tr} e^{-KM})^{-2} \left(\frac{d}{dt} \text{Tr} e^{-KM} \right) e^{-KM} \right\| &= K (\text{Tr} e^{-KM})^{-2} |\text{Tr} e^{-KM} \dot{M}| \|e^{-KM}\| \\ &\leq K (\text{Tr} e^{-KM})^{-2} \text{Tr} e^{-KM} \|\dot{M}\| \text{Tr} e^{-KM} \\ &= K \|\dot{M}\|, \end{aligned}$$

此处利用了对于半正定矩阵 A , $|\text{Tr} AB| \leq \|AB\|_1 \leq \text{Tr} A \|B\|$ 的事实, 其中 $\|\cdot\|_1$ 是迹范数。对于第一项, 使用矩阵导数公式:

$$\frac{d}{dt} \phi(M) = Q(R \odot (Q^\top \dot{M} Q)) Q^\top,$$

其中 $M = Q\Lambda Q^\top$ 是 M 的谱分解,

$$R_{ij} = \begin{cases} \frac{\phi(\lambda_i) - \phi(\lambda_j)}{\lambda_i - \lambda_j}, & \text{若 } \lambda_i \neq \lambda_j, \\ \phi'(\lambda_i), & \text{若 } \lambda_i = \lambda_j. \end{cases}$$

在这里, 注意到 $\phi(t) = e^{-Kt}$, $\phi'(t) = -Ke^{-Kt}$, 因此利用中值定理有

$$R_{ij} = \phi'(\xi_{ij}) = -Ke^{-K\xi_{ij}}, \quad |R_{ij}| \leq Ke^{-K\lambda_{\min}},$$

其中 ξ_{ij} 介于 λ_i 和 λ_j 之间, 且 $\lambda_{\min} = \min_i \lambda_i$ 。由此,

$$\begin{aligned} \left\| \frac{d}{dt} e^{-KM} \right\|_2 &= \|Q(R \odot (Q^\top \dot{M} Q)) Q^\top\|_2 = \|R \odot (Q^\top \dot{M} Q)\|_2 \\ &\leq Ke^{-K\lambda_{\min}} \|Q^\top \dot{M} Q\|_2 \leq Ke^{-K\lambda_{\min}} \|\dot{M}\|_2, \end{aligned}$$

进而

$$\|I\| \leq \|I\|_2 \leq K \frac{e^{-K\lambda_{\min}}}{\text{Tr} e^{-KM}} \|\dot{M}\|_2 \leq K \|\dot{M}\|_2.$$

总之, 有

$$\left\| \frac{d}{dt} f(M) \right\| \leq 2K \|\dot{M}\|_2.$$

最后, 可以断定 $k_{\zeta,r}(W, W')$ 的导数由 p, r 的多项式控制。应用引理 6.4 结合联合概率界, 就给出了所需的上界。 ■

6.5.4.3 训练动力学

命题 6.28: 在假设 6.4 下, 假设对于某个 $s > 0$, 有 $n \gtrsim d^{2m_0+1+s}$ 。设 (6.90) 和 (6.92) 成立。令 $\delta > 0$ 为定值。那么, 当 n, d 足够大时, 至少以概率 $1 - \delta$, 有

$$\min_{1 \leq i \leq p} \sigma_i^2(t) \geq 1/2, \quad \forall t \in [T^{\text{app}}, T^{\text{fin}}], \quad (6.93)$$

其中 $T^{\text{app}} \lesssim \log d + d^{m_0-1}$, 且 T^{fin} 可取为 d 的任意固定多项式, 且 $\lesssim$ 符号中的常数可以取决于 δ 。

证明 证明遵循与命题 6.9 和命题 6.22 类似的策略。不失一般性地, 可以假设对于所有 r 都有 $h_r \geq 0$ 。

首先, 根据命题 6.14, 存在某个 $c = c(\delta)$ 使得至少以概率 $1 - \delta$ 有:

$$\min_{1 \leq i \leq p} \sigma_i^2 \geq cd^{-1}.$$

取 (6.55) 中的 ω 和 ϕ_1 。根据命题 6.32, 通过取 $K \asymp Cd \log p$, 有:

$$\omega(0) \geq cd^{-1} - \frac{1}{K} \log p \geq cd^{-1}.$$

此外, 在这里断言在所关注的时间范围内, 有 $\omega(t) \geq \rho_0 := \frac{1}{2}cd^{-1}$, 将在稍后证明这一断言。

回顾命题 6.24 中的动力学。令 $r_0 = m_0/2$ 。注意到当 $\omega(t) \geq \rho_0$ 时, 由于 $n \gtrsim d^{2m_0+1+s}$, 有:

$$\phi_{r_0}(t) \geq \omega(t)^{r_0} \gtrsim d^{-r_0} \gtrsim \sqrt{\frac{d}{n}}.$$

因此, 有:

$$\dot{h}_{r_0} = \mu_{2r_0}(\phi_{r_0} h_{r_0}^* - h_{r_0} + e_{r_0}) \geq \mu_{2r_0}(\phi_{r_0} h_{r_0}^*/2 - h_{r_0}), \quad (6.94)$$

从而:

$$T_0^g := \inf \left\{ t \geq 0 : h_{r_0}(t) \geq \frac{1}{4} \rho_0^{r_0} h_{r_0}^* \right\} \lesssim 1.$$

现在证明在 $t \leq T_0^g$ 时的断言。回顾命题 6.24 中 ω 的动力学：

$$\dot{\omega} = \frac{1}{\sum_{j=1}^p e^{-K\sigma_j^2}} \sum_{i=1}^p e^{-K\sigma_i^2} (1 - \sigma_i^2) A_i + \zeta,$$

其中 $A_i = 4 \sum_{r \geq 1} h_r h_r^* \sum_{|r|=r} r_i v_r^2 \sigma^{2r}$ 。类似于命题 6.8，可以证明：

$$h_r(t) \geq -C \min(1, \mu_{2r} t) |e_r| \geq -C \min(1, \mu_{2r} t) \sqrt{\frac{d \log r}{n}}.$$

同样，类似于 (6.51)，有：

$$\sum_{r \geq 1} (r \log r)^{\frac{1}{2}} |h_r| \lesssim (1 + \log^+ t)^2, \quad (6.95)$$

因此误差项 (6.92) 进一步被界定为：

$$|\zeta(W, g)| \lesssim \sqrt{\frac{pd}{n}} \sqrt{\log(pK)} \sum_{r \geq 1} (r \log r)^{\frac{1}{2}} |h_r| \lesssim \sqrt{\frac{d \log d}{n}} (1 + \log^+ t)^2. \quad (6.96)$$

因此，

$$\begin{aligned} A_i &\geq - \sum_{r \geq 1} h_r^* \min(1, \mu_{2r} t) \sqrt{\frac{d \log r}{n}} \sum_{|r|=r} r_i v_r^2 \sigma^{2r} \\ &\geq -\sqrt{d/n} \sum_{r \geq 1} h_r^* r (\log r)^{\frac{1}{2}} \min(1, \mu_{2r} t) \\ &\geq -\sqrt{d/n} (1 + \log^+ t)^2, \end{aligned} \quad (6.97)$$

其中最后一步使用了 $\mu_r \lesssim e^{-\gamma r}$ 和命题 6.31。将这两个界限代入 $\dot{\omega}$ 的动力学，有：

$$\begin{aligned} \dot{\omega} &\gtrsim - \frac{1}{\sum_{j=1}^p e^{-K\sigma_j^2}} \sum_{i=1}^p e^{-K\sigma_i^2} \sqrt{d/n} (1 + \log^+ t)^2 - \sqrt{\frac{d \log d}{n}} (1 + (\log^+ t)^2) \\ &\gtrsim -\sqrt{\frac{d \log d}{n}} (1 + \log^+ t)^2. \end{aligned}$$

因此，由于 $T_0^g \lesssim 1$ 且 $n \gtrsim d^{2m_0+1+s}$ ，有：

$$\omega(T_0^g) \geq \omega(0) - \sqrt{\frac{d \log d}{n}} T_0^g \geq cd^{-1} - \sqrt{\frac{d \log d}{n}} \geq \rho_0,$$

这证明了直到 T_0^g 为止的断言。

在此之后，有：

$$\begin{aligned} h_{r_0} h_{r_0}^* \sum_{|r|=r_0} r_i v_r^2 \sigma^{2r} &\geq \rho_0^{r_0} (h_{r_0}^*)^2 \sum_{|r|=r_0} r_i v_r^2 \rho_0^{r_0} \\ &= c(h_{r_0}^*)^2 \rho_0^{2r_0} \gtrsim d^{-2r_0}, \end{aligned}$$

因此既然 $n \gtrsim d^{2m_0+1+s}$, 有:

$$\begin{aligned} A_i &= 4h_{r_0}h_{r_0}^* \sum_{|r|=r_0} r_i v_r^2 \sigma^{2r} + 4 \sum_{r \neq r_0} h_r h_r^* \sum_{|r|=r} r_i v_r^2 \sigma^{2r} \\ &\geq 4h_{r_0}h_{r_0}^* \sum_{|r|=r_0} r_i v_r^2 \sigma^{2r} - \sqrt{d/n}(1 + \log^+ t)^2 \\ &\geq 2h_{r_0}h_{r_0}^* \sum_{|r|=r_0} r_i v_r^2 \sigma^{2r}, \end{aligned} \quad (6.98)$$

前提是 t 最多为 d 和 n 的多项式量级。因此,

$$\dot{\omega} \geq \frac{2}{\sum_{j=1}^p e^{-K\sigma_j^2}} \sum_{i=1}^p e^{-K\sigma_i^2} (1 - \sigma_i^2) h_{r_0} h_{r_0}^* \sum_{|r|=r_0} r_i v_r^2 \sigma^{2r} \geq 0. \quad (6.99)$$

于是, 结合 (6.94) 和 (6.99), 可以应用与命题 6.22 相同的分析方法。定义类似的时间 T_k^g 和 T_k^ρ , 有 $T^{\text{app}} = T_L^\rho \lesssim \log d + d^{m_0-1}$, 其中 $L = 1 + \lceil \log_2 \rho_0^{-1} \rceil$. ■

命题 6.29: 在假设 6.4 下, 假设对于某个 $s > 0$, 有 $n \gtrsim d^{2m_0+1+s}$. 设 (6.90)、(6.91) 成立, 且取 $v = 1/n$. 假设 (6.93) 成立。那么, 存在某个 $t_1 \leq T^{\text{app}} + C$ 使得:

$$1 - \phi_1(t_1 + s) \lesssim e^{-cs},$$

前提是:

$$1 - \phi_1(t_1 + s) \gtrsim \frac{pd \cdot \text{polylog}(n, d, p)}{n}.$$

证明 首先, 利用动力学 (6.94), 当 $t_1 = T^{\text{app}} + C$ 之后, 有 $h_{r_0} \geq ch_{r_0}^*$. 那么, 利用 (6.98) 和 (6.93), 有:

$$A_i \gtrsim h_{r_0}h_{r_0}^* \sum_{|r|=r_0} r_i v_r^2 \sigma^{2r} \gtrsim 1.$$

于是, 命题 6.24 中 ϕ_1 的动力学给出:

$$\dot{\phi}_1 \geq c \sum_{i=1}^p (1 - \sigma_i^2) - |\xi| = c(1 - \phi_1) - |\xi|.$$

另一方面, 结合 (6.91) 和 (6.95), 以及考虑到时间最多为 d 的多项式量级, 有:

$$|\xi| \lesssim (1 - \phi_1)^{\frac{1}{2}} \sqrt{\frac{dpp \text{polylog}(n, d)}{n}}.$$

因此, 只要 $1 - \phi_1 \gtrsim \frac{dpp \text{polylog}(n, d, p)}{n}$, 就有 $|\xi| \leq c(1 - \phi_1)/2$, 从而 $\dot{\phi}_1 \geq c(1 - \phi_1)/2$, 得出所需结果。■

定理 6.4 的证明 初始特征误差度量是命题 6.14 的直接推论。对于特征误差度量的衰减, 应用命题 6.28 和命题 6.29, 并结合误差界命题 6.25、命题 6.26 和命题 6.27。对于最终的特征误差度量, 可以使用命题 6.19 以及 $1 - \phi_1 \leq v$ 意味着 $1 - \min_j \sigma_j^2 \leq pv$

的事实。 ■

6.6 辅助结果

6.6.1 随机过程的控制

定义 6.1: 令 $X(\cdot)$ 为度量空间 (T, d) 上的一个随机过程。如果满足以下条件, 称 $X(\cdot)$ 是 σ^2 -次高斯 (σ^2 -sub-Gaussian) 的:

$$\|X(t) - X(s)\|_{\psi_2}^2 \leq \sigma^2 d(s, t)^2, \quad \text{对于所有 } s, t \in T \text{ 均成立.}$$

定理 6.5 (Dudley 熵积分): 令 X 为度量空间 (T, d) 上的一个 σ^2 -次高斯随机过程, 且 $\mathbb{E}X(t) = 0$ 。令 $\mathcal{N}(\epsilon, T, d)$ 为 T 关于 d 的覆盖数 (covering number)。令 $R = \text{diam}(T, d)$ 为 T 关于度量 d 的直径, 定义 Dudley 熵积分为:

$$I = \int_0^R \sqrt{\log \mathcal{N}(\epsilon, T, d)} \, d\epsilon. \quad (6.100)$$

则有:

$$\mathbb{E} \sup_{t \in T} X(t) \leq C\sigma I. \quad (6.101)$$

此外, 对于任何 $u \geq 0$, 以至少 $1 - 2e^{-u^2}$ 的概率, 有:

$$\sup_{s, t \in T} |X(s) - X(t)| \leq C\sigma (I + Ru), \quad (6.102)$$

现在考虑一个均值为零的高斯过程 $X(t), t \in T$, 其中 $T \subseteq \mathbb{R}^p$ 。 $X(\cdot)$ 的协方差函数由 $k(s, t) = \mathbb{E}X(s)X(t)$ 给出。诱导度量 (induced metric) 由下式给出:

$$d(s, t) = \sqrt{k(s, s) + k(t, t) - 2k(s, t)}.$$

关于 T 相对于 d 的覆盖数, 有以下结果。

引理 6.4: 令 $X(t), t \in T \subseteq \mathbb{R}^p$ 为一个均值为零的高斯过程, 其协方差函数为 $k(s, t)$ 。假设对于所有 $t \in T$, $k(t, t) \leq 1$, 且 k 在对角线上是 Hölder 连续的, 指数为 α :

$$|k(t, t) - k(s, s)| \leq L \|t - s\|^\alpha, \quad \forall s, t \in T.$$

令 $R = \text{diam}(T, \|\cdot\|)$ 为 T 关于范数 $\|\cdot\|$ 的直径。则有:

$$\mathbb{E} \sup_{t \in T} X(t) \leq C\alpha^{-1/2} (\sqrt{\log R + \log L + C}) \sqrt{p}.$$

此外, 对于任何 $u \geq 0$, 以至少 $1 - 4e^{-u^2}$ 的概率, 有:

$$\sup_{t \in T} |X(t)| \leq C\alpha^{-1/2} (\sqrt{\log R + \log L + C}) \sqrt{p} + Cu.$$

证明 用 $d(s, t)$ 表示上述定义的诱导度量。由于 $k(t, t) \leq 1$, 核函数的基本性质给出

$|k(s, t)| \leq \sqrt{k(s, s)k(t, t)} \leq 1$, 故 $d(s, t) \leq 2$, 即 $\text{diam}(T, d) \leq 2$ 另一方面,

$$\begin{aligned} d(s, t) &= \sqrt{k(s, s) + k(t, t) - 2k(s, t)} \leq \sqrt{|k(s, s) - k(s, t)| + |k(t, t) - k(s, t)|} \\ &\leq \sqrt{2L \|t - s\|^\alpha} = \sqrt{2L} \|t - s\|^{\alpha/2}. \end{aligned}$$

这表明:

$$\mathcal{N}(\epsilon, T, d) \leq \mathcal{N}((\epsilon/\sqrt{2L})^{2/\alpha}, T, \|\cdot\|).$$

现在, 利用 $\mathbb{R}^p$ 覆盖数的标准结果: $\mathcal{N}(\delta, T, \|\cdot\|) \leq (CR/\delta)^p$, 因此:

$$\mathcal{N}((\epsilon/\sqrt{2L})^{2/\alpha}, T, \|\cdot\|) \leq (CR/(\epsilon/\sqrt{2L})^{2/\alpha})^p = (CR(2L)^{1/\alpha})^p \epsilon^{-2p/\alpha}.$$

将此代入 Dudley 熵积分, 有:

$$\begin{aligned} \int_0^2 \sqrt{\log \mathcal{N}(\epsilon, T, d)} \, d\epsilon &\leq \int_0^2 \sqrt{\log (CRL^{1/\alpha})^p (\epsilon/\sqrt{2})^{-2p/\alpha}} \, d\epsilon \\ &= \int_0^2 \sqrt{p(\log(CR) + \alpha^{-1} \log(2L)) + \frac{2p}{\alpha} \log(\sqrt{2}/\epsilon)} \, d\epsilon \\ &\leq \int_0^2 \left[\sqrt{p(\log(CR) + \alpha^{-1} \log(2L))} + \sqrt{\frac{2p}{\alpha} \log(\sqrt{2}/\epsilon)} \right] \, d\epsilon \\ &\lesssim \sqrt{p(\log(CR) + \alpha^{-1} \log(2L))} + \sqrt{2p/\alpha} \\ &\lesssim \alpha^{-1/2} (\sqrt{\log R + \log L + C}) \sqrt{p}. \end{aligned}$$

■

6.6.2 高斯分布

如下的高斯分部积分公式是熟知的:

$$\mathbb{E} X_i h(X) = \mathbb{E} \partial_i h(X), \quad \mathbb{E} X_i X_j h(X) = \delta_{ij} \mathbb{E} h(X) + \mathbb{E} \partial_{ij} h(X).$$

引理 6.5: 令 $X \sim N(0, I_d)$ 且 $u, v \in \mathbb{R}^d$. 令 $f, g: \mathbb{R}^d \rightarrow \mathbb{R}$ 为二阶可微函数, 且其二阶以内的导数均平方可积. 那么, 有

$$\begin{aligned} \mathbb{E} [\langle u, X \rangle \langle v, X \rangle f(X) g(X)] &= \langle u, v \rangle \mathbb{E} [f(X) g(X)] \\ &\quad + \mathbb{E} [\langle u, \nabla f(X) \rangle \langle v, \nabla g(X) \rangle] + \mathbb{E} [\langle v, \nabla f(X) \rangle \langle u, \nabla g(X) \rangle] \\ &\quad + \mathbb{E} [f(X) u^\top \nabla^2 g(X) v] + \mathbb{E} [g(X) v^\top \nabla^2 f(X) u]. \end{aligned}$$

证明 从 $\mathbb{E}[(u^T X) f(X) (v^T X) g(X)]$ 出发, 首先将 $(u^T X)(v^T X)$ 表示为

$\sum_{i,j=1}^d u_i v_j X_i X_j$ 。因此,

$$\mathbb{E}[(u^T X)f(X)(v^T X)g(X)] = \sum_{i,j=1}^d u_i v_j \mathbb{E}[X_i X_j f(X)g(X)].$$

利用高斯分部积分公式得到

$$\mathbb{E}X_i X_j f(X)g(X) = \delta_{ij} \mathbb{E}fg + \mathbb{E}\partial_{ij}(fg) = \delta_{ij} \mathbb{E}fg + \mathbb{E}(f_{ij}g + f_i g_j + f_j g_i + f g_{ij}),$$

其中用 f_i, f_{ij}, g_i, g_{ij} 表示 f 和 g 的导数。

将各项相加, 即得所求。 ■

引理 6.6: 令 f, g 为 $\mathbb{R}^d$ 上的平滑函数, 使得它们的三阶以内导数关于标准高斯测度 γ_d 是连续且平方可积的。令 A, B 为 $\mathbb{R}^{d \times d}$ 中的两个矩阵。假设 $A\nabla f = 0$ 且 $B\nabla g = 0$, 那么, 有

$$\mathbb{E}[(\nabla f(X))^T A X \cdot (\nabla g(X))^T B X] = \mathbb{E}(\nabla f)^T A B^T \nabla g + \mathbb{E} \text{Tr}(\nabla^2 f A \nabla^2 g B). \quad (6.103)$$

证明 令 $A = (a_{ij})$ 且 $B = (b_{kl})$ 。用 $f_i, f_{ij}, g_i, \dots$ 表示 f 和 g 的导数。那么, 可以展开结果为

$$I = \mathbb{E}[(\nabla f(X))^T A X \cdot (\nabla g(X))^T B X] = \mathbb{E} \sum_{ij} a_{ij} f_i x_j \sum_{kl} b_{kl} g_k x_l = \sum_{ijkl} a_{ij} b_{kl} \mathbb{E} x_j x_l (f_i g_k).$$

利用高斯分部积分公式, 有

$$\mathbb{E} x_j x_l (f_i g_k) = \delta_{jl} \mathbb{E} f_i g_k + \mathbb{E} (f_{ijl} g_k + f_{il} g_{jk} + f_{ij} g_{kl} + f_i g_{jkl}),$$

因此有

$$I = I_0 + I_1 + I_2 + I_3 + I_4 + I_5.$$

为了简化结果, 条件 $A\nabla f = B\nabla g = 0$ 写作

$$\sum_j a_{ij} f_j = 0, \quad \sum_l b_{kl} g_l = 0.$$

因此, 由于导数是线性的且是对易的, 有

$$I_1 = \mathbb{E} \sum_{ijkl} a_{ij} b_{kl} f_{ijl} g_k = \mathbb{E} \sum_{ikl} b_{kl} g_k \sum_j a_{ij} f_{jil} = \mathbb{E} \sum_{ikl} b_{kl} g_k \partial_{il} \sum_j a_{ij} f_j = 0.$$

类似地, $I_5 = 0$ 。此外,

$$I_3 = \mathbb{E} \sum_{ijkl} a_{ij} b_{kl} f_{ij} g_{kl} = \mathbb{E} \sum_{ikl} b_{kl} g_{kl} \sum_j a_{ij} f_{ji} = \mathbb{E} \sum_{ikl} b_{kl} g_{kl} \partial_i \sum_j a_{ij} f_j = 0.$$

剩余项可以写作

$$I_0 = \mathbb{E} \sum_{ijkl} a_{ij} b_{kl} \delta_{jl} f_i g_k = \mathbb{E} \sum_{ijk} a_{ij} b_{kj} f_i g_k = \mathbb{E}(\nabla f)^T A B^T \nabla g,$$

同时

$$\begin{aligned}
 I_2 &= \mathbb{E} \sum_{ijkl} a_{ij} b_{kl} f_{il} g_{jk} = \mathbb{E} \sum_l \sum_{ijk} f_{li} a_{ij} g_{jk} b_{kl} \\
 &= \mathbb{E} \sum_l (\nabla^2 f A \nabla^2 g B)_{ll} \\
 &= \mathbb{E} \text{Tr}(\nabla^2 f A \nabla^2 g B).
 \end{aligned}$$

■

6.6.3 Hermite 多项式

引理 6.7: 令 $x \sim \mathcal{N}(0, I_d)$, 并令 $P \in S(d, p)$ 和 $Q \in S(d, q)$ 为 Stiefel 矩阵。设 $m \in \mathbb{N}^p$ 和 $n \in \mathbb{N}^q$ 为多重指标。有

$$\mathbb{E} [H_m(P^\top x) H_n(Q^\top x)] = \sqrt{m!n!} [\alpha^m \beta^n] e^{\alpha^\top R \beta} = \delta_{|m|, |n|} \sum_{Y \in \Pi(m, n)} \frac{\sqrt{m!n!}}{Y!} R^Y, \quad (6.104)$$

其中 $R = P^\top Q \in \mathbb{R}^{p \times q}$, $\Pi(m, n) \subset \mathbb{N}^{p \times q}$ 是满足以下条件的整数矩阵 $Y = [Y_{i,j}]$ 的集合:

$$\sum_{j=1}^q Y_{i,j} = m_i \quad \text{对所有 } i \text{ 成立}, \quad \sum_{i=1}^p Y_{i,j} = n_j \quad \text{对所有 } j \text{ 成立}.$$

且 $Y! = \prod_{i,j} Y_{i,j}!$, $R^Y = \prod_{i,j} (R_{i,j})^{Y_{i,j}}$ 。

证明 令 $u = P^\top x \in \mathbb{R}^p$ 且 $v = Q^\top x \in \mathbb{R}^q$ 。那么 (u, v) 服从均值为 0、协方差矩阵为下式的联合高斯分布:

$$\Sigma = \begin{pmatrix} I_p & R \\ R^\top & I_q \end{pmatrix}, \quad \text{其中 } R = P^\top Q \in \mathbb{R}^{p \times q}.$$

对于向量对 $(u, v) \in \mathbb{R}^p \times \mathbb{R}^q$, 考虑联合生成函数:

$$e^{\langle \alpha, u \rangle + \langle \beta, v \rangle} = e^{\alpha^\top P^\top x + \beta^\top Q^\top x} = e^{\langle P\alpha + Q\beta, x \rangle}.$$

由于 $x \sim \mathcal{N}(0, I_d)$, 该指数函数的期望为:

$$\mathbb{E} [e^{\langle P\alpha + Q\beta, x \rangle}] = e^{\frac{1}{2} \|P\alpha + Q\beta\|^2}.$$

展开二次项:

$$\|P\alpha + Q\beta\|^2 = \alpha^\top P^\top P \alpha + \beta^\top Q^\top Q \beta + 2\alpha^\top P^\top Q \beta = \|\alpha\|^2 + \|\beta\|^2 + 2\alpha^\top R \beta,$$

因为 $P^\top P = I_p$ 且 $Q^\top Q = I_q$ 。因此,

$$\mathbb{E} [e^{\langle \alpha, u \rangle + \langle \beta, v \rangle}] = e^{\frac{1}{2} \|\alpha\|^2 + \frac{1}{2} \|\beta\|^2 + \alpha^\top R \beta}.$$

展开左边:

$$e^{\langle \alpha, u \rangle + \langle \beta, v \rangle} = e^{\frac{1}{2}\|\alpha\|^2 + \frac{1}{2}\|\beta\|^2} \sum_{m \in \mathbb{N}^p} \frac{\alpha^m}{\sqrt{m!}} H_m(u) \sum_{n \in \mathbb{N}^q} \frac{\beta^n}{\sqrt{n!}} H_n(v).$$

取期望:

$$\mathbb{E}[e^{\langle \alpha, u \rangle + \langle \beta, v \rangle}] = e^{\frac{1}{2}\|\alpha\|^2 + \frac{1}{2}\|\beta\|^2} \sum_{m \in \mathbb{N}^p} \sum_{n \in \mathbb{N}^q} \frac{\alpha^m}{\sqrt{m!}} \frac{\beta^n}{\sqrt{n!}} \mathbb{E}[H_m(u) H_n(v)].$$

由此得出:

$$\sum_{m \in \mathbb{N}^p} \sum_{n \in \mathbb{N}^q} \frac{\alpha^m \beta^n}{\sqrt{m! n!}} \mathbb{E}[H_m(u) H_n(v)] = e^{\alpha^\top R \beta} = \sum_{k=0}^{\infty} \frac{1}{k!} (\alpha^\top R \beta)^k.$$

通过匹配系数得到:

$$\mathbb{E}[H_m(u) H_n(v)] = \sqrt{m! n!} [\alpha^m \beta^n] e^{\alpha^\top R \beta}.$$

最后证明最后一个表达式。由于 $\alpha^\top R \beta = \sum_{i,j} R_{i,j} \alpha_i \beta_j$, 有

$$(\alpha^\top R \beta)^k = \sum_{Y: \sum_{i,j} Y_{ij} = k} \frac{k!}{\prod_{i,j} Y_{ij}!} \prod_{i,j} R_{i,j}^{Y_{ij}} \alpha_i^{Y_{ij}} \beta_j^{Y_{ij}},$$

因此

$$\begin{aligned} e^{\alpha^\top R \beta} &= \sum_{k=0}^{\infty} \sum_{Y: \sum_{i,j} Y_{ij} = k} \frac{1}{\prod_{i,j} Y_{ij}!} \prod_{i,j} R_{i,j}^{Y_{ij}} \alpha_i^{Y_{ij}} \beta_j^{Y_{ij}} \\ &= \sum_{Y \in \mathbb{N}^{p \times q}} \frac{1}{Y!} \prod_{i,j} R_{i,j}^{Y_{ij}} \alpha_i^{Y_{ij}} \beta_j^{Y_{ij}}. \end{aligned}$$

由此可知,

$$[\alpha^m \beta^n] e^{\alpha^\top R \beta} = \sum_{Y \in \mathbb{N}^{p \times q}, \sum_j Y_{ij} = m_i, \sum_i Y_{ij} = n_j} \frac{1}{Y!} \prod_{i,j} R_{i,j}^{Y_{ij}}.$$

■

推论 6.5: 令 $x \sim N(0, I_d)$ 且 $u, v \in \mathbb{S}^{d-1}$. 有

$$\mathbb{E} H_m(\langle u, x \rangle) H_n(\langle v, x \rangle) = \delta_{m,n} \langle u, v \rangle^m, \quad (6.105)$$

$$\mathbb{E} H_m(\langle u, x \rangle) H_n(x) = \delta_{m,|n|} \binom{m}{n}^{\frac{1}{2}} u^n. \quad (6.106)$$

证明 第一个等式通过在引理 6.7 中取 $P = u$ 且 $Q = v$ 容易得出, 此时 $p = q = 1$ 且 $\alpha, \beta \in \mathbb{R}$.

对于第二个等式, 在引理 6.7 中取 $P = u$ 且 $Q = I_d$, 因此 $p = 1, q = d$ 且 $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^d$. 计算得到

$$\mathbb{E} H_m(\langle u, x \rangle) H_n(x) = \sqrt{m! n!} [\alpha^m \beta^n] e^{\alpha^\top u^\top \beta}$$

$$\begin{aligned}
 &= \sqrt{m!n!}[\alpha^m \beta^n] \sum_{k \geq 0} \frac{1}{k!} (\alpha u^\top \beta)^k \\
 &= \sqrt{m!n!} \delta_{m,|n|} [\alpha^m \beta^n] \frac{\alpha^m (u^\top \beta)^m}{m!} \\
 &= \delta_{m,|n|} \sqrt{m!n!} \frac{1}{m!} [\beta^n] (u^\top \beta)^m \\
 &= \delta_{m,|n|} \sqrt{m!n!} \frac{1}{m!} \binom{m}{n} u^n \\
 &= \delta_{m,|n|} \binom{m}{n}^{\frac{1}{2}} u^n.
 \end{aligned}$$

■

推论 6.6: 令 $r \in \mathbb{N}$ 。空间 $\mathcal{H}_r = \text{span}\{H_m : m \in \mathbb{N}^p, |m| = r\}$ 在任何 $Q \in \mathcal{S}(p, p)$ 对应的 P_Q 作用下是不变的。

引理 6.8: 令 $\sigma \in [-1, 1]$ 且 $\xi \sim N(0, 1 - \sigma^2)$ 。那么,

$$\mathbb{E}_\xi H_m(\sigma y + \xi) = \sigma^m H_m(y).$$

证明 利用归一化 Hermite 多项式的生成函数 (6.3) 来证明该引理。定义

$$g(t) = \mathbb{E}_\xi e^{t(\sigma y + \xi) - \frac{t^2}{2}}.$$

利用 (6.3) 展开右式, 得到:

$$g(t) = \sum_{m=0}^{\infty} \frac{\mathbb{E}_\xi H_m(\sigma y + \xi)}{\sqrt{m!}} t^m.$$

另一方面, 可以将 $g(t)$ 重写为

$$\begin{aligned}
 g(t) &= e^{t\sigma y - \frac{t^2}{2}} \mathbb{E}_\xi e^{t\xi} = e^{t\sigma y - \frac{t^2\sigma^2}{2}} e^{\frac{t^2(1-\sigma^2)}{2}} \\
 &= e^{t\sigma y - \frac{t^2\sigma^2}{2}}.
 \end{aligned}$$

因此, 再次使用 (6.3), 有

$$g(t) = \sum_{m=0}^{\infty} \frac{H_m(\sigma y)}{\sqrt{m!}} (t\sigma)^m = \sum_{m=0}^{\infty} \frac{H_m(y)}{\sqrt{m!}} \sigma^m t^m.$$

通过对比两个展开式, 即可得到所需结果。 ■

以下是关于 Hermite 多项式的标准结果。

引理 6.9 (递推关系与导数): 令 $n \in \mathbb{N}$ 。有

$$\sqrt{n+1} H_{n+1} = x H_n - \sqrt{n} H_{n-1}. \quad (6.107)$$

此外, 对于任何多重指标 $m \in \mathbb{N}^d$,

$$\nabla_{x_i} H_m(x) = \sqrt{m_i} H_{m-e_i}(x), \quad (6.108)$$

其中 $e_i \in \mathbb{N}^d$ 是第 i 个单位向量, 约定若 $m_i = 0$ 则 $H_{m-e_i} = 0$ 。

令 $f = \sum_{m \in \mathbb{N}^d} f_m H_m$ 为 $\mathbb{R}^d$ 上某个函数的 Hermite 展开。那么, 结合引理 6.9, 有

$$\nabla_{x_i} f = \sum_m f_m \nabla_{x_i} H_m = \sum_m f_m \sqrt{m_i} H_{m-e_i} = \sum_m \sqrt{m_i+1} f_{m+e_i} H_m. \quad (6.109)$$

6.6.4 序列模型

引理 6.10: 考虑由一组正交基导出的序列模型 $z_j = f_j^* + \varepsilon_j$, $j \in N$ 。对于一个函数 f 及其在该基下的系数 f_j , 定义总体损失 (population loss) 和经验损失 (empirical loss) 分别为: $\mathcal{L} = \frac{1}{2} \sum_{j \in N} (f_j^* - f_j)^2 = \frac{1}{2} \|f^* - f\|_{L^2}^2$ 和 $\hat{\mathcal{L}} = \frac{1}{2} \sum_{j \in N} (z_j - f_j)^2$ 。则有:

$$-\nabla \hat{\mathcal{L}} = -\nabla \mathcal{L} + \sum_{j \in N} \varepsilon_j \nabla f_j. \quad (6.110)$$

证明 这可以直接从以下计算得出:

$$-\nabla \hat{\mathcal{L}} = \sum_{j \in N} (z_j - f_j) \nabla f_j = \sum_{j \in N} (f_j^* - f_j) \nabla f_j + \sum_{j \in N} \varepsilon_j \nabla f_j = -\nabla \mathcal{L} + \sum_{j \in N} \varepsilon_j \nabla f_j. \quad \blacksquare$$

6.6.5 级数

命题 6.30: 固定 $\alpha \in \mathbb{R}$ 。令 $\lambda_r \asymp r^{-\gamma}$, 其中 $\gamma > \max(0, \alpha + 1)$ 。则对于任何固定的 $s > 0$, 有:

$$\sum_{r \geq 0} r^\alpha (\log r)^q \min(1, \lambda_r t) \lesssim 1 + t^{\frac{1}{\gamma}(\alpha+1+s)}, \quad (6.111)$$

其中隐式常数取决于 α, q, s, γ 。

证明 令 $L = \inf\{r \geq 0 : \lambda_r t \leq 1\} \asymp t^{1/\gamma}$ 。那么有:

$$I = \sum_{r \geq 0} r^\alpha (\log r)^q \min(1, \lambda_r t) = \sum_{r \leq L} r^\alpha (\log r)^q + \sum_{r > L} r^\alpha (\log r)^q \lambda_r t = I_1 + I_2.$$

对于 I_1 , 有:

$$I_1 = \sum_{r \leq L} r^\alpha (\log r)^q \lesssim \begin{cases} 1, & \alpha < -1, \\ L^{\alpha+1+s} \lesssim t^{\frac{1}{\gamma}(\alpha+1+s)}, & \alpha \geq -1. \end{cases}$$

对于 I_2 , 有:

$$I_2 = \sum_{r > L} r^\alpha (\log r)^q \lambda_r t \lesssim t \sum_{r > L} r^\alpha (\log r)^q r^{-\gamma}$$

$$\begin{aligned}
 &= t \sum_{r>L} r^{-\gamma+\alpha} (\log r)^q \lesssim t L^{-\gamma+\alpha+1+s} \\
 &\lesssim t^{\frac{1}{\gamma}(\alpha+1+s)}.
 \end{aligned}$$

结论由上述两个不等式得出。 ■

命题 6.31: 固定 $p, q \in \mathbb{R}$ 。令 $\lambda_r \lesssim e^{-\gamma r}$, 其中 $\gamma > 0$ 。则对于任何固定的 $s > 0$, 有:

$$\sum_{r \geq 0} r^p (\log r)^q \min(1, \lambda_r t) \lesssim 1 + (\log^+ t)^{p+1+s}, \quad (6.112)$$

其中隐式常数取决于 p, q, s, γ 。

证明 不失一般性, 假设 $p, q \geq 0$ 且 $t \geq 1$ 。定义 $L = \inf \{r \geq 0 : \lambda_r t \leq 1\} \lesssim \gamma^{-1} \log t$ 。那么有:

$$I = \sum_{r \geq 0} r^p (\log r)^q \min(1, \lambda_r t) = \sum_{r \leq L} r^p (\log r)^q + \sum_{r > L} r^p (\log r)^q \lambda_r t = I_1 + I_2.$$

对于 I_1 , 有:

$$I_1 = \sum_{r \leq L} r^p (\log r)^q \lesssim \begin{cases} 1, & p < -1, \\ L^{p+1+s}, & p \geq -1. \end{cases}$$

对于 I_2 , 有:

$$I_2 = \sum_{r > L} r^p (\log r)^q \lambda_r t \lesssim t \sum_{r > L} r^p (\log r)^q e^{-\gamma r} \lesssim t L^{p+s} e^{-\gamma L} \lesssim L^{p+s}.$$

将上述两个不等式与 $L \lesssim \log t$ 结合, 即证得结果。 ■

6.6.6 一些初等函数

命题 6.32: 令 $x_i \in [0, 1], i = 1, \dots, p$ 。令 $K > 0$ 。定义函数:

$$\omega(x_1, \dots, x_p) = -\frac{1}{K} \log \left(\sum_{i=1}^p e^{-Kx_i} \right).$$

则有:

$$\min(x_1, \dots, x_p) \geq \omega(x_1, \dots, x_p) \geq \min(x_1, \dots, x_p) - \frac{1}{K} \log p.$$

此外:

$$\frac{\partial}{\partial x_i} \omega = \frac{e^{-Kx_i}}{\sum_{j=1}^p e^{-Kx_j}}.$$

命题 6.33: 对于任何 $x_i \in [0, 1], i = 1, \dots, p$ 和 $K > 0$, 以下不等式成立:

$$\frac{1}{\sum_{j=1}^p e^{-Kx_j}} \sum_{i=1}^p e^{-Kx_i(1-x_i)} \geq 1 - \frac{1}{p} \sum_{i=1}^p x_i.$$

第7章 总结与展望

本文围绕神经网络的泛化机制展开研究，整体脉络是从固定核的线性近似出发，逐步过渡到训练过程中动态演化的特征学习机制。一方面，本文系统研究了神经切线核及其对应核方法的谱性质与泛化行为；另一方面，本文进一步提出并分析了能够刻画动态特征学习的自适应特征模型框架。由此，全文试图在静态核视角与动态特征视角之间建立一条较为完整的理论链条。

在核方法的泛化性能方面，本文引入了解析谱算法的定义，并发展出一套新的解析算子演算证明方法，给出了精细的算子扰动分析，从而得到 $(1 + o_{\mathbb{P}}(1))$ 形式的泛化误差刻画。此结果还说明了核回归的泛化误差与其对应序列模型中收缩估计量的泛化误差等价，从而完整刻画了核回归的泛化表现。这些结论不仅能够作为推论导出几乎所有前人对于核方法泛化误差的结果，说明了核方法的优势和局限性，也为后续研究神经网络的动态特征学习提供了坚实的基础。

随后，为克服神经切线核框架下固定核函数的“惰性训练”无法解释神经网络特征学习的缺陷，本文提出了通用的自适应特征模型框架。在该框架下，本文引入参数化、可训练的特征映射，用以模拟真实网络中的特征学习过程。为了衡量特征学习的效果，本文结合固定核方法泛化能力的序列模型等价性提出了特征误差度量，用于刻画训练过程中某个特征映射与目标函数之间的匹配程度。此外，为了在简化分析的同时保留统计表现的丰富性，本文引入了序列模型简化，并使用序列损失作为样本损失的代理。本文也提供了一些数值实验，验证了序列模型简化在自适应特征模型中的有效性。

在自适应特征模型框架下，本文深入分析了两种具有代表性的自适应模型：对角参数化模型与方向参数化模型，揭示了梯度下降驱动的隐式正则化效应如何提升泛化性能。在对角参数化模型中，本文引入了可训练的特征值参数，证明了它们在训练过程中会自适应地调整以对齐目标函数的稀疏性结构，达到近乎最优的泛化能力；相比之下，无论如何选取正则化参数，固定核方法都无法达到这一效果。本文也说明了这种自适应特征学习的优势在于其能够动态调整特征分量的重要性，提升了特征映射与目标函数之间的匹配程度。在方向参数化模型中，利用 Hermite 基函数，本文引入了可训练的特征方向参数，分析了其在训练过程中的演化行为。本文发现，在单指标模型和多指标模型中，梯度下降会引导特征方向逐渐对齐目标函数的主方向，从而显著提升特征映射的表达能力和泛化性能。这两种模型的分析都涉及非线性、非凸、无显式解的复杂动力学演化。对此，本文给出了较为深

入的分析和精细的训练动态刻画，其中发展的技术方法也可为后续研究提供借鉴。

综上所述，本文的研究工作从固定核的线性机制出发，迈向自适应特征学习的非线性机制，系统地揭示了神经网络泛化性能提升的内在动力学机制。在自适应特征模型的框架下，神经网络的过参数化特性并不有害于它的泛化能力；相反，过参数化结构在梯度下降驱动的训练动态中会自动实现对隐式结构性特征的自适应调整与优化，对齐数据的内在结构，提升模型的泛化性能。这一系列的结论和观点不仅深化了对神经网络从线性机制向非线性特征学习过渡的理解，也为设计更高效、具备可解释性的深度学习架构提供了理论指导。

尽管本文在理解神经网络的特征学习机理方面取得了一些进展，但仍有许多有趣且重要的问题值得进一步探索：

- **更广泛的自适应特征模型：**本文主要分析了对角参数化模型与方向参数化模型，未来可以探索其他类型的自适应特征模型，如低秩参数化模型、卷积参数化模型以及 Transformer 中的注意力参数化等，以进一步理解不同结构对特征学习与泛化的影响。
- **组合式的特征学习：**本文对自适应特征模型的分析主要集中于单一的参数化结构。未来可以将研究扩展到多种参数化形式的组合结构，探讨多层结构的组合如何影响特征演化与泛化性能。
- **优化算法的影响：**本文主要考虑梯度流方法，这对应于梯度下降步长趋于零的极限。而在实际训练中，常用的是离散步长的梯度下降，这会带来额外的影响，例如稳定性边缘（Edge of Stability）现象。此外，未来也可以研究随机梯度下降、动量法或者自适应学习率方法等其他常用优化算法对特征学习与泛化的影响。
- **实际应用中的验证：**本文所给出的理论结果主要基于理想化的假设和简化的模型，未来可以将本文的理论结果应用于图像分类、自然语言处理等实际问题，验证其有效性并指导模型设计。另一方面，本文的数值实验主要验证了理论模型本身的训练动力学；未来还可以进一步比较自适应核梯度流、直接训练的神经网络以及经典自适应非参数回归方法，并为方向参数化模型补充更系统的数值实验，以评估理论机制在有限样本和实际数据中的表现。

参考文献

- [1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems: Vol. 25. 2012.
- [2] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
- [3] Kingma D P, Welling M. Auto-encoding variational bayes[A]. 2013. arXiv: 1312.6114.
- [4] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems: Vol. 27. 2014.
- [5] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]//Advances in Neural Information Processing Systems: Vol. 33. 2020: 6840-6851.
- [6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems: Vol. 30. 2017.
- [7] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems: Vol. 33. 2020: 1877-1901.
- [8] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873): 583-589.
- [9] Tsybakov A B. Introduction to nonparametric estimation[M]. 1st ed. New York ; London: Springer, 2009.
- [10] Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms [M]. New York, NY, USA: Cambridge University Press, 2014.
- [11] Zhang C, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization[A]. 2017. arXiv: 1611.03530.
- [12] Jacot A, Gabriel F, Hongler C. Neural tangent kernel: Convergence and generalization in neural networks[C]//Bengio S, Wallach H, Larochelle H, et al. Advances in Neural Information Processing Systems: Vol. 31. Curran Associates, Inc., 2018.
- [13] Steinwart I, Christmann A. Support vector machines[M]. New York, NY: Springer-Verlag New York, 2008.
- [14] Caponnetto A, De Vito E. Optimal rates for the regularized least-squares algorithm[J]. Foundations of Computational Mathematics, 2007, 7(3): 331-368.
- [15] Arora S, Du S S, Hu W, et al. On exact computation with an infinitely wide neural net[C]//Advances in Neural Information Processing Systems: Vol. 32. Curran Associates, Inc., 2019.
- [16] Lee J, Xiao L, Schoenholz S, et al. Wide neural networks of any depth evolve as linear models under gradient descent[C]//Advances in Neural Information Processing Systems: Vol. 32. Curran Associates, Inc., 2019.
- [17] Woodworth B, Gunasekar S, Lee J D, et al. Kernel and rich regimes in overparametrized models [C]//Conference on Learning Theory. PMLR, 2020: 3635-3673.

-
- [18] Yang G, Hu E J. Feature learning in infinite-width neural networks[A]. 2022. arXiv: 2011.14522.
 - [19] Wendland H. Scattered data approximation[M]. Cambridge: Cambridge University Press, 2004.
 - [20] Du S S, Zhai X, Poczos B, et al. Gradient descent provably optimizes over-parameterized neural networks[A]. 2018. arXiv: 1810.02054.
 - [21] Arora S, Du S, Hu W, et al. Fine-grained analysis of optimization and generalization for over-parameterized two-layer neural networks[C]//International Conference on Machine Learning. PMLR, 2019: 322-332.
 - [22] Du S S, Zhai X, Poczos B, et al. Gradient descent provably optimizes over-parameterized neural networks[A]. 2019. arXiv: 1810.02054.
 - [23] Allen-Zhu Z, Li Y, Song Z. A convergence theory for deep learning via over-parameterization [Z]. 2019.
 - [24] Lee J, Xiao L, Schoenholz S S, et al. Wide neural networks of any depth evolve as linear models under gradient descent[J]. Journal of Statistical Mechanics: Theory and Experiment, 2019, 2020(12): 124002.
 - [25] Hu T, Wang W, Lin C, et al. Regularization matters: A nonparametric perspective on over-parametrized neural network[C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2021: 829-837.
 - [26] Suh N, Ko H, Huo X. A non-parametric regression viewpoint: Generalization of over-parametrized deep ReLU network under noisy observations[C]//International Conference on Learning Representations. 2021.
 - [27] Lai J, Xu M, Chen R, et al. Generalization ability of wide neural networks on R[A]. 2023. arXiv: 2302.05933.
 - [28] Li Y, Yu Z, Chen G, et al. On the eigenvalue decay rates of a class of neural-network related kernel functions defined on general domains[J]. Journal of Machine Learning Research, 2024, 25(82): 1-47.
 - [29] Dai F, Xu Y. Approximation theory and harmonic analysis on spheres and balls[M]. New York, NY: Springer New York, 2013.
 - [30] Bietti A, Mairal J. On the inductive bias of neural tangent kernels[C]//Advances in Neural Information Processing Systems: Vol. 32. 2019.
 - [31] Geifman A, Yadav A, Kasten Y, et al. On the similarity between the Laplace and neural tangent kernels[C]//Advances in Neural Information Processing Systems: Vol. 33. 2020: 1451-1461.
 - [32] Chen L, Xu S. Deep neural tangent kernel and laplace kernel have the same RKHS[A]. 2020. arXiv: 2009.10683.
 - [33] Bietti A, Bach F. Deep equals shallow for relu networks in kernel regimes[A]. 2020. arXiv: 2009.14397.
 - [34] Steinwart I, Hush D R, Scovel C, et al. Optimal rates for regularized least squares regression. [C]//Conference on Learning Theory. 2009: 79-93.
 - [35] Fischer S R, Steinwart I. Sobolev norm learning rates for regularized least-squares algorithms [J]. Journal of Machine Learning Research, 2020, 21: 205:1-205:38.

-
- [36] Zhang H, Li Y, Lu W, et al. On the optimality of misspecified kernel ridge regression[C]// International Conference on Machine Learning. 2023.
 - [37] Zhang T, Yu B. Boosting with early stopping: Convergence and consistency[J]. The Annals of Statistics, 2005, 33(4): 1538-1579.
 - [38] Yao Y, Rosasco L, Caponnetto A. On early stopping in gradient descent learning[J]. Constructive Approximation, 2007, 26: 289-315.
 - [39] Bauer F, Pereverzev S, Rosasco L. On regularization algorithms in learning theory[J]. Journal of Complexity, 2007, 23(1): 52-72.
 - [40] Gerfo L L, Rosasco L, Odone F, et al. Spectral algorithms for supervised learning[J]. Neural Computation, 2008, 20(7): 1873-1897.
 - [41] Rosasco L, De Vito E, Verri A. Spectral methods for regularization in learning theory[J]. DISI, Universita degli Studi di Genova, Italy, Technical Report DISI-TR-05-18, 2005.
 - [42] Caponnetto A. Optimal rates for regularization operators in learning theory[Z]. 2006.
 - [43] Blanchard G, Mücke N. Optimal rates for regularization of statistical inverse learning problems [J]. Foundations of Computational Mathematics, 2018, 18(4): 971-1013.
 - [44] Lin J, Cevher V. Optimal rates of sketched-regularized algorithms for least-squares regression over Hilbert spaces[C]//International Conference on Machine Learning. PMLR, 2018: 3102-3111.
 - [45] Rakhlin A, Zhai X. Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon[A]. 2018. arXiv: 1812.11167.
 - [46] Buchholz S. Kernel interpolation in Sobolev spaces is not consistent in low dimensions[C]// Conference on Learning Theory. PMLR, 2022: 3410-3440.
 - [47] Beaglehole D, Belkin M, Pandit P. Kernel ridgeless regression is inconsistent in low dimensions [A]. 2022. arXiv: 2205.13525.
 - [48] Bordelon B, Canatar A, Pehlevan C. Spectrum dependent learning curves in kernel regression and wide neural networks[C]//International Conference on Machine Learning. PMLR, 2020: 1024-1034.
 - [49] Cui H, Loureiro B, Krzakala F, et al. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime[C]//Advances in Neural Information Processing Systems: Vol. 34. 2021: 10131-10143.
 - [50] Mallinar N, Simon J B, Abedsoltan A, et al. Benign, tempered, or catastrophic: A taxonomy of overfitting[A]. 2022. arXiv: 2207.06569.
 - [51] Liang T, Rakhlin A. Just interpolate: Kernel "ridgeless" regression can generalize[J]. The Annals of Statistics, 2020, 48(3).
 - [52] Ghorbani B, Mei S, Misiakiewicz T, et al. Linearized two-layers neural networks in high dimension[A]. 2020. arXiv: 1904.12191.
 - [53] Ghosh N, Mei S, Yu B. The three stages of learning dynamics in high-dimensional kernel methods[A]. 2021. arXiv: 2111.07167.
 - [54] Liu F, Liao Z, Suykens J A K. Kernel regression in high dimensions: Refined analysis beyond double descent[A]. 2021. arXiv: 2010.02681.

-
- [55] Lu W, Zhang H, Li Y, et al. Optimal rate of kernel regression in large dimensions[A]. 2023. arXiv: 2309.04268.
- [56] Zhang H, Li Y, Lu W, et al. Optimal rates of kernel ridge regression under source condition in large dimensions[J]. Journal of Machine Learning Research, 2025, 26(219): 1-63.
- [57] Lu W, Zhang H, Li Y, et al. On the saturation effects of spectral algorithms in large dimensions [C]//Advances in Neural Information Processing Systems. 2024.
- [58] Ba J, Erdogdu M A, Suzuki T, et al. High-dimensional asymptotics of feature learning: How one gradient step improves the representation[A]. 2022. arXiv: 2205.01445.
- [59] Moniri B, Lee D, Hassani H, et al. A theory of non-linear feature learning with one gradient step in two-layer neural networks[A]. 2024. arXiv: 2310.07891.
- [60] Cui H, Pesce L, Dandi Y, et al. Asymptotics of feature learning in two-layer networks after one gradient-step[A]. 2024. arXiv: 2402.04980.
- [61] Bordelon B, Atanasov A, Pehlevan C. How feature learning can improve neural scaling laws [A]. 2024. arXiv: 2409.17858.
- [62] LeJeune D, Alemohammad S. An adaptive tangent feature perspective of neural networks[A]. 2023. arXiv: 2308.15478.
- [63] Dandi Y, Pesce L, Cui H, et al. A random matrix theory perspective on the spectrum of learned features and asymptotic generalization capabilities[A]. 2024. arXiv: 2410.18938.
- [64] Damian A, Lee J D, Soltanolkotabi M. Neural networks can learn representations with gradient descent[A]. 2022. arXiv: 2206.15144.
- [65] Gatmiry K, Jegelka S, Kelner J. Optimization and adaptive generalization of three layer neural networks[C]//International Conference on Learning Representations. 2021.
- [66] Hoff P D. Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization[J]. Computational Statistics & Data Analysis, 2017, 115: 186-198.
- [67] Gunasekar S, Woodworth B, Bhojanapalli S, et al. Implicit regularization in matrix factorization [C]//Advances in Neural Information Processing Systems: 2017-December. 2017: 6152-6160.
- [68] Arora S, Cohen N, Hu W, et al. Implicit regularization in deep matrix factorization[C]//Advances in Neural Information Processing Systems: Vol. 32. 2019.
- [69] Li Z, Luo Y, Lyu K. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning[A]. 2021. arXiv: 2012.09839.
- [70] Razin N, Maman A, Cohen N. Implicit regularization in tensor factorization[A]. 2021. arXiv: 2102.09972.
- [71] Yun C, Krishnan S, Mobahi H. A unifying view on implicit bias in training linear neural networks[A]. 2021. arXiv: 2010.02501.
- [72] Nacson M S, Ravichandran K, Srebro N, et al. Implicit bias of the step size in linear diagonal neural networks[C]//International Conference on Machine Learning. PMLR, 2022: 16270-16295.
- [73] Kubo M, Banno R, Manabe H, et al. Implicit regularization in over-parameterized neural networks[A]. 2019. arXiv: 1903.01997.

-
- [74] Li Z, Wang T, Arora S. What happens after SGD reaches zero loss? –a mathematical framework [A]. 2022. arXiv: 2110.06914.
 - [75] Vivien L P, Reygner J, Flammarion N. Label noise (stochastic) gradient descent implicitly solves the Lasso for quadratic parametrisation[C]//Conference on Learning Theory. PMLR, 2022: 2127-2159.
 - [76] Pesme S, Pillaud-Vivien L, Flammarion N. Implicit bias of sgd for diagonal linear networks: A provable benefit of stochasticity[C]//Advances in Neural Information Processing Systems: Vol. 34. Curran Associates, Inc., 2021: 29218-29230.
 - [77] Zhao P, Yang Y, He Q C. High-dimensional linear regression via implicit regularization[J]. Biometrika, 2022, 109(4): 1033-1046.
 - [78] Li J, Nguyen T V, Hegde C, et al. Implicit sparse regularization: The impact of depth and early stopping[A]. 2021. arXiv: 2108.05574.
 - [79] Vaškevičius T, Kanade V, Rebeschini P. Implicit regularization for optimal sparse recovery[A]. 2019. arXiv: 1909.05122.
 - [80] Fan J, Yang Z, Yu M. Understanding implicit regularization in over-parameterized single index model[A]. 2021. arXiv: 2007.08322.
 - [81] Cho Y, Saul L. Kernel methods for deep learning[C]//Bengio Y, Schuurmans D, Lafferty J, et al. Advances in Neural Information Processing Systems: Vol. 22. Curran Associates, Inc., 2009.
 - [82] Aronszajn N. Theory of reproducing kernels[J]. Transactions of the American Mathematical Society, 1950, 68(3): 337-404.
 - [83] Steinwart I, Scovel C. Mercer’s theorem on general domains: On the interaction between measures, kernels, and RKHSs[Z]. 2012.
 - [84] Lin J, Rudi A, Rosasco L, et al. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces[J]. Applied and Computational Harmonic Analysis, 2018, 48: 868-890.
 - [85] Widom H. Asymptotic behavior of the eigenvalues of certain integral equations[J]. Transactions of the American Mathematical Society, 1963, 109(2): 278-295.
 - [86] Simon B. Operator theory[M]. Providence, Rhode Island: American Mathematical Society, 2015.
 - [87] Weyl H. On the volume of tubes[J]. American Journal of Mathematics, 1939, 61(2): 461-472.
 - [88] Li Y, Zhang H, Lin Q. On the asymptotic learning curves of kernel ridge regression under power-law decay[C]//Advances in Neural Information Processing Systems. 2023.
 - [89] Li Y, Zhang H, Lin Q. Kernel interpolation generalizes poorly[J]. Biometrika, 2024, 111(2): 715-722.
 - [90] Johnstone I M. Gaussian estimation: Sequence and wavelet models[Z]. 2017.
 - [91] Li Y, Zhang H, Lin Q. On the saturation effect of kernel ridge regression[C]//International Conference on Learning Representations. 2023.
 - [92] Brown L D, Cai T T, Low M G, et al. Asymptotic equivalence theory for nonparametric regression with random design[J]. The Annals of Statistics, 2002, 30(3): 688-707.

-
- [93] Reiß M. Asymptotic equivalence for nonparametric regression with multivariate and random design[J]. *The Annals of Statistics*, 2008, 36(4): 1957-1982.
 - [94] Hsing T, Eubank R. Theoretical foundations of functional data analysis, with an introduction to linear operators[M]. Wiley, 2015.
 - [95] Li Y, Gan W, Shi Z, et al. Generalization error curves for analytic spectral algorithms under power-law decay[A]. 2024. arXiv: 2401.01599.
 - [96] Zhang H, Li Y, Lin Q. On the optimality of misspecified spectral algorithms[J]. *Journal of Machine Learning Research*, 2024, 25(188): 1-50.
 - [97] Wainwright M J. High-dimensional statistics: A non-asymptotic viewpoint[M]. Cambridge University Press, 2019.
 - [98] Tropp J A. An introduction to matrix concentration inequalities[A]. 2015. arXiv: 1501.01571.
 - [99] Li Y, Lin Q. Diagonal over-parameterization in reproducing kernel hilbert spaces as an adaptive feature model: Generalization and adaptivity[A]. 2025. arXiv: 2501.08679.
 - [100] Zhang H, Lai J, Li Y, et al. Towards a statistical understanding of neural networks: Beyond the neural tangent kernel theories[A]. 2024. arXiv: 2412.18756.
 - [101] Li Y, Lin Q. Improving adaptivity via over-parameterization in sequence models[C]//Advances in Neural Information Processing Systems. 2024.
 - [102] Vershynin R. High-dimensional probability: An introduction with applications in data science: Vol. 47[M]. Cambridge university press, 2018.
 - [103] Pinelis I F, Sakhanenko A I. Remarks on inequalities for large deviation probabilities[J]. *Theory of Probability & Its Applications*, 1986, 30(1): 143-148.
 - [104] Bietti A, Bruna J, Pillaud-Vivien L. On learning gaussian multi-index models with gradient flow[A]. 2023. arXiv: 2310.19793.
 - [105] Dudeja R, Hsu D. Learning single-index models in gaussian space[C]//Conference on Learning Theory. PMLR, 2018: 1887-1930.
 - [106] Damian A, Pillaud-Vivien L, Lee J D, et al. Computational-statistical gaps in Gaussian single-index models[A]. 2024. arXiv: 2403.05529.
 - [107] Bietti A, Bruna J, Sanford C, et al. Learning single-index models with shallow neural networks [C]//Advances in Neural Information Processing Systems: Vol. 35. 2022: 9768-9783.
 - [108] Kuchibhotla A K, Patra R K. Efficient estimation in single index models through smoothing splines[A]. 2019. arXiv: 1612.00068.
 - [109] Arous G B, Gheissari R, Jagannath A. Online stochastic gradient descent on non-convex losses from high-dimensional inference[J]. *Journal of Machine Learning Research*, 2021, 22(106): 1-51.
 - [110] Kalai A T, Sastry R. The Isotron algorithm: High-dimensional isotonic regression.[C]//Conference on Learning Theory: Vol. 1. 2009: 9.
 - [111] Kakade S M, Kanade V, Shamir O, et al. Efficient learning of generalized linear and single index models with isotonic regression[C]//Advances in Neural Information Processing Systems: Vol. 24. 2011.

- [112] Arous G B, Gheissari R, Jagannath A. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling[A]. 2023. arXiv: 2206.04030.
- [113] Abbe E, Boix-Adsera E, Misiakiewicz T. The merged-staircase property: A necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks [A]. 2024. arXiv: 2202.08658.
- [114] Abbe E, Boix-Adsera E, Misiakiewicz T. SGD learning on neural networks: Leap complexity and saddle-to-saddle dynamics[A]. 2023. arXiv: 2302.11055.
- [115] Rasmussen C E, Williams C K I. Gaussian processes for machine learning[M]. Cambridge, Mass: MIT Press, 2006.
- [116] Absil P A, Edelman A, Koev P. On the largest principal angle between random subspaces[J]. Linear Algebra and Its Applications, 2006, 414(1): 288-294.