Pure Mathematics 理论数学, **2026, 16(5), 80-98**
Published Online May 2026 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2026.165133

Hans汉斯

多核插值估计中的自适应隐式正则化与泛化分析

梁维俊，崔文泉*

中国科学技术大学管理学院，安徽 合肥

收稿日期：2026 年 3 月 24 日；录用日期：2026 年 4 月 29 日；发布日期：2026 年 5 月 25 日

摘 要

核插值估计凭借隐式正则化带来的良性过拟合特性，已成为统计学习中突破经典偏差 - 方差权衡框架的重要研究方向。然而，单核函数难以适配复杂数据的多尺度特征，也无法自适应匹配核矩阵未知的特征值衰减规律，导致其在实际应用中性能受限。针对上述问题，本文将核插值的隐式正则化理论框架扩展至多核学习场景，提出一种自适应多核插值估计方法。该方法利用 M 个正定内积核的线性组合构造多核函数，通过自适应优化核权重，克服单核插值在复杂数据表示与带宽选择上的局限，使模型具备自适应实现最优隐式正则化的能力。本文进一步建立了多核插值估计量积分平方风险的数据依赖型上界，并证明在高维情形下，该估计量以高概率满足泛化误差的收敛性。

关键词

隐式正则化，良性过拟合，核插值估计，多核学习，可再生核希尔伯特空间

Large Adaptive Implicit Regularization and Generalization Analysis for Multi-Kernel Interpolation

* 通讯作者。

文章引用: 梁维俊, 崔文泉. 多核插值估计中的自适应隐式正则化与泛化分析 [J]. 理论数学, 2026, 16(5): 80-98.
DOI: 10.12677/pm.2026.165133

Weijun Liang, Wenquan Cui*

School of Management, University of Science and Technology of China, Hefei Anhui

Received: March 24, 2026; accepted: April 29, 2026; published: May 25, 2026

Abstract

Kernel interpolation estimators have emerged as a prominent direction in statistical learning for overcoming the classical bias-variance trade-off, due to the benign overfitting enabled by implicit regularization. Nevertheless, single kernel functions are inadequate for capturing the multi-scale features in complex data and cannot adaptively match the unknown eigenvalue decay of the kernel matrix, which restricts their practical performance. To tackle these limitations, this paper extends the theoretical framework of implicit regularization for kernel interpolation to the setting of multiple kernel learning and proposes an adaptive multiple kernel interpolation estimator. The proposed estimator constructs a composite kernel as a linear combination of M positive definite inner-product kernels. By adaptively optimizing the kernel weights, it alleviates the drawbacks of single-kernel interpolation in representing complex data and selecting bandwidths, allowing the model to achieve optimal implicit regularization in a data-driven manner. Furthermore, we derive a data-dependent upper bound for the integrated squared risk of the multiple kernel interpolation estimator and prove that the estimator achieves convergence of the generalization error with high probability under high-dimensional settings.

Keywords

Implicit Regularization, Benign Overfitting, Kernel Interpolation Estimator, Multiple Kernel Learning, Reproducing Kernel Hilbert Space

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

在经典统计学习理论中，泛化误差的分析高度依赖于模型复杂度。Vapnik 构建的统计学习基本框架 [1] 以及 Geman 等人提出的偏差 -方差困境 (Bias-Variance Dilemma) [2] 是这一领域的基石。后续研究借助 Rademacher 复杂度与 VC 维等数学工具 [3] [4] 进一步完善了误差上界的理论体系，形成了模型复杂度必须与可用样本规模相匹配的传统认知。

然而，近年来的一些研究打破了这一传统认知。Belkin 等人 [5] 通过实验指出，过度参数化的高维模型在实现零训练误差 (即完美插值数据) 后，其测试误差并未如传统理论预测那般发散，反而随着模型复杂度的进一步增加而再次下降。Nakkiran 等人 [6] 将这一突破传统 U 型风险曲线的现象形式化为“双下降”(Double Descent)。Bartlett 等人 [7] 首次在无岭线性回归 (Ridgeless Linear Regression) 中证明了“良性过拟合”(Benign Overfitting) 的存在条件；随后，Tsigler 和 Bartlett [8] 将该结论推广至一般性的岭回归模型，而 Frei 等人 [9] 则揭示了非线性模型中同样存在良性过拟合的理论边界。

在这一非传统泛化现象的理论探究中，基于再生核希尔伯特空间 (RKHS，$\mathcal{H}$) 的模型由于其完备的数学性质，成为了重要的分析工具 [10] [11]。在经典的核估计中，高维情形下通常必须引入显式的 Tikhonov 正则化项 $\lambda\|f\|_{\mathcal{H}}^2$ 以规避过拟合 [12]。

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2+\lambda\|f\|_{\mathcal{H}}^2, \tag{1.1}$$

然而，Liang 和 Rakhlin [13] 的性研究证实，在去除显式正则化 (即设 $\lambda\to 0$) 的情形下，核“无岭”回归 (Kernel Ridgeless Regression) 不仅能够精确插值训练数据，还能实现渐进一致收敛。本文后续讨论中将这种方法称之为核插值 (Kernel Interpolation) 方法。

这种无需人工设定惩罚项的泛化能力，本质上源于模型的“隐式正则化”(Implicit Regularization)。Neyshabur 等人 [14] 最早探讨了深度学习中隐式正则化的作用；随后，Jacot 等人 [15] 提出的神经正切核 (Neural Tangent Kernel) 理论为分析宽泛模型的插值行为提供了基础；Arora 等人 [16] 进一步给出了细粒度的优化与泛化分析；而 Rakhlin 与 Zhai [17] 则严格证明了在拉普拉斯核中，这种插值一致性是由高维空间中核矩阵特征值衰减规律主导的高维现象。

尽管单核插值估计的理论框架已初具规模，但在真实的复杂应用场景中，数据往往蕴含多尺度的几何特征。固定的单一核函数难以自适应地匹配未知数据的谱特征衰减规律。多核学习 (Multiple Kernel Learning, MKL) 因其能够融合不同尺度的特征表示，长期以来被视为解决核选择难题的有效途径 [18]- [20]，且其在经典非插值状态下的泛化边界已有充分讨论 [21]。由此，本文旨在结合单核插值的现有研究，将多核学习方法引入核插值估计的理论框架，提出一种自适应多核插值估计方法。考虑由 M 个基础内积核经凸组合构成的多核函数：

$$K_{\text{multi}}=\sum_{m=1}^{M}\omega_m K_m,\quad \sum_{m=1}^{M}\omega_m=1,\quad \omega_m\geq 0. \tag{1.2}$$

本文将探究在多核插值设定下，模型如何在训练过程中自适应优化组合权重 ω_m，通过匹配真

实数据的特征谱衰减，从而达到最优的隐式正则化效果。这一研究不仅填补了多核学习在良性过拟合领域的理论空白，也为高维复杂数据的精准预测提供了更为鲁棒的估计方法。

2. 方法及理论

2.1. 符号说明

为了行文的简洁与严谨，本文在后续的理论推导中统一采用以下数学符号约定：

- **基本数学与矩阵：**对任意给定的矩阵 A，分别记 $A^\top$ 为其转置。$\mathrm{Tr}(A)$ 表示方阵 A 的迹。$\lambda_j(A)$ 表示半正定矩阵 A 的第 j 大特征值。I 表示适当维度的单位矩阵，$\mathbf{1}$ 表示全 1 向量。δ_{ij} 为克罗内克函数，当且仅当 $i=j$ 时取值为 1，否则为 0。
- **范数与空间：**对于向量 v，$||v||$ 默认表示其欧几里得范数 (l_2 范数)。对于矩阵 A，$||A||_{\mathrm{op}}$ 表示其算子范数。$||A||_F$ 表示其 Frobenius 范数。L^2_μ 表示关于概率测度 μ 的平方可积函数空间。
- **渐近记号：**对于两个实值序列 a_n 与 b_n，当存在绝对常数 $C>0$ 使得 $|a_n| \le C|b_n|$ 成立时，记为 $a_n = O(b_n)$。
- **伴随算子：**对于算子 A ，A^* 用于表示其伴随算子。

2.2. 模型介绍

本文考虑标准的非参数回归设定。假设我们观测到 n 个独立同分布 (i.i.d.) 的样本 $\{(x_i,y_i)\}_{i=1}^n$，其中协变量 x_i 取值于紧致域 $\Omega \subset \mathbb{R}^d$，响应变量 $y_i \in \mathbb{R}$。假设数据生成自某一未知概率分布 $\mu(x,y)$，我们的核心目标是估计条件期望函数 $f_*(x) = \mathbb{E}[y|x]$。

在基于再生核希尔伯特空间 (RKHS，记为 $\mathcal{H}$) 的单核估计中，最小范数插值解定义为：

$$\tilde{f} = \arg\min_{f\in\mathcal{H}} ||f||_{\mathcal{H}}, \quad \text{s.t.} \quad f(x_i) = y_i, \forall i. \tag{2.1}$$

设 $X \in \mathbb{R}^{n\times d}$ 是行为 $(x_1,x_2,...,x_n)$ 的矩阵，Y 是取值为 $y_1,y_2,...,y_n$ 的向量，我们记 $K(X,X) = [K(x_i,x_j)]_{ij} \in \mathbb{R}^{n\times n}$ 为核矩阵，对 $x\in\mathbb{R}$ 记 $K(x,X) = [K(x,x_1),...,K(x,x_n)] \in \mathbb{R}^{1\times n}$，当核矩阵 $K(X,X)\in\mathbb{R}^{n\times n}$ 可逆时，该插值估计存在闭式解 $\hat{f}(x) = K(x,X)K(X,X)^{-1}Y$。

然而，单一核函数难以自适应复杂数据的多尺度特性。为此，我们引入多核学习 (MKL) 机制。考虑由 M 个不同的基础核函数线性组合构成的多核函数：

$$K_{\mathrm{multi}}(x,x') = \sum_{m=1}^{M} \omega_m K_m(x,x'). \tag{2.2}$$

其中 K_m 为不同类型或带宽的正定基础核，$\omega_m \ge 0$ 为对应的权重 (通常约束 $\sum \omega_m = 1$)。多核函数 K_{multi} 是 M 个正定的函数的线性组合，也是正定核，由 Moore-Aronszajn 定理可知，存

在唯一的 RKHS(记为 $\mathcal{H}_{multi}$) 以多核函数 K_{multi} 为再生核。由此，我们定义多核插值估计为：

$$\hat{f}_{\text{multi}} = \arg\min_{f\in\mathcal{H}_{\text{multi}}} ||f||_{\mathcal{H}_{\text{multi}}} \quad \text{s.t.} \quad f(x_i) = y_i, \forall i. \tag{2.3}$$

其对应的闭式解扩展为：

$$\hat{f}_{\text{multi}}(x) = K_{\text{multi}}(x, X)K_{\text{multi}}(X, X)^{-1}Y. \tag{2.4}$$

本文旨在探究该多核插值估计量 $\hat{f}_{\text{multi}}$ 在测试集上的积分平方风险 (泛化误差) 上界，即寻找一个数据依赖的界限 $\phi_{n,d}(X, f^*)$，使得 $\mathbb{E}(\hat{f}_{\text{multi}}(x) - f_*(x))^2 \le \phi_{n,d}(X, f^*)$ 成立。

2.3. 多核积分算子与特征谱表示

为量化泛化误差，我们需要分析多核函数在特征空间中的行为。定义关于边缘概率测度 $\mu(x)$ 的多核积分算子 $T_\mu : L^2_\mu \to L^2_\mu$ 为：

$$T_\mu f(z) = \int K_{\text{multi}}(z, x)f(x)d\mu(x). \tag{2.5}$$

由谱定理，存在特征函数族 $\{e_i(x)\}_{i=1}^p$ 及其对应的非负特征值对角阵 $T = \text{diag}(t_1, ..., t_p)$，满足满足

$$T_\mu e_i = t_i e_i, \quad \text{且} \quad \int e_i(x)e_j(x)d\mu(x) = \delta_{ij}. \tag{2.6}$$

利用 Mercer 谱表示，多核函数可分解为：

$$K_{\text{multi}}(x, z) = e(x)^\top Te(z). \tag{2.7}$$

对应的，我们在有限样本上定义经验多核算子 $\hat{T} := \frac{1}{n}K_{\text{multi}X}K^*_{\text{multi}X}$，其非零特征值与规范化核矩阵 $\frac{1}{n}K_{\text{multi}}(X, X)$ 完全一致。此时，多核插值估计量在谱空间中可等价表述为：

$$\hat{f}_{\text{multi}}(x) = e(x)^*Te(X)[e(X)^*Te(X)]^{-1}Y. \tag{2.8}$$

这一表示表明，多核插值的拟合能力与稳定性，本质上由基础核组合后的联合特征谱 T 以及数据的经验投影 $e(X)$ 共同决定。

2.4. 理论假设与多核曲率特征

为推导多核插值估计的泛化误差的数据依赖上界，我们首先给出数据分布与模型结构的基础假设。

假设 2.1 (高维机制). 存在常数 $c, C \in (0, \infty)$，使得维度与样本量之比满足 $c \le d/n \le C$。记总体协方差矩阵 $\Sigma_d = \mathbb{E}_\mu[xx^\top]$，其算子范数有界 $||\Sigma_d||_{op} \le 1$。

假设 2.2 (高阶矩条件). 数据处理后的随机向量 $z_i = \Sigma_d^{-1/2}x_i$ 的各分量 $z_i(k)(1 \le k \le d)$ 独立同分布，满足 $\mathbb{E}z_i(k) = 0$，$\text{Var}(z_i(k)) = 1$，且存在常数 $m > 0$，使得 $|z_i(k)| \le C \cdot d^{\frac{2}{8+m}}$。

假设 2.3 (噪声有界). 存在常数 $\sigma > 0$ 使得 $\mathbb{E}[(f_*(x) - y)^2 | x = x] \leq \sigma^2$ 几乎处处成立。

上述三条为处理高维随机矩阵的标准设定。而决定多核插值估计泛化特性的核心，在于基础核函数的非线性映射结构：

假设 2.4 (有界非线性多核组合). 假设所有基础核 K_m 均为内积核，即具有形式 $K_m(x, x') = h_m(\frac{1}{d}\langle x, x'\rangle)$，其中 $h_m : \mathbb{R} \to \mathbb{R}$ 在零点邻域光滑。此时，多核函数必然继承内积核的结构：

$$K_{\text{multi}}(x, x') = \sum_{m=1}^{M} \omega_m h_m \left(\frac{1}{d}\langle x, x'\rangle\right) \triangleq h_{\text{multi}} \left(\frac{1}{d}\langle x, x'\rangle\right) \tag{2.9}$$

其中，非线性映射函数 $h_{\text{multi}}(\cdot)$ 即为各基础映射函数的加权组合。且对任意 $x \in \Omega$，存在常数 $A_m \in (0, \infty)$，使得 $K_m(x, x) \leq A_m, \quad 1 \leq m \leq M$, 则对于多核函数 $K_{multi}(x, x) = \sum_{m=1}^{M} \omega_m K_m(x, x) \leq \sum_{m=1}^{M} \omega_m A_m, \quad \sum_{m=1}^{M} \omega_m = 1$, 记 $A_{multi} = \sum_{m=1}^{M} \omega_m A_m$.

接下来定义函数 h_m 与 h_{multi} 的曲率参数：

$$\begin{aligned}
&\alpha_m := h_m(0) + h_m''(0)\frac{\operatorname{Tr}(\Sigma_d^2)}{d^2}, \alpha_{multi} := \sum_{m=1}^{M} \omega_m \alpha_m, \\
&\beta_m := h_m'(0), \beta_{multi} := \sum_{m=1}^{M} \omega_m \beta_m, \\
&\gamma_m := h_m\left(\frac{\operatorname{Tr}(\Sigma_d)}{d}\right) - h_m(0) - h_m'(0)\frac{\operatorname{Tr}(\Sigma_d)}{d}, \gamma_{multi} := \sum_{m=1}^{M} \omega_m \gamma_m.
\end{aligned} \tag{2.10}$$

在单核插值理论中，参数 γ 代表了非线性核函数偏离线性映射的“弯曲程度”。单核插值的局限在于其 γ 值是固定的：一旦核函数及其带宽选定，模型在面对具有不同噪声水平或谱衰减特性的复杂数据时，便失去了调节方差的能力。本文提出的多核插值估计通过引入自适应权重 ω_m，使得模型能够在训练过程中动态合成出一个最优的隐式正则化参数 γ_{multi}。当数据噪声较大时，模型可倾向于赋予高曲率核更大的权重以增强隐式正则化；当目标函数极为平滑时，又可降低曲率以减小逼近偏差。多核学习为核插值模型提供了自适应的隐式正则化能力。

2.5. 多核插值泛化误差数据依赖上界

本文提出的多核插值估计量，其泛化能力本质上取决于对偏差与方差的动态权衡。一方面，面对高噪声或极端谱衰减的复杂数据，多核模型能够自适应地提升高曲率核函数 (如小带宽核) 的权重。这一机制动态提高了复合多核的隐式曲率下界，从而在不依赖任何显式正则化参数的情况下，稳健地压制了奇异方向上的方差波动。另一方面，通过重塑经验核矩阵的谱结构，多核优化加速了尾部特征值的衰减，使得模型能够在更小的截断维度上取得最优解，从而为偏差项提供了更紧致的上界。接下来的定理2.1将严格量化这一过程。

定理 2.1. 在假设 2.1-2.4 成立且维度 d 充分大的高维设定下，定义多核插值估计泛化误差的方差项 (V) 与偏差项 (B) 之和为 $\phi_{n,d}(X, f_*)$：

$$\begin{aligned}\phi_{n,d}(X,f_*) &= V+B \\ &= \frac{8\sigma^2||\Sigma_d||_{op}}{d}\sum_j \frac{\lambda_j\left(\frac{XX^\top}{d}+\frac{\alpha_{\text{multi}}}{\beta_{\text{multi}}}\mathbf{1}\mathbf{1}^\top\right)}{\left[\frac{\gamma_{\text{multi}}}{\beta_{\text{multi}}}+\lambda_j\left(\frac{XX^\top}{d}+\frac{\alpha_{\text{multi}}}{\beta_{\text{multi}}}\mathbf{1}\mathbf{1}^\top\right)\right]^2} \\ &\quad + ||f_*||_{\mathcal{H}}^2 \inf_{0\le k\le n}\left\{\frac{1}{n}\sum_{j>k}\lambda_j(K_{\text{multi}X}K_{\text{multi}X}^\top)+2A_{\text{multi}}\sqrt{\frac{k}{n}}\right\}.\end{aligned} \tag{2.11}$$

则对于设计矩阵 X 的随机抽样，以至少 $1-2\delta-d^{-2}$ 的高概率，多核插值估计满足：

$$\mathbb{E}_{Y|X}||\hat{f}_{\text{multi}}-f_*||_{L^2_\mu}^2 \le \phi_{n,d}(X,f_*)+\epsilon(n,d). \tag{2.12}$$

其中高阶余项 $\epsilon(n,d)=O(d^{-\frac{m}{8+m}}\log^{4.1}d)+O(n^{-\frac{1}{2}}\log^{0.5}(n/\delta))$。

定理给出的多核插值估计的泛化误差上界 $\phi_{n,d}(X,f_*)$ 分为两部分：第一项对应于多核插值估计量的方差项上界 (Variance Bound, V)，第二项对应于多核插值估计偏差项上界 (Bias Bound, B)。

与经典的基于 Tikhonov 正则化 (如岭回归) 的理论结果截然不同，该上界完全是一个数据依赖 (Data-dependent) 且无需显式正则化参数的非渐近界。并且，多核学习 (MKL) 在良性过拟合机制下的两大核心贡献：

1. 对关于方差项 (V)： 多核方差项的上界高度依赖于隐式曲率比值 $\frac{\gamma_{\text{multi}}}{\beta_{\text{multi}}}$ 以及经验协方差矩阵特征值 λ_j 的衰减规律。在极端情况下 (例如采用线性核 $h(a)=a$)，曲率参数 $\gamma=0$；此时若数据矩阵存在极小的特征值 $\lambda_j\to 0$，分母将趋于零，导致方差项发散 (即经典的过拟合崩溃)。多核学习方法的引入，为模型提供了自适应的隐式正则化。由于 $\gamma_{\text{multi}}=\sum_{m=1}^M\omega_m\gamma_m$，在面对高噪声数据或恶劣的谱衰减时，多核优化算法可以通过提升高曲率核函数 (如小带宽的高斯核) 的权重 ω_m，提高 γ_{multi} 的下界，从而在完全没有引入显式正则化参数 λ 的情况下，稳定了核插值估计在奇异方向上的波动，极大地增强了模型对方差的鲁棒性。

2. 对关于偏差项 (B) 多核插值估计偏差项的上界主要由两部分决定：目标函数在再生核希尔伯特空间中的范数 $\|f_*\|_{\mathcal{H}}^2$，以及经验多核算子的特征值尾部衰减速度 $\inf_k\{\lambda_k(K_{\text{multi}X}K_{\text{multi}X}^*)+\dots\}$。对于单一固定的核函数，如果其特征空间与真实目标函数 f_* 的频率成分不匹配, 不仅会导致 $\|f_*\|_{\mathcal{H}}^2$ 极大 (甚至无穷大)，而且其特征值衰减轨迹 λ_k 无法有效截断，使得偏差界居高不下。多核学习通过改变权重 ω_m ，重新调整了经验核矩阵的谱结构 (Spectral Structure) 使得截断指标 k 可以在更小的维度上取得最优解，加速了尾部特征值 λ_k 的衰减，从而使得偏差项具备更紧致的上界。

2.6. 自适应多核插值估计算法

前文的理论分析表明，多核插值估计的泛化能力从根本上依赖于非线性曲率参数 γ_{multi} 对数据未知谱衰减规律的自适应匹配。在实际计算中，如何摆脱人工先验，通过数据驱动的方式寻找最优的基础核函数及其凸组合权重，是实现上述隐式正则化机制的关键。为此，本节提出一种自适应多核插值算法。该算法包含两个核心模块：基于数据距离分布的带宽候选集生成，以及基于

验证集误差的权重凸优化。

2.6.1. 基于数据距离分布的候选带宽生成

在构建多核集合 $\{K_m\}_{m=1}^{M}$ 时，基础核 (如高斯核) 的带宽参数直接决定了特征映射的尺度，固定带宽或单一的网格搜索往往难以覆盖高维复杂数据的多尺度几何结构，为避免人工指定带宽参数导致的先验偏差，本文基于 Schölkopf 等 [22] 和 Gretton 等 [23] 的思路, 提出完全由数据驱动的自适应带宽选择策略。

该自适应带宽生成策略基于样本间距离分布，给定训练集数据矩阵 $X \in \mathbb{R}^{n\times d}$，首先计算所有样本对之间的非零成对欧氏距离集合 $\mathcal{D} = \{\|x_i - x_j\|_2\}_{1\le i<j\le n}$，并提取其 5% 与 95% 分位数，分别记为 $d_{5\%}$ 与 $d_{95\%}$ 。为确保待选的基础核函数中既包含能捕捉高频剧烈变化的“小尺度核”，也包含能刻画全局平滑趋势的“大尺度核”，候选带宽在此边界的基础之上向两端进行扩展，设定实际的带宽搜索区间为 $[0.1\times d_{5\%}, 3\times d_{95\%}]$ 。随后，在该扩展区间内呈几何级数生成 M 个候选带宽序列 $\{\eta_m\}_{m=1}^{M}$ 。

该自适应带宽生成策略生成的带宽集合无需任何人工预设区间，锚定于当前数据集的内在几何分布，从而为后续权重的稀疏选择提供了完备的候选基础核函数集合。

2.6.2. 验证集误差驱动的权重凸优化

在确立了 M 个基础核之后，多核插值估计模型的拟合完全由权重向量 $\omega = (\omega_1, \ldots, \omega_M)^\top$ 决定。为避免在训练集上直接优化权重导致退化解，本文引入验证集机制进行权重凸优化。将原数据集划分为训练集 (X_{tr}, Y_{tr}) 与验证集 (X_{val}, Y_{val})，对于给定的权重向量 ω，多核插值估计量在训练集上的闭式解为：

$$\hat{f}_{multi}(x;\omega) = K_{multi}(x, X_{tr};\omega) K_{multi}(X_{tr}, X_{tr};\omega)^{-1} Y_{tr},$$

权重优化的目标是最小化该估计量在验证集上的均方误差，由于多核插值估计具有尺度不变性，若不对权重施加限制，优化目标将出现无数个最优解，面临尺度模糊性，导致梯度下降算法无法稳定收敛。因此，该问题需要被严格表述为单纯形约束下的优化问题：

$$\begin{aligned} \min_{\omega} \quad & \mathcal{L}_{val}(\omega) = \frac{1}{|X_{val}|} \left\| Y_{val} - \hat{f}_{multi}(X_{val};\omega) \right\|_2^2 \\ \text{s.t.} \quad & \sum_{m=1}^{M} \omega_m = 1, \quad \omega_m \ge 0, \forall m. \end{aligned}$$

由于 K_{multi} 是 ω 的仿射组合，上述目标函数是关于 ω 的连续可微函数，本文采用投影梯度下降法 (PGD) 进行求解，优化结束时，那些与数据真实特征谱衰减不匹配的基础核将被赋予接近于零的权重，实现了多核插值估计的稀疏化选择。

3. 理论证明

在展开严密的数学推导之前，本节首先简要概述定理 2.1 的证明路线图。要证明多核插值估计量在高维设定下的有效性，必须保证其方差与偏差的非渐近上界均足够小且收敛。因此，证明分

为以下三个核心步骤：

偏差 -方差精确分解：首先，利用再生核希尔伯特空间的再生性质，将多核插值估计的期望泛化误差严格分解为方差项 (V) 与偏差项 (B) 。

基于线性逼近的方差界估计：为了处理复杂的非线性多核算子，我们引入高维随机矩阵理论。核心逻辑在于证明高维机制下，非线性多核矩阵的谱行为可以被其一阶泰勒展开 (即线性多核近似矩阵) 以高概率紧密逼近，从而通过分析线性主导项的迹运算来获得方差的非渐近上界。

基于经验过程理论的偏差界估计：偏差项的分析涉及在无限维函数空间中对目标函数投影误差的界定。为此，我们借助于经验过程理论，通过构建投影矩阵函数类，并利用 Rademacher 对称化技术与复杂度控制，将总体分布下的偏差转化为有限样本特征值尾部衰减的经验界。

为证明定理 2.1，我们首先利用引理(3.1) 将多核插值估计的泛化误差分解为偏差项 (Bias) 和方差项 (Variance)。随后，我们将结合高维随机矩阵理论与经验过程理论，分别在定理 3.1和定理 3.2中给出这方差项和偏差项的数据依赖型上界。

3.1. 偏差 -方差分解 (Bias-Variance Decomposition)

回顾多核插值估计的闭式解为 $\hat{f}_{\text{multi}}(x) = K_{\text{multi}}(x, X)K_{\text{multi}}(X, X)^{-1}Y$。其中 $K_{multi}(X, X) = \sum_{m=1}^{M} \omega_m K_m(X, X)$ 为多核矩阵。由再生核希尔伯特空间的再生性，多核 K_{multi} 在点 x 的特征映射 $K_{\text{multi}_x} : \mathbb{R} \to \mathcal{H}_{\text{multi}}$ 满足 $f(x) = \langle K_{\text{multi}_x}, f\rangle_{\mathcal{H}_{\text{multi}}} = K^*_{\text{multi}_x} f$。

下述引理给出了多核插值估计精确的偏差 -方差分解。只有当方差项与偏差项的非渐近上界均足够小且收敛时，才能保证多核插值估计的有效性。

引理 3.1. 多核插值估计量 $\hat{f}_{\text{multi}}$ 满足如下均方误差分解：

$$\mathbb{E}_{Y|X}||\hat{f}_{\text{multi}} - f_*||^2_{L^2_\mu} = V + B. \tag{3.1}$$

其中，方差项 V 与偏差项 B 分别定义为：

$$V := \int \mathbb{E}_{Y|X}|K^*_{\text{multi}_x} K_{\text{multi}X}(K^*_{\text{multi}X} K_{\text{multi}X})^{-1}(Y - \mathbb{E}[Y|X])|^2 d\mu(x), \tag{3.2}$$

$$B := \int |K^*_{\text{multi}_x}[K_{\text{multi}X}(K^*_{\text{multi}X} K_{\text{multi}X})^{-1} K^*_{\text{multi}X} - I]f_*|^2 d\mu(x). \tag{3.3}$$

证明. 定义观测噪声带来的残差向量为 $E = Y - \mathbb{E}[Y|X] = Y - f_*(X)$。由条件期望性质，显然有 $\mathbb{E}_{Y|X}[E] = 0$，可得估计误差为：

$$\begin{aligned}\hat{f}_{\text{multi}}(x) - f_*(x) = {} & K^*_{\text{multi}_x} K_{\text{multi}X}(K^*_{\text{multi}X} K_{\text{multi}X})^{-1}(f_*(X) + E) - f_*(x) \\ & + K^*_{\text{multi}_x}[K_{\text{multi}X}(K^*_{\text{multi}X} K_{\text{multi}X})^{-1} K^*_{\text{multi}X} - I]f_*.\end{aligned} \tag{3.4}$$

对上式两端平方后取关于条件分布 $Y|X$ 的期望。由于 $\mathbb{E}_{Y|X}[E] = 0$，上述两项的交叉项期望严格为零。

由 Fubini 定理，交换积分与期望的顺序，得

$$\mathbb{E}_{Y|X}\left\|\hat{f}_{multi}-f_*\right\|_{L^2_\mu}^2=\int \mathbb{E}_{Y|X}\left|K_{multi_x}^* K_{multiX}\left(K_{multi_X}^* K_{multiX}\right)^{-1}E\right|^2 d\mu(x)$$
$$+\int\left|K_{multi_x}^*\left[K_{multiX}\left(K_{multi_X}^* K_{multiX}\right)^{-1}K_{multi_X}^*-I\right]f_*\right|^2 d\mu(x).$$

证毕。 □

3.2. 方差项的上界估计

为估计多核插值估计的方差项 V，我们首先需要引入两个关于高维核矩阵谱范数逼近的辅助引理。在高维机制下，非线性核矩阵可以被其一阶泰勒展开项 (即线性核部分) 紧密逼近。

结合本文对多核函数的定义，我们构造如下线性多核近似矩阵及向量：

$$\begin{aligned}K^{\text{lin}}(X,X)&:=\gamma_{\text{multi}}I+\alpha_{\text{multi}}\mathbf{1}\mathbf{1}^\top+\beta_{\text{multi}}\frac{XX^\top}{d}\in\mathbb{R}^{n\times n},\\ K^{\text{lin}}(X,x)&:=\beta_{\text{multi}}\frac{Xx^\top}{d}\in\mathbb{R}^{n\times 1}.\end{aligned}\tag{3.5}$$

引理 3.2 (多核矩阵的线性近似)**.** 在假设 2.1、2.2 与 2.4 成立的条件下，令 $\theta=\frac{1}{2}-\frac{2}{8+m}$。对于设计矩阵 X 的随机抽样，若维度 d 充分大，则以至少 $1-\delta-d^{-2}$ 的概率有：

$$||K_{\text{multi}}(X,X)-K^{\text{lin}}(X,X)||_{op}\le d^{-\theta}(\delta^{-1/2}+\log^{0.51}d).\tag{3.6}$$

该引理的核心思想源自 El Karoui [24] 关于高维核随机矩阵谱逼近的经典结果，并由 Liang 等 [13] 在弱矩条件下给出了非渐近界 (详见其补充材料 Proposition A.2)。本文将其结论推广至基础核函数的凸组合空间中，由于 $\sum\omega_m=1$ 且 $\omega_m\ge 0$，线性可加性保证了多核矩阵同样满足该逼近误差率。

引理 3.3. 在与引理 3.2 相同的条件下，对于任意固定的样本外点 $x\sim\mu$，以至少 $1-d^{-2}$ 的高概率，多核向量与其线性近似在期望意义下满足：

$$\mathbb{E}_\mu||K_{\text{multi}}(x,X)-K^{\text{lin}}(x,X)||^2\le d^{-(4\theta-1)}\log^{4.1}d.\tag{3.7}$$

该引理的证明逻辑参考了文献 [13] 补充材料中的 Lemma B.2。通过对多核函数进行泰勒展开，并结合 Bernstein 不等式 [25] 与高维随机向量的范数集中性，可将误差控制在多项式衰减级别。

基于上述两个引理，我们现在给出方差项的上界估计。

定理 3.1. 设 $\delta\in(0,1)$。在假设 2.1-2.4 成立的条件下，对于设计矩阵 X 的随机抽样，以至少 $1-\delta-d^{-2}$ 的概率，当维度 d 充分大时，方差项 V 满足：

$$V\le\frac{8\sigma^2||\Sigma_d||_{op}}{d}\sum_j\frac{\lambda_j\left(\frac{XX^\top}{d}+\frac{\alpha_{\text{multi}}}{\beta_{\text{multi}}}\mathbf{1}\mathbf{1}^\top\right)}{\left[\frac{\gamma_{\text{multi}}}{\beta_{\text{multi}}}+\lambda_j\left(\frac{XX^\top}{d}+\frac{\alpha_{\text{multi}}}{\beta_{\text{multi}}}\mathbf{1}\mathbf{1}^\top\right)\right]^2}+\frac{8\sigma^2}{\gamma_{\text{multi}}^2}d^{-(4\theta-1)}\log^{4.1}d.\tag{3.8}$$

证明. 根据 V 的定义式 (3.2)，由噪声的独立性可知，当 $i \neq j$ 时，$\mathbb{E}_{Y|X}[(Y_i - f_*(X_i))(Y_j - f_*(X_j))] = 0$，且由假设 2.3 可得该协方差矩阵的算子范数 $||\mathbb{E}_{Y|X}[EE^\top]||_{op} \leq \sigma^2$。

利用矩阵迹的循环性质与算子范数放缩，方差项可化简为：

$$\begin{aligned} V &\leq \sigma^2 \int ||(K^*_{\text{multi}X} K_{\text{multi}X})^{-1} K^*_{\text{multi}X} K_{\text{multi}_x}||^2 d\mu(x) \\ &= \sigma^2 \mathbb{E}_\mu ||K_{\text{multi}}(X,X)^{-1} K_{\text{multi}}(X,x)||^2. \end{aligned} \tag{3.9}$$

通过引入式 (3.5) 定义的 $K^{\text{lin}}(X,X)$ 和 $K^{\text{lin}}(X,x)$，并利用三角不等式的平方展开，我们将目标项拆分为线性主导项与高阶余项：

$$\begin{aligned} V \leq\ & 2\sigma^2 \mathbb{E}_\mu ||K_{\text{multi}}(X,X)^{-1} K^{\text{lin}}(X,x)||^2 \\ & + 2\sigma^2 ||K_{\text{multi}}(X,X)^{-1}||^2_{op} \cdot \mathbb{E}_\mu ||K_{\text{multi}}(X,x) - K^{\text{lin}}(X,x)||^2. \end{aligned} \tag{3.10}$$

由于当 d 充分大时，由引理 3.2和 Weyl 不等式可得 $||K_{\text{multi}}(X,X)^{-1}||_{op} \leq \frac{2}{\gamma_{\text{multi}}}$。代入引理 3.3 对 x 向量的逼近结果，式 (3.10) 的第二项以 $\frac{8\sigma^2}{\gamma^2_{\text{multi}}} d^{-(4\theta-1)} \log^{4.1} d$ 为上界。

对于第一项 (线性主导项)，我们有：

$$||K_{\text{multi}}(X,X)^{-1} K^{\text{lin}}(X,X)||_{op} \leq 1 + ||K_{\text{multi}}(X,X)^{-1}||_{op} ||K_{\text{multi}} - K^{\text{lin}}||_{op} \leq 2, \tag{3.11}$$

将 K^{lin} 的定义代入迹运算，并利用 $\mathbb{E}_\mu[xx^*] = \Sigma_d$ 可得：

$$\begin{aligned} & \mathbb{E}_\mu \left\| K_{\text{multi}}(X,X)^{-1} K^{\text{lin}}(X,x) \right\|^2 \\ & \leq \left\| K_{\text{multi}}(X,X)^{-1} K^{\text{lin}}(X,X) \right\|^2_{\text{op}} \cdot \mathbb{E}_\mu \left\| K^{\text{lin}}(X,X)^{-1} K^{\text{lin}}(X,x) \right\|^2 \\ & \leq \frac{4}{d} \left\| \Sigma_d \right\|_{\text{op}} \sum_j \frac{\lambda_j \left(\frac{XX^\top}{d} + \frac{\alpha_{\text{multi}}}{\beta_{\text{multi}}} \mathbf{1}\mathbf{1}^\top \right)}{\left(\frac{\gamma_{\text{multi}}}{\beta_{\text{multi}}} + \lambda_j \left(\frac{XX^\top}{d} + \frac{\alpha_{\text{multi}}}{\beta_{\text{multi}}} \mathbf{1}\mathbf{1}^\top \right) \right)^2}, \end{aligned} \tag{3.12}$$

综上，对于设计矩阵 X 的随机抽样，以至少 $1 - \delta - d^{-2}$ 的概率有

$$V \leq 2\sigma^2 \mathbb{E}_\mu \left\| K^{\text{lin}}(X,X)^{-1} K^{\text{lin}}(X,x) \right\|^2 + \frac{8\sigma^2}{\gamma^2_{multi}} d^{-(4\theta-1)} \log^{4.1} d,$$

且当 d 充分大时，

$$V \leq \frac{8\sigma^2 \left\| \Sigma_d \right\|_{\text{op}}}{d} \sum_j \frac{\lambda_j \left(\frac{XX^\top}{d} + \frac{\alpha_{multi}}{\beta_{multi}} \mathbf{1}\mathbf{1}^\top \right)}{\left[\frac{\gamma_{multi}}{\beta_{multi}} + \lambda_j \left(\frac{XX^\top}{d} + \frac{\alpha_{multi}}{\beta_{multi}} \mathbf{1}\mathbf{1}^\top \right) \right]^2} + \frac{8\sigma^2}{\gamma^2_{multi}} d^{-(4\theta-1)} \log^{4.1} d.$$

证毕。 □

3.3. 偏差项的上界估计

为处理偏差项中的无限维函数空间操作，我们需要借助于经验过程理论。在此之前，先引入经验风险最小化分析中常用的关键引理。

引理 3.4. 设函数族 $\mathcal{G}$ 满足对于任意 $g \in \mathcal{G}$ 有 $|g(x)| \leq M$。则对于 i.i.d. 抽样的 $x_i \sim \mu$，以至少 $1-2\delta$ 的高概率有：

$$\sup_{g\in\mathcal{G}} |\mathbb{E}g(x) - \hat{\mathbb{E}}_n g(x)| \leq 2\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(x_i) + 3M\sqrt{\frac{\log(1/\delta)}{2n}}, \tag{3.13}$$

其中 ϵ_i 为独立的 Rademacher 随机变量，$\mathbb{E}_\epsilon$ 为关于 ϵ_i 的期望。

这是经验过程与统计学习理论中的标准结果，可通过 McDiarmid 不等式与 Rademacher 对称化技术直接推导得出 (参考 [4] 及 [13] 补充材料 Lemma B.4)。

定理 3.2. 令 $\delta \in (0,1)$。在假设 2.4 的条件下，多核函数一致有界 $\sup_x K_{\text{multi}}(x,x) \leq A_{\text{multi}}$。对于设计矩阵 X 的随机抽样，以至少 $1-2n\delta$ 的概率，偏差项 B 满足上界：

$$B \leq ||f_*||^2_{\mathcal{H}_{\text{multi}}} \cdot \inf_{0\leq k\leq n} \left\{ \frac{1}{n}\sum_{j>k} \lambda_j(K_{\text{multi}}(X,X)) + 2A_{\text{multi}}\sqrt{\frac{k}{n}} \right\} + 3A_{\text{multi}}\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{3.14}$$

证明. 我们将目标函数展开为其在特征基 $\{e_i(x)\}$ 下的级数 $f(x) = e(x)^\top f_*$。利用核积分算子的谱分解表示 $K_{\text{multi}}(x,z) = e(x)^\top T e(z)$，我们将偏差项式 (3.3) 改写为

$$B \leq \|f_*\|^2_{\mathcal{H}_{multi}} \int \left\| \left[T^{1/2}e(X) \left(e^\top(X)Te(X)\right)^{-1} e^\top(X)T^{1/2} - I \right] T^{1/2}e(x) \right\|^2 d\mu(x). \tag{3.15}$$

回顾经验核算子的谱分解 $\hat{T} = \hat{U}\hat{\Lambda}\hat{U}^*$，其中对角元 $\hat{\Lambda}_{jj} = \frac{1}{n}\lambda_j(K(X,X))$。我们考虑前 k 个最大特征值, 记 $\hat{U}$ 的前 k 列为 $\hat{U}_k$，$P^\perp_{\hat{U}_k}$ 为正交于 $\hat{U}_k$ 张成空间的投影算子, 因此对所有 $k \leq n$，有

$$B \leq \|f_*\|^2_{\mathcal{H}_{multi}} \int \left\| P^\perp_{\hat{U}_k} \left(T^{1/2}e(x)\right) \right\|^2 d\mu(x). \tag{3.16}$$

接下来利用 [26] 的方法分析(3.16)的积分项。

对任意秩为 k 的投影矩阵 U_k，定义函数：

$$g_{U_k}(x) := \|P_{U_k}(T^{1/2}e(x))\|^2 = \text{Tr}(e^\top(x)T^{1/2}U_kU_k^TT^{1/2}e(x)). \tag{3.17}$$

显然有 $\|U_kU_k^T\|_F = \sqrt{k}$。据此定义函数类：

$$\mathcal{G}_k := \left\{ g_{U_k}(x) : U_k^TU_k = I_k \right\}. \tag{3.18}$$

易知对应于经验多核算子的投影 $\hat{g}_{\hat{U}_k} \in \mathcal{G}_k$，且 $g_{\hat{U}_k}$ 为依赖于数据 X 的随机函数。我们将利用

经验过程理论对偏差项作非渐近界估计。

首先，直接计算总体期望与经验期望可得：

$$\mathbb{E}_{x\sim\mu}\|P^{\perp}_{\hat{U}_k}(T^{1/2}e(x))\|^2=\int\|P^{\perp}_{\hat{U}_k}(T^{1/2}e(x))\|^2d\mu(x), \tag{3.19}$$

$$\hat{\mathbb{E}}_n\|P^{\perp}_{\hat{U}_k}(T^{1/2}e(x))\|^2=\frac{1}{n}\sum_{i=1}^{n}\|P^{\perp}_{\hat{U}_k}(T^{1/2}e(x_i))\|^2=\frac{1}{n}\sum_{j>k}\lambda_j(K_{\mathrm{multi}}(X,X)). \tag{3.20}$$

令 $A_{\mathrm{multi}}=\sup_{x\in\Omega}K_{\mathrm{multi}}(x,x)$。由于 $\sup_{g\in\mathcal{G}_k}|g(x)|\le\|T^{1/2}e(x)\|^2=K_{\mathrm{multi}}(x,x)\le A_{\mathrm{multi}}$，满足有界性条件。由引理(3.4)，对于设计矩阵 X 的随机抽样，以至少 $1-2\delta$ 的概率有：

$$\begin{aligned}&\int\|P^{\perp}_{\hat{U}_k}(T^{1/2}e(x))\|^2d\mu(x)-\frac{1}{n}\sum_{j>k}\lambda_j(K_{\mathrm{multi}}(X,X))\\&\le 2\mathbb{E}_{\epsilon}\sup_{U_k:U_k^TU_k=I_k}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\|T^{1/2}e(x_i)\|^2-\|P_{U_k}(T^{1/2}e(x_i))\|^2)+3A_{\mathrm{multi}}\sqrt{\frac{\log(1/\delta)}{2n}}.\end{aligned} \tag{3.21}$$

由于 Rademacher 随机变量 ϵ_i 对称且期望为 0，且 $\|T^{1/2}e(x_i)\|^2$ 与 U_k 无关，故上式中包含 $\|T^{1/2}e(x_i)\|^2$ 的项在取上确界时可被消去，简化为：

$$2\mathbb{E}_{\epsilon}\sup_{g\in\mathcal{G}_k}\frac{1}{n}\sum_{i=1}^{n}\epsilon_ig(x_i)+3A_{\mathrm{multi}}\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{3.22}$$

接下来对函数类 $\mathcal{G}_k$ 的 Rademacher 复杂度作界估计。利用矩阵内积的性质，我们有：

$$\mathbb{E}_{\epsilon}\sup_{g\in\mathcal{G}_k}\frac{1}{n}\sum_{i=1}^{n}\epsilon_ig(x_i)\le\frac{\sqrt{k}}{n}\mathbb{E}_{\epsilon}\left\|\sum_{i=1}^{n}\epsilon_iT^{1/2}e(x_i)e^{\top}(x_i)T^{1/2}\right\|_F, \tag{3.23}$$

进一步，利用 ϵ_i 的独立性，计算 Frobenius 范数平方的期望：

$$\frac{\sqrt{k}}{n}\left\{\mathbb{E}_{\epsilon}\left\|\sum_{i=1}^{n}\epsilon_iT^{1/2}e(x_i)e^{\top}(x_i)T^{1/2}\right\|_F^2\right\}^{1/2}=\sqrt{\frac{k}{n}}\sqrt{\frac{\sum_{i=1}^{n}K_{\mathrm{multi}}(x_i,x_i)^2}{n}}. \tag{3.24}$$

综上所述，通过对所有可能的截断指标 $k\le n$ 运用并集界 (Union Bound)，对于设计矩阵 X 的随机抽样，以至少 $1-2n\delta$ 的概率有：

$$\begin{aligned}B\le\|f_*\|^2_{\mathcal{H}_{\mathrm{multi}}}\cdot\inf_{0\le k\le n}\Bigg\{&\frac{1}{n}\sum_{j>k}\lambda_j(K_{\mathrm{multi}}(X,X))\\&+2\sqrt{\frac{k}{n}}\sqrt{\frac{\sum_{i=1}^{n}K_{\mathrm{multi}}(x_i,x_i)^2}{n}}+3A_{\mathrm{multi}}\sqrt{\frac{\log(1/\delta)}{2n}}\Bigg\}.\end{aligned} \tag{3.25}$$

考虑到多核函数的有界性，对所有样本点均满足 $K_{\text{multi}}(x_i, x_i) \leq A_{\text{multi}}$，因此该上界可进一步被一致放缩为：

$$B \leq \|f_*\|^2_{\mathcal{H}_{\text{multi}}} \cdot \left(\inf_{0\leq k\leq n} \left\{ \frac{1}{n} \sum_{j>k} \lambda_j(K_{\text{multi}}(X,X)) + 2A_{\text{multi}} \sqrt{\frac{k}{n}} \right\} + 3A_{\text{multi}} \sqrt{\frac{\log(1/\delta)}{2n}} \right). \tag{3.26}$$

证毕。 □

4. 模拟与真实数据分析

为验证本文提出的多核插值估计理论，本节将采用高斯核 ($K(x,x') = exp(\frac{-\|x-x'\|^2}{2\sigma^2})$) 分别在模拟数据与实际数据集进行实验，并通过与单核插值模型进行对比，探讨基于多核学习的核插值方法的改进以及自适应隐式正则化的动态实现。

4.1. 模拟数据

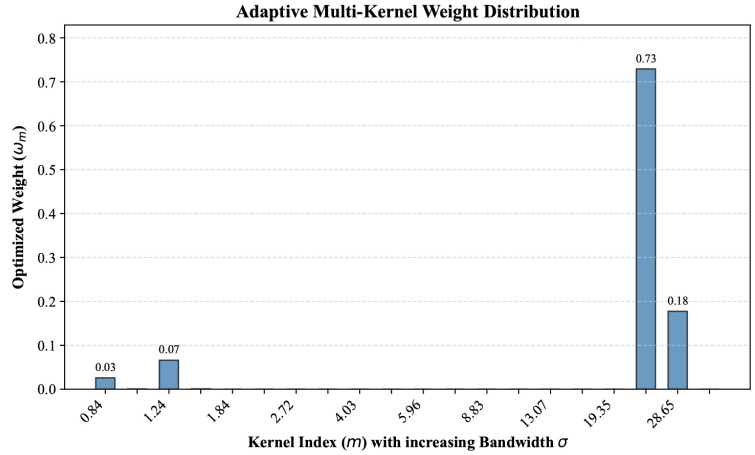
为了直观展示多核插值估计对比单核插值估计在多尺度数据处理上的优势，我们首先设计了一个多尺度回归任务。数据生成过程如下：协变量 $x \in \mathbb{R}^{50}$ 服从标准正态分布，维数 $d = 50$，样本量为 $n = 1000$(按 5:2:3 划分为训练集、验证集与测试集)。真实的条件期望函数 $f_*(x)$ 由两个不同尺度的高斯核函数线性组合生成 (真实尺度分别设定为 $\sigma_{\text{small}} = 2$ 与 $\sigma_{\text{large}} = 25$)。观测标签引入了方差为 $\sigma^2 = 1.0$ 的独立高斯噪声。这构成了一个典型的高维噪声插值问题。

$$f^*(x) = \sum_{j=1}^{10} w_j^{(s)} \exp\left(-\frac{\|x - \mathbf{c}_j^{(s)}\|^2}{2\sigma_{small}^2} \right) + \sum_{k=1}^{10} w_k^{(l)} \exp\left(-\frac{\|x - \mathbf{c}_k^{(l)}\|^2}{2\sigma_{large}^2} \right) \tag{4.1}$$

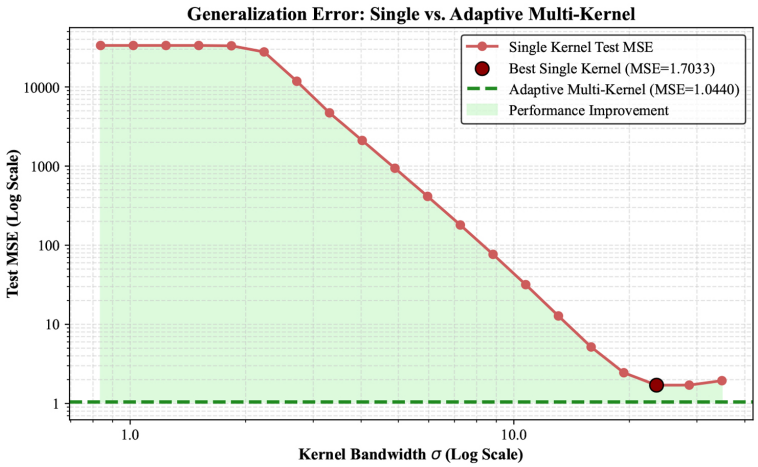
其中 $\omega_j^{(s)}, \omega_k^{(l)}$ 分别表示对应的随机权重，$c_j^{(s)}, s_k^{(l)}$ 分别表示对应尺度的中心点，以上参数均服从标准正态分布。

基于数据分布计算得到的带宽搜索边界为 $[0.83, 34.86]$，覆盖了潜在的真实尺度。经过自适应多核插值估计优化，模型在 $M = 20$ 个基础核中呈现出显著的稀疏性与双峰分布，集中分布于极小尺度与极大尺度两端，表明模型自发进行了特征选择。(图 1(a)所示)。

最终，测试集上的量化表现如图 1(b)所示：在遍历所有 20 个独立单核插值模型后，最优单核插值的测试均方误差仅为 1.7033；而自适应多核插值模型的测试均方误差降至 1.0440。在保持训练集 0 误差 (完全插值) 的条件下，多核模型逼近了不可缩减的噪声下界 (Noise floor ≈ 1.0)，相对于最优单核实现了 **38.7%** 的误差降低。



(a) 多核权重的自适应分布



(b) 自适应多核与单核的性能比较

Figure 1. Simulation experiment

图 1. 模拟实验

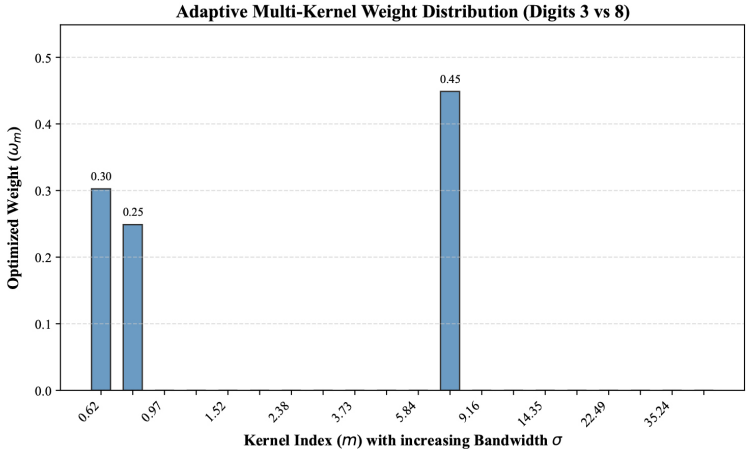
4.2. 真实数据

为进一步在真实数据中验证多核插值估计的有效性，本文选用了 UCI 光学手写数字识别数据集 (Optical Recognition of Handwritten Digits)。该数据集的每个样本为 8×8 的像素矩阵，展平后的特征维度 $d = 64$。

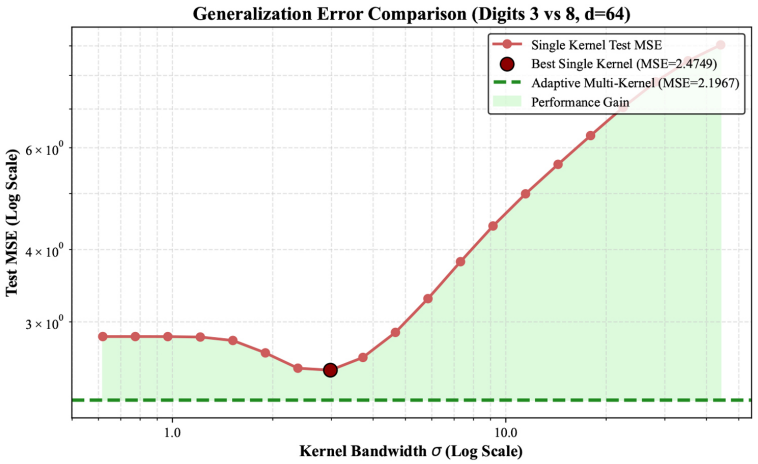
本文提取了数字对“3”和“8”进行实验，将其分类标签映射为 $\{-1, 1\}$ 以转化为连续回归目标。在总计约 357 个目标样本中，随机抽取 150 个样本作为训练集，107 样本作为验证集，100 个样本作为测试集。

此外，实验对图像像素特征进行了标准化处理，并主动向训练标签中注入了方差为 $\sigma^2 = 1.5^2$ 的高斯噪声。这一高噪、高维且高度非线性的设定，构成了验证良性过拟合现象与隐式正则化机制的理想环境。

面对未知的图像像素特征谱衰减规律，固定带宽的单核模型表现出显著的局限性。如图 2(a)所示的误差变化曲线表明，当强制要求单核模型在含有强噪声的训练集上实现零误差插值时，即便是遍历候选库后选出的最优单核进行插值估计，其测试集 MSE 较自适应多核插值估计依然处于较高水平。并且通过验证集驱动的权重的选择，模型在 $M = 20$ 个基础核中实现了权重的稀疏分配(如图图 2(b)所示)。该结果不仅在真实高维数据上复现了核插值的良性过拟合现象，更有力地证明了多核学习在权衡偏差与方差、实现自适应隐式正则化方面的核心理论价值。



(a) 多核权重的自适应分布



(b) 自适应多核与单核的性能比较

Figure 2. Optical recognition of handwritten digits experiment

图 2. 光学手写数据集实验

5. 总结与展望

本文针对高维统计学习中单核插值估计难以自适应适配复杂数据几何特征的问题，成功将核插值的隐式正则化理论拓展至多核学习框架。通过在再生核希尔伯特空间中严格推导偏差 -方差分解，我们证实了多核凸组合并非简单的线性叠加，而是通过权重的自适应调整，动态合成了最优

的隐式正则化。这一机制在无需显式 Tikhonov 正则化项的情况下，有效克服了单核模型因核函数曲率固定而导致的方差发散问题。基于高维随机矩阵理论，我们进一步推导了多核插值估计在有限样本下的非渐近泛化误差上界，从理论上揭示了多核模型通过提升小尺度核权重以抑制高频噪声、利用大尺度核加速特征值衰减来降低泛化误差的机制。

实证方面，本文提出了一种自适应选取带宽生成策略，并在多尺度模拟数据与真实数据集上进行了验证。实验结果表明，在强制训练误差为零的插值条件下，自适应多核模型能够自发地进行稀疏权重分配，其测试集均方误差相比最优单核模型有明显降低，有力印证了多核隐式正则化理论的正确性。

尽管本文在理论与应用上取得了进展，但仍存在一定局限。首先，多核插值涉及核矩阵求逆，计算复杂度较高，未来可结合 Nyström 近似等技术提升计算效率以适应大规模数据。其次，目前的理论推导依赖于亚高斯分布假设，未来研究可尝试将泛化边界推广至重尾分布或非独立同分布数据。此外，鉴于多核学习与宽神经网络在数学形式上的相似性，未来的工作可进一步探究多核插值理论与深度学习泛化机制 (如神经正切核 NTK) 的深层联系，从而为解释深度神经网络中的良性过拟合现象提供新的理论视角。

参考文献

[1] Vapnik, V. (2013) The Nature of Statistical Learning Theory. Springer.

[2] Geman, S., Bienenstock, E. and Doursat, R. (1992) Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, **4**, 1-58. https://doi.org/10.1162/neco.1992.4.1.1

[3] Shalev-Shwartz, S. and Ben-David, S. (2014) Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. https://doi.org/10.1017/cbo9781107298019

[4] Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018) Foundations of Machine Learning. MIT Press.

[5] Belkin, M., Hsu, D., Ma, S. and Mandal, S. (2019) Reconciling Modern Machine-Learning Practice and the Classical Bias-Variance Trade-Off. *Proceedings of the National Academy of Sciences of the United States of America*, **116**, 15849-15854. https://doi.org/10.1073/pnas.1903070116

[6] Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B. and Sutskever, I. (2021) Deep Double Descent: Where Bigger Models and More Data Hurt. *Journal of Statistical Mechanics: Theory and Experiment*, **2021**, Article ID: 124003. https://doi.org/10.1088/1742-5468/ac3a74

[7] Bartlett, P.L., Long, P.M., Lugosi, G. and Tsigler, A. (2020) Benign Overfitting in Linear Regression. *Proceedings of the National Academy of Sciences of the United States of America*, **117**, 30063-30070. https://doi.org/10.1073/pnas.1907378117

[8] Tsigler, A. and Bartlett, P.L. (2023) Benign Overfitting in Ridge Regression. *Journal of Machine Learning Research*, **24**, 1-76.

[9] Frei, S., Chatterji, N.S. and Bartlett, P. (2022) Benign Overfitting without Linearity: Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data. *35th Annual Conference on Learning Theory*, London, 2-5 July 2022, 2668-2703.

[10] Aronszajn, N. (1950) Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, **68**, 337-404. https://doi.org/10.1090/s0002-9947-1950-0051437-7

[11] Berlinet, A. and Thomas-Agnan, C. (2011) Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer.

[12] Wahba, G. (1990) Spline Models for Observational Data. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970128

[13] Liang, T. and Rakhlin, A. (2020) Just Interpolate: Kernel "Ridgeless" Regression Can Generalize. *The Annals of Statistics*, **48**, 1329-1347. https://doi.org/10.1214/19-aos1849

[14] Neyshabur, B., Tomioka, R. and Srebro, N. (2014) In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. arXiv: 1412.6614.

[15] Jacot, A., Gabriel, F. and Hongler, C. (2018) Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montréal, 3-8 December 2018, 8580-8589.

[16] Arora, S., Du, S., Hu, W., Li, Z.Y. and Wang, R.S. (2019) Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. *International Conference on Machine Learning*, Long Beach, 9-15 June, 322-332.

[17] Rakhlin, A. and Zhai, X.Y. (2019) Consistency of Interpolation with Laplace Kernels Is a High-Dimensional Phenomenon. *Conference on Learning Theory*, Phoenix, 25-28 June 2019, 2595-2623.

[18] Lanckriet, G.R.G., Cristianini, N., Bartlett, P., El Ghaoui, L. and Jordan, M.I. (2004) Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research*, **5**, 27-72.

[19] Bach, F.R., Lanckriet, G.R.G. and Jordan, M.I. (2004) Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. *Twenty-First International Conference on Machine Learning—ICML'04*, Banff, 4-8 July 2004, 6. https://doi.org/10.1145/1015330.1015424

[20] Gönen, M. and Alpaydın, E. (2011) Multiple Kernel Learning Algorithms: A Review. *Journal of Machine Learning Research*, **12**, 2211-2268.

[21] Cortes, C., Mohri, M. and Rostamizadeh, A. (2010) Generalization Bounds for Learning Kernels. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Haifa, 21-24 June 2010, 247-254.

[22] Schölkopf, B. and Smola, A.J. (2001) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press. https://doi.org/10.7551/mitpress/4175.001.0001

[23] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B. and Smola, A.J. (2012) A Kernel Two-Sample Test. *Journal of Machine Learning Research*, **13**, 723-773.

[24] El Karoui, N. (2010) The Spectrum of Kernel Random Matrices. *The Annals of Statistics*, **38**, 1-50.

[25] Boucheron, S., Lugosi, G. and Massart, P. (2013) Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press.

[26] Shawe-Taylor, J. and Cristianini, N. (2004) Kernel Methods for Pattern Analysis. Cambridge University Press.