超越交叉验证：基于核的梯度下降的自适应参数选择

Xiaotong Liu ARIESOOMOON@GMAIL.COM
西安交通大学管理学院智能决策与机器学习中心，中国西安

Yunwen Lei LEIYW@HKU.HK
数学系，香港大学，中国香港

Xiangyu Chang XIANGYUCHANG@GMAIL.COM
Shao-Bo Lin* SBLIN1983@GMAIL.COM
西安交通大学管理学院智能决策与机器学习中心，中国西安

arXiv:2603.03401v1 [stat.ML] 3 Mar 2026

编辑：

摘要

本文提出了一种针对基于核的梯度下降（KGD）算法的新型参数选择策略，将偏差-方差分析与分割方法相结合。我们引入经验有效维数这一概念，用以量化 KGD 中的迭代增量，并由此推导出一种可实现的自适应参数选择策略。在学习理论框架下，我们给出了理论验证。借助近期发展的积分算子方法，我们严格证明，采用所提出的自适应参数选择策略的 KGD 能够达到最优泛化误差界，并能有效适应不同的核、目标函数和误差度量。因此，与现有的 KGD 参数选择方法相比，该策略展现出显著优势。

Ke关键词： 参数选择、非参数回归、梯度下降、核 1
me方法

1. 引言

参数选择通常也称为超参数选择，是统计学和机器学习中的一个关键且长期存在的研究课题。在学习算法中选择适当的参数，会显著影响所得模型的准确率、效率和泛化能力。例如，选择合适的正则化参数，可以使支持向量机（Steinwart and Christmann，2008）在偏差与方差之间保持平衡，从而确保其在未见数据上表现出色；而在深度学习中调整随机梯度下降的适当步长，则能够在训练时间与模型质量之间实现平衡（Bengio 等人，2017）。

*。通讯作者

© Xiaotong Liu、Yunwen Lei、Xiangyu Chang 和 Shao-Bo Lin。许可协议：知识共享署名 4.0 国际许可协议，详见 https://creativecommons.org/licenses/by/4.0/。署名要求见 http://jmlr.org/papers/v/.html。

参数选择通常需要采用具有策略性的方法，以确保所选参数与学习任务的具体目标和约束相一致，这也推动了过去二十年来的大量研究（Chapelle 等人，2002；De Vito 等人，2010；Caponnetto 和 Yao，2010；Raskutti 等，2014；Wei 等，2019；Blanchard 等人，2019；Lu 等，2020；Gizewski 等人，2022；Dinu 等人，2023；Lin，2024）。总体而言，参数选择有三种基本思路：信息熵方法、分割方法和偏差-方差分析方法。信息熵方法通过在拟合精度与熵相关复杂度约束之间取得平衡来确定最优参数，在多种线性方法中尤其常用（Reiss and Todd Ogden，2009）。信息熵方法的典型例子包括赤池信息准则（Akaike，1974年；Hurvich 等人，1998）、贝叶斯信息准则（Schwarz，1978；Mathias and Martyn，2017）和广义信息准则（Nishii，1984年）。该方法的主要优点是易于实现，但一个显著缺点是难以为非线性算法推导出可证明的泛化误差界。分割方法侧重于通过将样本划分为训练集和验证集来选择参数，使用前者构建模型，后者进行参数选择。该方法的典型例子包括留出法（Caponnetto 和 Yao，2010）、交叉验证（Gy¨orfi 等人，2002，第8章）和留一法（Zhang，2003年）。分割方法的优点在于适用性广，几乎适用于所有学习算法；其缺点则源于训练过程中排除了一部分样本，这可能导致泛化误差增大。偏差-方差分析方法通过平衡偏差与方差来选择参数，其中偏差与方差使用依赖于数据且可实施的量进行评估。该方法的典型例子包括平衡原理（Lu 等，2020）、Lepskii 原理（Blanchard 等人，2019）和偏差原则（Celisse 和 Wahl，2021）。该方法有利于理论分析，但在实现方面存在挑战，因为要为偏差和方差推导出带有最优常数的精确界限非常困难。

由于其可实现性、快速性与通用性，分割方法，尤其是交叉验证和留出法，在现代统计学和机器学习中占据主导地位。留出法、交叉验证和广义交叉验证的统计性质已分别由Caponnetto 和 Yao（2010年）、Xu等人（2019年）以及Gy¨orfi等人（2002，第8章）充分发展。然而，分割方法常常要求经验超额风险是总体风险的可获得无偏估计（Blanchard 等人，2019），这使得相应的参数选择在应对训练样本与测试样本分布不同的协变量偏移问题时较为薄弱。此外，常用的截断（或裁剪）算子将所得估计投影到有限区间上（Gy¨orfi等人，2002），不仅使参数选择仅适用于有界样本，而且使最终模型偏离原始假设类。这些缺陷的主要原因在于缺乏偏差-方差分析来揭示参数对偏差与方差的具体影响，而这种分析对于解决协变量偏移问题（Ma等，2023）以及处理无界样本（Blanchard和Kr¨amer，2016）至关重要。

本文旨在为经典分割方法配备精细的偏差–方差分析，以规避上述缺陷。以基于核的梯度下降（KGD）（Lin 和 Zhou，2018a；Yao 等人，2007）（文献中也称为基于核的提升

（Raskutti 等，2014；Wei 等，2019））为例，我们提出一种新颖的混合选择方案（HSS），结合留出法（Caponnetto 和 Yao，2010）与 Lepskii 型原理（Blanchard 等人，2019；Lin，2024），自适应地确定 KGD 的迭代次数，并提供最优的理论验证。根据精细的偏差-方差分析，HSS 首先通过连续两次迭代之间的增量量化 KGD 的偏差，并通过经验有效维数（Zhang，2002）量化其方差。随后，考虑到相关量的单调性，它以逆向方式搜索最优迭代次数，即从 $|D|, |D|-1$到 1。由此，我们提出了一种新颖的偏差–方差分析方法，其中包含一个与数据规模无关且未指定的常数。最后，HSS 使用留出法（或交叉验证）等分割方法，在一部分样本上确定该常数。在此情况下，我们得到一种参数选择策略，能够适应不同的核、目标函数和度量，同时不舍弃任何样本。

我们的主要贡献可总结如下。从理论角度来看，我们证明了采用所提出的 HSS 的 KGD 能够达到（Lin 和 Zhou，2018a）所确立的最优泛化误差界，这表明了 HSS 的有效性和最优性。这克服了众多参数选择方案的瓶颈，例如平衡原理（Lu 等，2020）和偏差原则（Celisse 和 Wahl，2021）；这些方法只能得到次优的泛化误差界。我们在证明中的主要创新，是引入了若干半自适应停止规则，用以量化 HSS 所确定的迭代次数范围。从应用角度来看，所提出的 HSS 结合了偏差-方差分析与分裂方法，发挥了它们在参数选择方面的优势，并在实践中表现良好。我们通过玩具模拟和真实数据实验，展示了 HSS 的出色表现，并将其与留出法（Caponnetto 和 Yao，2010）、赤池信息准则（Akaike，1974）、贝叶斯信息准则（Schwarz，1978）、平衡原理（Lu 等，2020）、Lepskii 原理（Blanchard 等人，2019）、早停规则（Raskutti 等，2014）以及偏差原则（Celisse 和 Wahl，2021）进行了比较。

本文其余部分安排如下。下一节中，我们将介绍用于 KGD 的 HSS，并将我们的方法与现有参数选择策略进行比较。第 3 节探讨 HSS 的若干重要性质，并推导采用 HSS 的 KGD 的最优泛化误差界。第 4 节通过数值实验验证所提出停止规则的可行性。第 5 节将作进一步讨论。理论结果的证明见附录。

2. 采用混合选择策略的 KGD

本节介绍采用混合选择策略的 KGD，并将其与一些相关工作进行比较。

2.1 采用混合选择策略的 KGD

设 $D = \{(x_i, y_i)\}_{i=1}^{|D|} \subset \mathcal{X} \times \mathcal{Y}$ 是一个样本集，其 $|D|$ 基数为 D，$\mathcal{X}$是一个紧致输入空间，$\mathcal{Y} \subseteq \mathbb{R}$ 是一个输出空间。设 $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 是一个 Mercer核，$\mathcal{H}_K$ 是其对应的再生核希尔伯特空间（RKHS），并赋予内积

$\langle\cdot,\cdot\rangle_K$ 和范数 $\|\cdot\|_K$。由于 $\mathcal{X}$ 是紧致的且 K 是一个 Mercer核，容易验证 $\kappa := \sqrt{\sup_{x\in\mathcal{X}} K(x,x)} < \infty$。核方法旨在通过最小化经验风险来寻找一个函数 $f \in \mathcal{H}_K$

$$\mathcal{L}(f) := \frac{1}{|D|}\sum_{i=1}^{|D|}(y_i - f(x_i))^2. \tag{1}$$

T关于 $f \in \mathcal{H}_K$ 的 $\mathcal{L}(f)$ 梯度可由 Yao 等人（2007 年 ）

$$\nabla_f \mathcal{L}(f) = \frac{2}{|D|}\sum_{i=1}^{|D|}(f(x_i) - y_i)K_{x_i}. \tag{2}$$

给定步长 $\beta > 0$，式（1）的 KGD 可迭代地表述为

$$f_{t+1,\beta,D} = f_{t,\beta,D} - \frac{\beta}{|D|}\sum_{i=1}^{|D|}(f_{t,\beta,D}(x_i) - y_i)K_{x_i}, \tag{3}$$

其中 $f_{0,\beta,D} = 0$ 和 $K_{x_i}(\cdot) = K(x_i, \cdot)$。显然，$f_{t,\beta,D} \in \mathcal{H}_K$ 具有特殊形式 $f_{t,\beta,D} = \sum_{i=1}^{|D|} c_i^t K_{x_i}$。令 $\mathbf{c}_t = (c_1^t, \ldots, c_{|D|}^t)^\top$、$y_D = (y_1, \ldots, y_{|D|})^\top$ 和 $\mathbb{K} = (K(x_i, x_{i'}))_{x_i, x_{i'}\in D}$ 为核矩阵。由式（3）可得

$$\mathbf{c}_{t+1} = \mathbf{c}_t - \frac{\beta}{|D|}(\mathbb{K}\mathbf{c}_t - y_D) \tag{4}$$

其中 $\mathbf{c}_0 = (0, \ldots, 0)^\top$。因此，KGD 在每次迭代中需要进行 $\mathcal{O}(|D|^2)$ 次浮点计算。Lin 和 Zhou（2018a）以及 Yao 等人（2007）表明，只要适当调整迭代次数，KGD 就是一个能够达到最优泛化误差界的最优学习器。它已广泛应用于回归（Yao 等人，2007）、分类（Wei 等，2019）、最小误差熵原理分析（Hu 等，2020）和检验（Hagrass 等，2024）。

如图1所示，迭代次数 t至关重要，因为通过设置最优的 t 来平衡偏差与方差，KGD 才能实现最佳泛化性能。我们的目的是提出一种新颖的策略，自适应地选择 t，以逼近最优的 t_*。我们的方法基于一种新颖的偏差-方差分析原理，称为反向选择原理（BSP）。该原理由核矩阵的所谓经验有效维数（Zhang，2002）决定，其定义为

$$\mathcal{N}_D(\lambda) := \mathrm{Tr}[(\lambda|D|I + \mathbb{K})^{-1}\mathbb{K}], \quad \forall\lambda > 0, \tag{5}$$

其中，Tr 表示矩阵（或迹类算子）的迹。记

$$\mathcal{W}_{D,t} := \left(\frac{\sqrt{t}}{|D|} + \frac{\sqrt{\max\{\mathcal{N}_D(t^{-1}), 1\}}(1 + \sqrt{t/|D|})}{\sqrt{|D|}}\right) \tag{6}$$

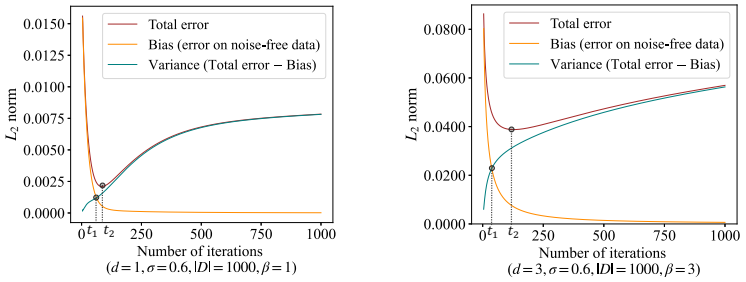

图1： L_2 范数下迭代次数对控制 KGD 的偏差、方差和总误差的作用。训练样本 $\{x_i\}_{i=1}^{1000}$是根据均匀分布独立抽取的，该分布定义在具有 $d=1,3$ 的（超）立方体 $[0,1]^d$ 上。相应的输出由模型 $y_i=g_j(x_i)+\varepsilon_i, j=1,2$ 生成，其中 g_1 和 g_2 分别由式（43）和式（44）定义，ε_i 是独立的高斯噪声 $\mathcal{N}(0,\sigma^2)$，且 $\sigma=0.6$。偏差与方差由下文的式（23）定义。

and

$$\mathcal{U}_{D,t,\delta}:=\left(\frac{\log\left(1+8\log\frac{64}{\delta}\frac{\sqrt{t}}{\sqrt{|D|}}\max\{1,\mathcal{N}_D(t^{-1})\}\right)}{t^{-1}|D|}+\sqrt{\frac{\log\left(1+8\log\frac{64}{\delta}\frac{\sqrt{t}}{\sqrt{|D|}}\max\{1,\mathcal{N}_D(t^{-1})\}\right)}{t^{-1}|D|}}\right) \tag{7}$$

对任意给定的 $\delta\in(0,1)$。进一步定义

$$T:=T_{\delta,D}:=\max_{1\le t\le|D|}\{C_1^*\mathcal{U}_{D,t,\delta}\le 1/2\} \tag{8}$$

其中 $C_1^*:=\max\{(\kappa^2+1)/3,2\sqrt{\kappa^2+1}\}$。由于 $\mathcal{U}_{D,t,\delta}$ 随 t增大，所以上述T 定义良好。接下来，我们为 KGD 定义 BSP，如下所示。

定义1（反向选择原理（BSP））给定置信水平 $\delta\in(0,1)$ 和步长 β，定义$\hat{}t:=\hat{}t_{D,\beta}\in[1,T]$ 为满足以下条件的最大整数

$$t\|f_{t+1,\beta,D}-f_{t,\beta,D}\|_D+t^{1/2}\|f_{t+1,\beta,D}-f_{t,\beta,D}\|_K\ge\tilde{C}\mathcal{W}_{D,t}\log^2\frac{16}{\delta}, \tag{9}$$

$\tilde{}$其中 C 是仅取决于 κ 和噪声方差的常数。若不存在任何$t\in[1,T]$满足（9），则令$\hat{}t=T$。

$\tilde{}$图2说明了 BSP 的可行性，即给定的常数 C 遵循特定的$\tilde{}t_{\tilde{C}}$。然而，过小或过大的 C 会直接导致 $t_{\tilde{C}}=T$。不难验证，所提出的 BSP 是 Lepskii 原理的一种特殊形式，唯一的区别在于，它移除了经典原理中的逐项比较（Blanchard 等人，2019；Lu 等，

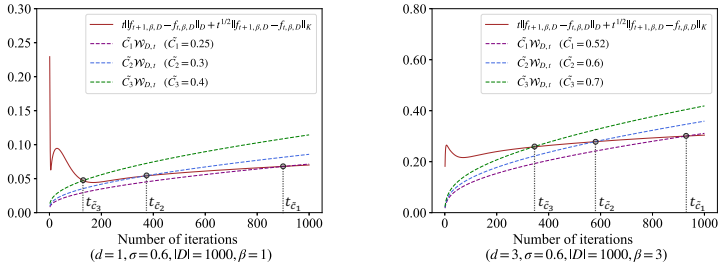

图2： BSP 下由不同常数 $\tilde{C}$ 的取值决定的步骤 $t_{\tilde{C}}$。实验设置与图1相同。

2020）。还应强调，BSP 并非像（Raskutti 等，2014）中的方案那样采用提前停止，因为 BSP 要求首先从 $t=1$ 到 $t=T$ 运行 KGD，然后找到满足（9）的最大 t。根据（4），存在 $\mathbf{c}^{t+1}=(c_1^{t+1},\ \ldots\ldots,\ c_{|D|}^{t+1})^\top$和 $\mathbf{c}^t=(c_1^t,\ \ldots\ldots,\ c_{|D|}^t)^\top$，使得

$$f_{t+1,\beta,D}-f_{t,\beta,D}=\sum_{i=1}^{|D|}(c_i^{t+1}-c_i^t)K_{x_i}.$$

于是，

$$\|f_{t+1,\beta,D}-f_{t,\beta,D}\|_K^2=\sum_{i=1}^{|D|}\sum_{j=1}^{|D|}(c_i^{t+1}-c_i^t)(c_j^{t+1}-c_j^t)K(x_i,x_j)=(\mathbf{c}^{t+1}-\mathbf{c}^t)^\top\mathbb{K}(\mathbf{c}^{t+1}-\mathbf{c}^t)$$

and

$$\|f_{t+1,\beta,D}-f_{t,\beta,D}\|_D^2=\frac{1}{|D|}\sum_{i=1}^{|D|}(f_{t+1,\beta,D}(x_i)-f_{t,\beta,D}(x_i))^2=\frac{1}{|D|}(\mathbf{c}^{t+1}-\mathbf{c}^t)^\top\mathbb{K}^2(\mathbf{c}^{t+1}-\mathbf{c}^t),$$

结合 (5) 和 (6)，定义1中的 BSP 即可实现 e.

由于 (6)，参数选择策略需要针对不同的 t 计算有效维数$\mathcal{N}_D(t^{-1})=\mathrm{Tr}[(t^{-1}|D|I+\mathbb{K})^{-1}\mathbb{K}]$。注意到矩阵 $\mathbb{K}$ 和 $(t^{-1}|D|I+\mathbb{K})^{-1}$ 的特征向量相同，如果已获得 $\mathbb{K}$ 的特征值（按降序记为 $\{\sigma_1^{\mathbf{x}},\ldots,\sigma_{|D|}^{\mathbf{x}}\}$），则可以通过 $\mathcal{N}_D(t^{-1})=\sum_{i=1}^{|D|}\frac{\sigma_i^{\mathbf{x}}}{\sigma_i^{\mathbf{x}}+t^{-1}|D|}$ 得到任意 t 对应的经验有效维数。因此，BSP 需要计算 $\mathbb{K}$ 的特征值，并需要 $\mathcal{O}(|D|^3)$；这与 (Raskutti 等，2014) 提出的早停规则类似，尽管 (Raskutti 等，2014) 中的早停会在 $t=|D|$ 之前强制终止迭代。

在实现 BSP 时，我们需要确定常数 κ, M 和置信水平δ。众所周知，估计噪声方差 M 并非易事，但我们可以采用（Raskutti 等，2014）中的方法对其进行近似计算，从而使 BSP˜在数值上可行。主要问题在于，（9）中推导出的常数 C 通常并非最优，导致 BSP 远非最优。幸运的是，我们发现该常数，如

算法 1 采用 HSS 的 KGD

输入： $D = \{(x_i, y_i)\}_{i=1}^m$，常数集合 $\mathcal{C}_U := \{\hat{C}_j\}_{j=1}^U$，其中 $U \in \mathbb{N}$，步长 $0 < \beta \leq \frac{1}{\kappa}$ 和核 K。

数据划分：对于 $L \leq |D|$，根据均匀分布从 D 中随机选取 L 个样本。将选取的样本划分为训练集 $D_{tr,\ L}$ 和验证集 $D_{val,\ L}$。

计算经验有效维数：计算两个核矩阵 $\mathbb{K} = (K(x_i,\ x_{i'}))_{x_i,\ x_{i'} \in D}$ 和 $\mathbb{K}_L = (K(x_i,\ x_{i'}))_{x_i,\ x'_i \in D_{tr,\ L}}$ 的特征值，并分别记为 $\{\sigma_i^{\mathbf{x}}\}_{i=1}^{|D|}$ 和 $\{\sigma_{i,L}^{\mathbf{x}}\}_{i=1}^{|D|}$。根据公式 (5) 和 (6)，计算 $\mathcal{W}_{D,\ t}$ 和 $\mathcal{W}_{D_{tr,\ L},\ t}$，其中 $t = 1,\ \ldots,\ |D|$。

计算突变停止值 T**：** 给定置信水平 $\delta \in (0,\ 1)$，根据公式 (7) 计算 $\mathcal{U}_{D,\ t,\ \delta}$，然后根据公式 (8) 确定 T。

KGD：设置 $f_{0,\ \beta,\ D} = 0$ 和 $f_{0,\ \beta,\ D_{tr,\ L}} = 0$。根据公式 (3) 得到两组 KGD估计量序列 $\{f_{t+1,\beta,D}\}_{t=0}^{|D|-1}$ 和 $\{f_{t+1,\beta,D_{tr,L}}\}_{t=0}^{|D|-1}$。

常数选择：定义 $\hat{t}_j \in [1, T]$ 为满足条件的最大整数

$$t\|f_{t+1,\beta,D_{tr,L}} - f_{t,\beta,D_{tr,L}}\|_{D_{tr,L}} + t^{1/2}\|f_{t+1,\beta,D_{tr,L}} - f_{t,\beta,D_{tr,L}}\|_K \geq \hat{C}_j \mathcal{W}_{D_{tr,L},t}. \quad (10)$$

如果 (10) 对所有 $t \in [1, T]$都不成立，则令 $\hat{t}_j = T$。定义

$$j^* := \arg\min_{j=1,\ldots,U} \frac{1}{|D_{val,L}|} \sum_{(x,y)\in D_{val,L}} (f_{\hat{t}_j,\beta,D_{tr,L}}(x) - y)^2. \quad (11)$$

最终参数的 BSP：定义 $\hat{t}^* \in [1, |D|]$ 为满足条件的最大整数

$$t\|f_{t+1,\beta,D} - f_{t,\beta,D}\|_D + t^{1/2}\|f_{t+1,\beta,D} - f_{t,\beta,D}\|_K \geq \hat{C}_{j^*} \mathcal{W}_{D,t}. \quad (12)$$

输出：$f_{\hat{t}^*,\ \beta,\ D}$。

以及置信水平 δ，该水平可在学习过程开始前指定，以反映用户的置信度；二者均与数据规模无关。因此，我们可以使用留出法或交叉验证方法，通过在一个小数据集上运行带有 BSP 的 KGD 来选择它，例如从 D中抽取 $|D|/10$ 个样本。通过这种方式，我们实际上结合 BSP 与交叉验证，获得了一种自适应参数选择策略。我们将这一新颖策略称为混合选择策略（HSS），如算法 1 所示。

接下来，我们对算法 1 作如下几点说明。

◇ **候选常数**：HSS 应在训练之前量化集合 $\mathcal{C}_U$。回顾 (9)，我们注意到 $\mathcal{C}_U$ 具有上界 $4(1+\beta)\tilde{C}\log^4\frac{16}{\delta}$。一个有利的设定是取 $\mathcal{C}_U = \{4(1+\beta)\tilde{C}\log^4\frac{16}{\delta} q^k :\ k = 0, 1, \ldots, \}$，其中 $0 < q < 1$ 为某个值。然而问题在于，$4(1+\beta)\tilde{C}\log^4\frac{16}{\delta}$ 本身并不容易计算，因为它需要获得数据的噪声。在数值上，我们可以设定

$$\mathcal{C}_U := \left\{\tilde{C}_0 q^k : k = 0, 1, \ldots, U\right\} \quad (13)$$

$\tilde{}$ $\tilde{}$取一个足够大的 C_0，即例如 $C_0 = 100$、$q = 9/10$ 和 $U = 20$。

◇ **计算简化**：由于（9）中的常数$4(1+\beta)\tilde{C}\log^4\frac{16}{\delta}$与数据无关，且参数选择策略中对 $L \leq |D|$ 个样本的选择仅用于选取一个合适的常数以输入BSP，因此数据划分步骤中的子抽样方案在理论上并不影响最终常数的选择。在计算经验有效

维数步骤中，不必计算全部$\mathcal{W}_{D,\ t}$和$\mathcal{W}_{D_{tr},\ L,\ t}$。事实上，我们可以从 $|D|$ 向后选取t 直至 1，并在（10）和（12）得到满足时停止搜索。在此情形下，我们只需对不小于由（10）和（12）所确定数值的 t 计算$\mathcal{W}_{D,\ t}$和$\mathcal{W}_{D_{tr},\ L,\ t}$。注意到 $\mathcal{N}_D(t^{-1}) = \sum_{i=1}^{|D|} \frac{\sigma_i^{\mathbf{x}}}{\sigma_i^{\mathbf{x}}+t^{-1}|D|}$，对所有 t 计算$\mathcal{W}_{D,\ t}$和$\mathcal{W}_{D_{tr},\ L,\ t}$并不会带来太多额外计算，因此我们建议在算法中将它们全部计算出来。一旦常数 C_j^* 确定，我们便使用全部样本导出最终参数$\hat{}t^*$ 以输入KGD。若设定 $L \leq |D|/U^{1/3}$，则容易验证，采用HSS实现KGD时总共有 $\mathcal{O}(|D|^3)$次浮点计算。

◇ **提前停止与 BSP**：不同于 Raskutti 等，2014 为 KGD 配备提前停止方案，我们在算法 1 中提出的策略要求对所有 $t \in [1, T]$ 运行 KGD。不过需要强调的是，所有现有的提前停止参数选择策略（Raskutti 等，2014；Wei 等，2019）都要求计算 $\sigma_i^{\mathbf{x}}$ 的特征值，计算范围为 $i = 1, \ldots\ldots, |D|$，其计算复杂度为 $\mathcal{O}(|D|^3)$。在这种情况下，是否采用提前停止参数选择策略，本质上并不会影响计算复杂度。与 Raskutti 等，2014 和 Wei 等，2019 的策略相比，我们提出的 HSS 策略能够更准确地反映数据的正则性。

◇ **Lepskii 原理与偏差原则**：Lepskii 原理和偏差原则是用于为核方法设计参数选择策略的两种常用原则（De Vito 等人，2010；Lu 等，2020；Blanchard 等人，2019；Celisse 和 Wahl，2021；Lin，2024）。前者通过限制两个连续估计之间的差异来选择参数，而后者则致力于用一些可计算的量来量化拟合误差。对于一般的核方法，它们本质上是不同的方法（Blanchard 等人，2019；Celisse 和 Wahl，2021）。然而，由于 KGD 的特殊谱性质（Lin 和 Zhou，2018a），即 $f_{t+1,\ \beta,\ D} - f_{t,\ \beta,\ D} = \beta\left(L_{K,\ D} f_{t,\ \beta,\ D} - S_D^\top y_D\right)$，Lepskii 原理和偏差原则对于 KGD 是相同的。也就是说，HSS 背后的理念既是 Lepskii 原理，也是偏差原则。

2.2 相关研究

在本部分中，我们将所提出的自适应停止规则与一些相关工作进行比较。KGD 最初由 Yao 等人（2007；Gerfo 等人，2008）作为一类特殊的基于核的谱算法进行研究，旨在克服核岭回归的饱和现象（Caponnetto 和 De Vito，2007）。在对数据和核作出不同假设的条件下，Yao 等人（2007）、Raskutti 等（2014）以及 Lin 和 Zhou（2018a）建立了 KGD 的最优泛化误差界。然而，所有这些最优泛化误差界都是在迭代次数选择恰当这一条件下建立的。尽管已有若干停止规则被提出，用于在逆问题框架下确定迭代次数，包括偏差原则（Engl 等人，1996，第 6 章）和平衡原理（Lazarov 等人，2007），但对于非参数回归领域而言，为 KGD 设计停止规则并非易事。接下来，我们将介绍现有的 KGD 迭代次数选择策略。

- 基线（BS）：迭代次数由以下方式确定

$$t_{BS} := \arg\min_{t\in[0,|D|]} \|f_{t,\beta,D} - f_\rho\|_D^2. \tag{14}$$

式 (14) 定义的基线是在已知目标函数f_ρ 精确知识的假设下，用于确定 KGD 中 t 的理论最优参数选择策略。该策略具有良好的数值性能，但需要获取 f_ρ，这在实际中难以实现。我们引入这一方法仅是为了提供一个参考，以判断其他可实施参数选择策略的质量。

- 留出法（HO）：迭代次数由以下公式确定

$$t_{HO} := \arg\min_{t\in[0,|D|]} \mathrm{M}_{HO}(f_{t,\beta,D_{tr}}). \tag{15}$$

具体而言，留出法（或交叉验证 CV）（Caponnetto 和 Yao，2010年；Lin 和 Zhou，2018b；Zhang，2003年）是一种在数值上可行的方法，用于从$[0,|D|]$ 中选择 t ，以获得 KGD 的最优泛化误差界。该方法需要将样本集 D 划分为训练集和验证集，分别记为 D_{tr} 和 D_{val}，其中 $D = D_{tr} \cup D_{val}$、 $D_{tr} \cap D_{val} = \emptyset$ 和 D_{tr} 包含全部样本数据的一半。随后，它通过 $\mathrm{M}_{HO}(f_{t,\beta,D_{tr}})$ 在 D_{val} 上评估 KGD 的性能：$= \frac{1}{|D_{val}|}\sum_{(x_i,y_i)\in D_{val}}(y_i - f_{t,\beta,D_{tr}}(x_i))^2$ ，并选择 t 以最小化验证误差。利用与（Lin 和 Zhou，2018b）相同的方法，可以很容易地推导出 HO 对 KGD 的最优性。然而，问题在于，样本划分会降低 KGD 本身的数值性能。

- 平衡原理（BP）：给定置信水平 δ，迭代次数由以下公式确定

$$t_{BP} := \min\{t \in [0,|D|] : \|f_{t',\beta,D} - f_{t,\beta,D}\|_D \le C_{BP}\mathcal{W}_{D,t'}\log^4\frac{16}{\delta},\ t' = t+1,\dots,|D|\Big\}, \tag{16}$$

其中，$\mathcal{W}_{D,t}$ 由（6）定义，且 $C_{BP} > 0$ 是一个与 $|D|$、 δ 或 t无关的常数。平衡原理最初由 De Vito 等人，2010 在学习理论框架下提出，用于为 KRR 选择正则化参数，随后在 Lu 等，2020 中得到发展，用于确定基于核的谱算法的参数。从（16）可以看出，平衡原理不需要拆分样本集。

- Lepskii 原理（LP）：给定 $q > 1$ 和 $0 < \delta < 1$，定义 $T_q := \{t_i = q^i/(\kappa^2) : i \in \mathbb{N}\}$,$L_{\delta,q} = 2\log\left(\frac{8}{\log\delta}\frac{|D|}{\log q}\right)$和 $T_{\mathbf{x}} := \left\{t \in T_q : t \le \max\left\{\frac{|D|}{100\kappa^2 L_\delta^{2q}},\frac{|D|}{3\kappa^2\mathcal{N}_D((t-1)+1)}\right\}\right\}$,迭代次数由以下公式确定

$$t_{LP} := \min\left\{t \in T_{\mathbf{x}} : \|(L_{K,D}+t'^{-1}I)^{1/2}(f_{t',\beta,D} - f_{t,\beta,D})\|_K \le C_{LP}\mathcal{W}^*_{D,t'},\ t' \ge t,\ t' \in T_{\mathbf{x}}\right\}. \tag{17}$$

其中 $\mathcal{W}^*_{D,t} = \frac{\sqrt{t}}{(\mathcal{N}_D(t^{-1})+1)\sqrt{|D|}}$. Lepskii 原理最初由 (Blanchard等人，2019) 提出，旨在提升监督学习中平衡原理的性能。(Blanchard等人，2019) 验证了 Lepskii 原理的预言机性质。具体而言，(Blanchard等人，2019) 表明，采用 Lepskii 原理的 KGD 能够实现近乎最优的泛化误差界。与反向传播类似，线性规划也需要针对给定常数 C_{LP}逐项进行比较。由于 C_{LP} 在实际中未知，因此通常需要大量计算。

• 早停规则（ESR）：定义核矩阵 $\mathbb{K}$ 的局部经验拉德马赫复杂度为 $\hat{\mathcal{R}}_{\mathbb{K}}(\epsilon) := \left[\frac{1}{|D|}\sum_i \min\{\sigma_{\mathbf{x}i}, \epsilon^2\}\right]^{1/2}$. 记 $\eta_t = t\beta$。设 t_{ESR} 为满足以下条件的第一个正数

$$\hat{\mathcal{R}}_{\mathbb{K}}(1/\sqrt{\eta_t}) > C_{ESR}(\tau\eta_t)^{-1}, \tag{18}$$

其中，$\tau > 0$ 是噪声的标准差，可通过（Hall 和 Marron，1990）中的经典方法估计；$C_{ESR} > 0$ 是一个调节参数，Raskutti 等人（2014）建议将其设为 $2e$ 。据我们所知，Raskutti 等人（2014）的研究首次为 KGD 设计了专门的早停规则，并给出了理论保证。与上述四种方法不同，ESR 无需在$[1, |D|]$上运行 KGD。然而，这种早停规则并不能自适应目标函数的正则性，因为它仅在假设目标函数属于 $\mathcal{H}_K$时才成立。

• 偏差原则（DP）：给定噪声水平 σ^2，偏差原则是一种经典且广泛用于逆问题参数选择的策略，由 Engl 等人（1996，第 4.3 节）定义：

$$t_{DP} := \inf_{t\in[1,|D|]}\{\|y_D - S_D f_{t,\beta,D}\|_{\ell^2} \le C_{DP}\sigma^2\},$$

其中，C_{DP} 是一个独立于 t和 $|D|$的常数。然而，对于监督学习而言，量化噪声水平 σ^2 并非易事，这限制了动态规划的应用。有趣的是，Celisse 和 Wahl（2021）通过引入噪声水平的计算和紧急停止规则，修改了偏差原则以适用于学习任务。Celisse 和 Wahl（2021）还在学习理论框架下进行了理论分析，以验证偏差原则的有效性。然而，问题在于，推导出的泛化误差界并非最优（Celisse 和 Wahl，2021，定理20）。

• 赤池信息准则（AIC）：令$W_{D,\,t}$ 由（6）定义，以模拟核学习框架中的自由度。Akaike 于1974年针对线性方法提出了 AIC，旨在引入额外的惩罚项，以平衡拟合精度和模型复杂度，其数学定义如下：

$$t_{AIC} := \inf_{t\in[1,|D|]} \|y_D - S_D f_{t,\beta,D}\|_{\ell^2} + C_{AIC}\mathcal{W}_{D,t}.$$

尽管上述定义中的惩罚项与（Demyanov 等人，2012年）中的略有不同，但其主要思想保持不变，即利用某些可实现的量来度量自由度，并将其作为惩罚项。此类赤池信息准则的一些统计性质已在（Claeskens 等人，2008年）中得到研究，但为相应算法推导可证明的泛化误差界仍然具有挑战性。在这种情况下，我们无法从理论上验证所定义的赤池信息准则的自适应性。

• 贝叶斯信息准则（BIC）：Schwarz（1978）最初针对线性方法提出了 BIC，并可将其扩展至 KGD，具体如下

$$t_{BIC} := \inf_{t\in[1,|D|]} \|y_D - S_D f_{t,\beta,D}\|_{\ell^2} + C_{BIC}\mathcal{W}_{D,t}\log|D|.$$

与赤池信息准则相比，贝叶斯信息准则的惩罚项中多了一个 $\log|D|$ 项。其统计性质与赤池信息准则类似，已在（Demyanov 等人，2012年）中讨论。从数学上很容易推导出 $t_{BIC} \leq t_{AIC}$。然而，与赤池信息准则类似，要通过推导相应 KGD 的最优泛化误差界来从理论上验证贝叶斯信息准则的自适应性也很困难。

3. 理论性质

In 本节中，我们对所提出的用于 KGD 的 HSS 进行理论验证 .

3.1 KGD 的偏差与方差分析

我们在非参数回归的框架下对 KGD 进行偏差-方差分析（Gy¨orfi 等，2002）。在该框架中，假设样本 $\{(x_i, y_i)\}_{i=1}^{|D|}$ 是从未知分布 $\rho := \rho_X \times \rho(y|x)$ 中独立同分布抽取的，其中 ρ_X 为边缘分布， $\rho(y|x)$ 为 x 上的条件分布。学习的目标是找到一个估计量来近似回归函数 $f_\rho = \int_{\mathcal{Y}} y d\rho(y|x)$，该函数使泛化误差 $\mathcal{E}(f) := \int (f(x)-y)^2 d\rho$ 最小。令 $L^2_{\rho_X}$表示由范数 $\|\cdot\|_\rho$赋范的 ρ-平方可积函数空间。容易验证

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2, \qquad f \in L^2_{\rho_X}. \tag{19}$$

KGD 的学习性能已在（Yao 等人，2007；Raskutti 等，2014；Lin 和 Zhou，2018a）中，在对噪声、回归函数的正则性以及核的容量作出假设的条件下得到了广泛研究。

假设1 （噪声假设：） 假设 $\int_{\mathcal{Y}} y^2 d\rho < \infty$ 且

$$\int_{\mathcal{Y}} \left(e^{\frac{|y-f_\rho(x)|}{M}} - \frac{|y-f_\rho(x)|}{M} - 1 \right) d\rho(y|x) \leq \frac{\gamma^2}{2M^2}, \forall x \in X, \tag{20}$$

其中 M 和 γ 为正常数。

条件（20）是关于噪声的标准 Bernstein 假设（Caponnetto and De Vito，2007；Blanchard 和 Kr¨amer，2016）；当噪声一致有界、服从高斯分布或次高斯分布时，该假设均可满足。为引入正则性假设，我们将以下正算子 $L_K : \mathcal{H}_K \to \mathcal{H}_K$ （或在无歧义时记作 $L^2_{\rho_X} \to L^2_{\rho_X}$ ）定义为 $L_K(f) := \int_X f(x) K_x d\rho_X$. 由于 K是 Mercer核，因此该算子的正性得以保证。

假设2（正则性假设） 存在一个 $r > 0$ 使得

$$f_\rho = L_K^r(h_\rho), \qquad h_\rho \in L^2_{\rho_X}, \quad r > 0, \tag{21}$$

其中， L_K^r 是 L_K的 r次幂。

假设2已在文献中得到广泛应用（Caponnetto and De Vito，2007；Yao 等人，2007；Blanchard 和 Kr¨amer，2016；Guo 等人，2017；Lin 等人，2017），指数 r量化了 f_ρ 的正则性。需要注意的是，(21) 在 $r = 1/2$ 条件下蕴含 $f_\rho \in \mathcal{H}_K$，反之亦然（ Caponnetto and De Vito，2007）。此外， r 越大，回归函数的正则性越好。

接下来，我们从偏差与方差的角度研究迭代在学习过程中的作用，以寻求为 KGD 选择最优 t 。令$f^{\diamond}_{t,\beta,D} := E[f_{t,\beta,D}|x]$ 为$f_{t,\beta,D}$的无噪声版本。由 (3) 可得

$$f^{\diamond}_{t+1,\beta,D} = f^{\diamond}_{t,\beta,D} - \frac{\beta}{|D|}\sum_{i=1}^{|D|}(f^{\diamond}_{t,\beta,D}(x_i) - f_\rho(x_i))K_{x_i}. \tag{22}$$

于是，由三角不等式可得

$$\|f_{t+1,\beta,D} - f_\rho\|_\rho \leq \overbrace{\|f^{\diamond}_{t+1,\beta,D} - f_\rho\|_\rho}^{Bias} + \overbrace{\|f^{\diamond}_{t+1,\beta,D} - f_{t+1,\beta,D}\|_\rho}^{Variance}. \tag{23}$$

以下引理可见于（Lu 等，2020；Lin 等人，2019），它建立了 $\|f\|_\rho$、 $\|f\|_D$ 和 $\|f\|_K$ 之间的关系。

引理 1 设 $f \in \mathcal{H}_K$。则

$$\|f\|_\rho \leq \mathcal{Q}_{D,t}(\|f\|_D + t^{-1/2}\|f\|_K), \tag{24}$$

其中

$$\mathcal{Q}_{D,t} := \|(L_{K,D} + t^{-1}I)^{-1/2}(L_K + t^{-1}I)^{1/2}\| \tag{25}$$

且 $\|A\|$ 表示算子 A 的谱范数。

根据上述引理，要研究 t 在界定$\|f_{t,\beta,D} - f_\rho\|_\rho$中的作用，只需分析其对$\|f_{t,\beta,D} - f_\rho\|_D + t^{-1/2}\|f_{t,\beta,D} - f_\rho\|_K$的影响。由(23)，偏差与方差可界定为

$$\mathcal{B}_{t,\beta,D} := \|f^{\diamond}_{t,\beta,D} - f_\rho\|_D + t^{-1/2}\|f^{\diamond}_{t,\beta,D} - f_\rho\|_K \tag{26}$$

and

$$\mathcal{V}_{t,\beta,D} := \|f^{\diamond}_{t,\beta,D} - f_{t,\beta,D}\|_D + t^{-1/2}\|f^{\diamond}_{t,\beta,D} - f_{t,\beta,D}\|_K. \tag{27}$$

以下命题通过(6)和有效维数 $\mathcal{N}(\lambda)$（Zhang，2002）说明 t 如何控制偏差与方差，其中有效维数定义为 $\mathcal{N}(\lambda)$: $=\mathrm{Tr}[(\lambda I + L_K)^{-1}L_K]$，其中$\lambda > 0$。

命题 2 设 $0 < \delta < 1$, $f_{t,\beta,D}$ 和$f^{\diamond}_{t,\beta,D}$ 由*(3)*和*(22)*定义，其中 $t \leq T$ 且$0 < \beta \leq \kappa^{-1}$。在假设*1*和假设*2*以及 $r \geq 1/2$成立的条件下，以 $1-\delta$的置信度，有

$$\mathcal{B}_{t,\beta,D} \leq \log^2\frac{16}{\delta}\begin{cases} 2^{r-\frac{1}{2}}\,t^{-r}, & \text{if } \frac{1}{2} \leq r \leq 1, \\ t^{-r} + 4\kappa^2 t^{-\frac{1}{2}}|D|^{-\min\{1/2,(r/2-1/4)\}}, & \text{if } r > 1 \end{cases} \tag{28}$$

and

$$\mathcal{V}_{t,\beta,D} \quad \leq \quad 2\sqrt{2}(\kappa M + \gamma)(1+\beta)\mathcal{A}_{D,\lambda}\log\frac{8}{\delta} \leq 2(1+\beta)\mathcal{W}_{D,t}\log^2\frac{16}{\delta}, \tag{29}$$

其中

$$\mathcal{A}_{D,t} := \frac{1}{\sqrt{|D|}}\left(\frac{\sqrt{t}}{\sqrt{|D|}} + \sqrt{\mathcal{N}(t^{-1})}\right). \tag{30}$$

命题 2 展示了偏差与方差随迭代次数 t变化的趋势。由 (6) 和 (29) 可知，方差的上界随迭代次数增加而增大。相反，(28) 表明偏差的上界相对于 t单调递减。因此，存在一个使 KGD 泛化误差最小的最优 t_*。设 $\{(\sigma_\ell, \varphi_\ell)\}_\ell$ 是 L_K 在 $\mathcal{H}_K$上的一组归一化特征对，其中 $\{\varphi_\ell\}_\ell$ 是 $\mathcal{H}_K$ 的正交归一基，且 $\{\sigma_\ell\}_\ell$ 按非递增顺序排列。基于命题 2，我们可以推导出以下推论，以展示在 σ_ℓ具有不同衰减率的情况下，为参数 t选择合适的调优值时 KGD 的泛化误差界。

推论 3 令 $0 < \delta < 1$, $f_{t,\beta,D}$ 由 *(3)* 定义，其中 $t \le T$ 和 $0 < \beta \le \kappa^{-1}$。在假设*1*和假设*2*以及 $r \ge 1/2$成立的条件下，若

$$t \sim \begin{cases} (|D|/L)^{\frac{1}{2r}}, & \text{if } \sigma_\ell = 0, \ell \ge L+1, \\ |D|^{\frac{1}{2r+s}}, & \text{if } \sigma_\ell \le c_0 \ell^{-1/s}, \\ (|D|/\sqrt{\log|D|})^{\frac{1}{2r}}, & \text{if } \sigma_\ell \le c_1 e^{-c_2 \ell^2}, \end{cases} \tag{31}$$

则以 $1-\delta$ 的置信度，有

$$\begin{aligned} & \max\{\|f_{D,t} - f_\rho\|_\rho, \|f_{D,t} - f_\rho\|_D, t^{1/2}\|f_{D,t} - f_\rho\|_K\} \\ \le\ & \tilde{C} \log^2 \frac{16}{\delta} \begin{cases} (|D|/L)^{-1/2}, & \text{if } \sigma_\ell = 0, \ell \ge L+1, \\ |D|^{-r/(2r+s)} & \text{if } \sigma_\ell \le c_0 \ell^{-1/s}, \\ (|D|/\sqrt{\log|D|})^{-1/2}, & \text{if } \sigma_\ell \le c_1 e^{-c_2 \ell^2}, \end{cases} \end{aligned} \tag{32}$$

~其中 C 是一个与 δ、$|D|$或 t无关的常数。

Caponnetto and De Vito (2007) 和 Raskutti 等 (2012) 指出，已建立的 KGD 泛化误差界在极小极大意义下是最优的，并且 KGD 的学习性能与步长 β 无关，前提是 $0 < \beta \le \kappa^{-1}$。上述理论验证表明 KGD 的优越性，但如何选择合适的参数 t 以优化其学习能力仍未解决，因为满足 (31) 的理论最优参数 t_* 在实践中无法直接获得。从 (31) 可知，对于有限秩和指数衰减的 L_K ，理论最优的 t 与 s 无关。因此，若核和 ρ_X 已指定，使得 L_K 为有限秩或指数衰减，则无需使停止规则自适应于核。回顾 (32)，有限秩和指数衰减 L_K 在 $\|\cdot\|_\rho$ 范数下的泛化误差界与 r 无关。然而，对于 $\mathcal{H}_K$ 范数，指标$_1$$_{2r}$至关重要。例如，若 L_K 是有限秩的 L，则由 (32) 可得$\|f_{D,\ t} - f_\rho\|_K \le \tilde{C}\left(\frac{L}{|D|}\right)^{\ 2}$ $\log^2 \frac{16}{\delta}$， 而 Raskutti 等 (2014) 和 Zhang 等人 (2015) 并未考虑这一点，因为他们只分析了 $r = 1/2$ 。

由于命题 2 中偏差与方差的计算需要用到实际中未知的回归函数，我们转而采用一种替代方法，利用 KGD 的迭代特性。我们的基本思路是在以下命题中控制连续迭代的误差增量。

命题 4 设 $0 < \delta < 1$ 和 $f_{t,\ \beta,\ D}$ 由 *(3)* 定义，其中 $t \in \mathbb{N}$ 和 $0 < \beta \le \kappa^{-1}$。在假设*1*和假设*2*成立且满足 $r \ge 1/2$的条件下，以置信度 $1-\delta$，有

$$\begin{aligned}&\|f_{t+1,\beta,D} - f_{t,\beta,D}\|_D + t^{-1/2}\|f_{t+1,\beta,D} - f_{t,\beta,D}\|_K \\ &\le 2\sqrt{2}\log^2\frac{16}{\delta}\,(\beta+1)\,t^{-1}\,\mathcal{W}_{D,t} + C_1'\log^2\frac{16}{\delta}\begin{cases}2^{r-1/2}\,t^{-r-1}, & \text{if } \frac{1}{2} \le r \le 1,\\ t^{-r-1}+4\kappa^2 t^{-3/2}|D|^{-\min\{1/2,\,r/2-1/4\}}, & \text{if } r > 1,\end{cases}\end{aligned} \tag{33}$$

其中， C_1' 是一个独立于 $|D|$、 t或 δ的常数。

比较命题

4 2, $t\|f_{t+1\,\beta\,D} - f_{t\,\beta\,D}\|_D +$ 与命题可知，就而言，,,,,,,,与,,,的表现相同。唯一的区别在于，$t^{1/2}\|f_{t+1\,\beta\,D} - f_{t\,\beta\,D}\|_K$ $\mathcal{V}_{t\,\beta\,D} + \mathcal{B}_{t\,\beta\,D}$ t $\|f_{t+1,\beta,D} - f_{t,\beta,D}\|_D$ 和$\|f_{t+1,\beta,D} - f_{t,\beta,D}\|_K$ 在数值上可实现，而$\|f_{t,\beta,D} - f_\rho\|_D$ 或$\|f_{t,\beta,D} - f_\rho\|_K$ 则无法计算。因此，可以量化连续迭代的增量，以制定类似于(9)的参数选择策略，从而获得泛化误差$\|f_{t,\beta,}$D$\|_K$的理想理论结果。

3.2 半自适应停止规则与迭代次数上界

为了研究 HSS 的性质，尤其是 $\hat{}t$ 的上界，我们基于命题 4 为 KGD 引入一种半自适应停止规则。定义 $t^* := t^*_{D,\ \beta} \in [1, T]$ 为满足以下条件的最大整数

$$\mathcal{W}_{D,t} \le C_2'\begin{cases}2^{r-1/2}t^{-r}, & \text{if } \frac{1}{2} \le r \le \frac{3}{2},\\ t^{-r} + 4\kappa^2 t^{-1/2}|D|^{-1/2}, & \text{if } r > \frac{3}{2},\end{cases} \tag{34}$$

其中 $C_2' := \max\{\tilde{C}$,(2(β + 1) $\tilde{C}$)$^{-1}C_1'\}$。容易看出，(34) 的左侧随 t 增大而增大，而右侧随 t 增大而减小。因此， t^* 的定义表明

$$\mathcal{W}_{D,t} \le C_2'\begin{cases}2^{r-1/2}t^{-r}, & \text{if } \frac{1}{2} \le r \le \frac{3}{2},\\ t^{-r} + 4\kappa^2 t^{-1/2}|D|^{-1/2}, & \text{if } r > \frac{3}{2},\end{cases} \forall t \le t^*, \tag{35}$$

and

$$\mathcal{W}_{D,t} \ge C_2'\begin{cases}2^{r-1/2}t^{-r}, & \text{if } \frac{1}{2} \le r \le \frac{3}{2},\\ t^{-r} + 4\kappa^2 t^{-1/2}|D|^{-1/2}, & \text{if } r > \frac{3}{2},\end{cases} \quad \forall t > t^*. \tag{36}$$

下面的命题给出了 $\hat{}t$ 与 t^* 之间的关系，并为 $\hat{}t$ 提供了一个上界。

命题5 设 $0 < \delta < 1$。对于

$$t_0 = \tilde{c}\log^{a(\sigma)}\frac{16}{\delta}\begin{cases}(|D|/L)^{\frac{1}{2r}}, & \text{if } \sigma_\ell = 0, \ell \ge L+1,\\ |D|^{\frac{1}{2r+s}}, & \text{if } \sigma_\ell \le c_0\ell^{-1/s},\\ (|D|/\sqrt{\log|D|})^{\frac{1}{2r}}, & \text{if } \sigma_\ell \le c_1 e^{-c_2\ell^2},\end{cases} \tag{37}$$

其中某个 $\tilde{c}$仅取决于 κ、 M、 δ、 γ 和 $a(\sigma) := -\frac{s}{2r+s}$当 $\sigma_\ell \le c_0\ell^{-1/s}$且$a(\sigma) = 0$否则，若假设 *1* 和 *2* 在 $r \ge 1/2$ 下成立，则以 $1-\delta$ 的置信度，有

$$\max\{\hat{t}, t_0\} \le t^* \le T \tag{38}$$

and

$$\begin{aligned}&\max\{\|f_{D,t^*}-f_\rho\|_\rho,\|f_{D,t^*}-f_\rho\|_D,(t^*)^{1/2}\|f_{D,t^*}-f_\rho\|_K\}\\ \leq\ & \tilde{C}'\log^2\frac{16}{\delta}\begin{cases}(|D|/L)^{-1/2}, & \text{if } \sigma_\ell=0,\ell\geq L+1,\\ |D|^{-r/(2r+s)}, & \text{if } \sigma_\ell\leq c_0\ell^{-1/s},\\ (|D|/\sqrt{\log|D|})^{-1/2}, & \text{if } \sigma_\ell\leq c_1e^{-c_2\ell^2},\end{cases}\end{aligned}\tag{39}$$

~其中 C' 是一个与 $|D|$、t 或 δ 无关的常数。

由于 $t_0\leq t^*$，由（39）可得

$$\begin{aligned}\|f_{D,t^*}-f_\rho\|_K &\leq t_0^{-1/2}\tilde{C}'\log^2\frac{16}{\delta}\begin{cases}(|D|/L)^{-1/2}, & \text{if } \sigma_\ell=0,\ell\geq L+1,\\ |D|^{-r/(2r+s)}, & \text{if } \sigma_\ell\leq c_0\ell^{-1/s},\\ (|D|/\sqrt{\log|D|})^{-1/2}, & \text{if } \sigma_\ell\leq c_1e^{-c_2\ell^2}\end{cases}\\ &\leq \tilde{C}_1'\log^2\frac{16}{\delta}\begin{cases}(|D|/L)^{-\frac{(1-1/2r)}{2}}, & \text{if } \sigma_\ell=0,\ell\geq L+1,\\ |D|^{-\frac{r-1/2}{2r+s}}, & \text{if } \sigma_\ell\leq c_0\ell^{-1/s},\\ \left(\frac{|D|}{\sqrt{\log|D|}}\right)^{-\frac{(1-1/2r)}{2}}, & \text{if } \sigma_\ell\leq c_1e^{-c_2\ell^2},\end{cases}\end{aligned}$$

~这与推论 3 中推导出的界一致，其中 C_1' 是一个与 δ，$|D|$，t无关的常数。命题5为采用所提出的半自适应停止规则的 KGD 推导了最优泛化误差界。此外，它还给出了 $\hat{}t$ 的一个上界，该上界对于验证定义1中的 BSP 规则的最优性至关重要。

3.3 采用 HSS 的 KGD 的最优泛化误差界

在本小节中，我们对所提出的参数选择策略进行理论验证。首先，我们推导由定义1给出的采用 BSP 的 KGD 的最优泛化误差界。

定理 6 设 $0<\delta<1$ 和 $0<\beta\leq\kappa^{-1}$。在假设*1*和假设*2*成立且$r\geq 1/2$的条件下，若$\hat{}t$ 由定义*1*定义，且 $\tilde{C}\geq 32\sqrt{2}(1+\beta)(\kappa M+\gamma)$，则以$1-\delta$的置信度，有

$$\max\left\{\|f_{\hat{t},\beta,D}-f_\rho\|_\rho,\|f_{\hat{t},\beta,D}-f_\rho\|_D\right\}\leq C\log^4\frac{16}{\delta}\begin{cases}\frac{\sqrt{L}\log|D|}{\sqrt{|D|}}, & \text{if } \sigma_\ell=0,\ell\geq L+1,\\ |D|^{-r/(2r+s)}, & \text{if } \sigma_\ell\leq c_0\ell^{-1/s},\\ \frac{\log|D|}{\sqrt{|D|}}, & \text{if } \sigma_\ell\leq c_1e^{-c_2\ell^2},\end{cases}\tag{40}$$

and

$$\|f_{\hat{t},\beta,D}-f_\rho\|_K\leq C\log^4\frac{16}{\delta}\begin{cases}\left(\frac{\sqrt{L}\log|D|}{\sqrt{|D|}}\right)^{1-1/2r}, & \text{if } \sigma_\ell=0,\ell\geq L+1,\\ |D|^{-\frac{r-1/2}{2r+s}}, & \text{if } \sigma_\ell\leq c_0\ell^{-1/s},\\ \left(\frac{\log|D|}{\sqrt{|D|}}\right)^{(1-1/2r)}, & \text{if } \sigma_\ell\leq c_1e^{-c_2\ell^2},\end{cases}\tag{41}$$

其中 C 是一个与 $|D|$ 或 δ无关的常数。

表 1：不同参数选择方法的比较

KGD 参数选择方法的比较					
	方法	泛化误差界	适用于核	Ada. 到函数	Ada. 到范数 $\|\cdot\|_\rho$ 和 $\|\cdot\|_K$
信息熵方法	赤池信息准则（*Demyanov* 等人，*2012*年）	N/A	✓	×	×
	贝叶斯信息准则（*Demyanov* 等人，*2012*年）	N/A	✓	×	×
分割方法	高阶（*Caponnetto* 和 *Yao*，*2010*年）	最优	✓	✓	×
	交叉验证（*Gy¨orfi* 等，*2002*）	最优	✓	✓	×
	留一法（*Zhang*，*2003*年）	次优	✓	✓	×
偏差-方差分析方法	反向传播（*Lu* 等，*2020*年）	近乎最优	✓	✓	×
	线性规划（*Blanchard* 等人，*2019*年）	近乎最优	✓	✓	✓
	ESR（*Raskutti* 等，*2014*年）	最优	✓	×	×
	动态规划（*Celisse* 和 *Wahl*，*2021*年）	次优	✓	✓	×
Ours	*BSP*	最优	✓	✓	✓

Note：“Ada.”表示自适应。“N/A”表示未找到与 KGD 相关的文献 .

定理 6 表明，对于多项式衰减的 L_K，采用 BSP 的 KGD 能够达到推论 3 中 KGD 的最优泛化误差界。表 1 从理论性能的角度比较了 KGD 的不同参数选择策略。我们有以下观察。所有方法都能适应核，但只有 HO（Caponnetto 和 Yao，2010）、交叉验证（Gy¨orfi 等人，2002）、ESR（Raskutti 等，2014）和 BSP 能达到最优泛化误差界。值得注意的是，分割方法无法适应误差度量。KGD 的 ESR 无法适应 f_ρ 的正则性，因为它仅在 $r = 1/2$时成立。事实上，它的表现与式（34）定义的半自适应停止规则类似。对于 De Vito 等人（2010；Lu 等，2020）提出的反向传播以及 Celisse 和 Wahl（2021）提出的动态规划，两者都无法适应误差度量，这表明不同范数需要采用不同的参数选择策略。只有 Blanchard 等人（2019）提出的线性规划和 BSP 能够同时适应正则性指数 r、容量指数 s 以及误差度量。然而，线性规划需要在数值层面反复进行比较，并会在泛化误差界中引入额外的对数项，因此推论 3 与 Blanchard 等人（2019，定理 1）之间存在差距。在近期的一项工作中（Lin，2024），研究者为核岭回归设计了一种新颖的 Lepskii 型原理，并推导出了最优泛化误差界。然而，问题在于，由于核岭回归的饱和现象（Bauer 等人，2007；Gerfo 等人，2008），其泛化误差界会在 $r \leq 1$处饱和，这意味着更大的 r 不再带来更优的泛化误差界。定理 6 中的理论结果表明，BSP 能够适应正则性指数 $r \in [1/2, \infty$、容量指数 $s \in (0, 1]$ 以及不同的误差度量。这意味着，采用统一的参数选择策略时，对于所有 $r \in [1/2, \infty$、 $s \in (0, 1]$，在三种不同范数下： $\|\cdot\|_\rho$、 $\|\cdot\|_D$和 $\|\cdot\|_K$，均可同时得到最优泛化误差界。这体现了 BSP 的优势：它充分利用了 KGD 的迭代特性。

定理 6 给出了带 BSP 的 KGD 的可行性与最优性。一个关键点是，这些断言对任意 $\tilde{C} \geq \tilde{C}_0 := 32\sqrt{2}(1+\beta)(\kappa M + \gamma)$ 均成立。一方面，从我们的证明中可以看出，常数 C_0 略显悲观。我们相信，借助更精细的分析技术可以将其减小。另一方面， C 的最佳选择注定要通过最小化 (40) 和 (41) 中的常数C 来优化参数选择策略。回顾 BSP 能够适应不同的误差度量，HSS 随后侧重于选择 C 的一个良好候选，以在 $\|\cdot\|_\rho$ 度量下最小化误差（Caponnetto 和 Yao，2010）。下面这条可直接由定理 6 导出的推论，给出了算法 1 中带 HSS 的 KGD 的理论保证。

推论 7 设 $0 < \delta < 1$ 与 $0 < \beta \leq \kappa^{-1}$。在带有 $r \geq 1/2$的假设*1*和假设*2*下，若 $\hat{}t^*$ 由算法 *1* 在 $\min_{\hat{}C_j \in C_U} \geq \tilde{C}_0$下给出，则以置信度 $1-\delta$，(40) 与 (41) 对 $f_{\hat{t^*}}$成立。

推论 7 表明算法 1 中所提出的 HSS 策略具有最优性。~尽管对参数选择
~ ~策略额外施加了假设 $\min_{\hat{C}_j \in C_U} \geq C_0$ ，这在理论上不可避免地缩小了 C 的范围，但如上所述，对 C_0 的估计过于悲观，并且在数值上可以显著降低。上述所有理论结果表明，在不划分数据的情况下，可以设计一种精巧且完全自适应的参数选择策略，使 KGD 达到其最优泛化误差界。

由于我们的结果对 $\|\cdot\|_K$ 和 $\|\cdot\|_\infty \leq \kappa\|\cdot\|_K$ 成立，我们由定理 6 直接得到如下推论。

推论 8 设 $0 < \delta < 1$ 与 $0 < \beta \leq \kappa^{-1}$。在带有 $r \geq 1/2$ 的假设*1*和假设*2*下，则

$$\|f_{\hat{t},\beta,D} - f_\rho\|_\infty \quad \leq \quad C\kappa \log^4 \frac{16}{\delta} \begin{cases} \left(\frac{\sqrt{L}\log|D|}{\sqrt{|D|}}\right)^{1-1/2r}, & \text{if } \sigma_\ell = 0, \ell \geq L+1, \\ |D|^{-\frac{r-1/2}{2r+s}}, & \text{if } \sigma_\ell \leq c_0 \ell^{-1/s}, \\ \left(\frac{\log|D|}{\sqrt{|D|}}\right)^{(1-1/2r)}, & \text{if } \sigma_\ell \leq c_1 e^{-c_2 \ell^2}. \end{cases} \tag{42}$$

不同于 $\|\cdot\|_\rho$ 中对泛化误差的描述，上述推论在 $\|\cdot\|_\infty$ 中推导出了误差界，这意味着该方法能够处理协变量偏移问题（Ma 等人，2023年），即测试数据与训练数据的边缘分布不同。实际上，对于任意 ρ'_X，都很容易推导出 $\|\cdot\|_{\rho'} \leq \|\cdot\|_\infty$。我们应当强调，所推导出的 $\|\cdot\|_{\rho'}$ 界不如 Ma 等人（2023年）的结果，后者的算法利用了分布信息。我们处理协变量偏移问题的结果与 Ma 等人（2023年）的不同。一方面，我们的算法与分布无关，且结果不受 ρ_X 与 ρ'_X 之间相似性的影响。另一方面，我们引入了更严格的度量 $\|\cdot\|_K$ 或 $\|\cdot\|_\infty$ 来衡量泛化误差；至少对于 $\|\cdot\|_K$ 和多项式衰减的 L_K，该度量在极小极大意义下是最优的（Fischer 和 Steinwart，2020年）。

4. 数值分析

本节中，我们通过玩具模拟和两项真实世界数据实验，展示采用算法 1 所示 HSS 参数选择策略的 KGD 的数值特性，并验证定理 6、推论 7 和推论 8。除模拟 2 中的对比实验外，所有其他实验均使用 Python 3.7，在配备英特尔酷睿 i5 2 GHz 处理器的个人电脑上运行。复现实验的脚本可在 https://github.com/Ariesoomoon/HSS_experiments 获取。

4.1 模拟实验

我们进行了三次模拟。第一次模拟展示了所提出的自适应停止规则 BSP 的可行性与效力。第二次模拟说明了

HSS 的有效性与优越性能，方法是将其与其他参数选择策略进行比较，包括基准策略 BS、信息熵方法赤池信息准则和贝叶斯信息准则、分割类型策略高阶，以及偏差–方差分析策略，如反向传播、线性规划、ESR 和动态规划。第三次模拟凸显了 HSS 在克服协变量偏移方面的能力。为充分展示 HSS 的数值最优性，我们采用两种误差度量： L_2 范数和 L_∞ 范数。每次模拟进行 10 次以取平均。

我们生成 $|D| \in \{200, 400, \ldots, 6000\}$ 个样本用于训练，其中 $|D|$ 表示数据集 $D = \{(x_i, y_i)\}_{i=1}^{|D|}$ 的基数。输入 $\{x_i\}_{i=1}^{|D|}$ 是从边长为 $[0,1]^d$、且 $d = 1, 3$ 的（超）立方体上的均匀分布中独立抽取的。相应的输出由回归模型 $y_i = g_j(x_i) + \varepsilon_i, \ i = 1, \ldots, |D|, \ j = 1, 2$ 生成，其中 ε_i 是独立高斯噪声 $\mathcal{N}(0, \sigma^2)$，且 $\sigma = 0.6$,

$$g_1(x) := \begin{cases} x, & \text{if } 0 \le x \le 0.5, \\ 1 - x, & \text{if } 0.5 < x, \end{cases} \tag{43}$$

and

$$g_2(\|x\|_2) := \begin{cases} (1-\|x\|_2)^6(35\|x\|_2^2+18\|x\|_2+3), & \text{if } 0 \le \|x\|_2 \le 1, x \in \mathbb{R}^3, \\ 0, & \text{if } \|x\|_2 > 1. \end{cases} \tag{44}$$

此外，还生成了 $|D'| = 500$ 为 $\{(x_i', y_i')\}_{i=1}^{|D'|}$ 的测试样本，其中 $\{x_i'\}_{i=1}^{|D'|}$ 是根据均匀分布独立抽取的，且 $y_i' = g_j(x_i') \ j = 1, 2$。采用的核函数为 $K_1(x, x') = 1 + \min(x, x')$，适用于 $d = 1$；而 $K_2(x, x') = h(\|x - x'\|_2)$，其满足

$$h(\|x\|_2) := \begin{cases} (1 - \|x\|_2)^4(4\|x\|_2 + 1), & \text{if } 0 < \|x\|_2 \le 1, x \in \mathbb{R}^3, \\ 0, & \text{if } \|x\|_2 > 1, \end{cases} \tag{45}$$

对于 $d = 3$。(Wu，1995年；Schaback 和 Wendland，2006年) 中指出， $g_1 \in W_1^1$和 $g_2 \in W_3^4$，其中 W_d^α表示 $[0,1]^d$上的 α阶 Sobole夫空间； K_1和 K_2分别是 W_1^1和 W_3^2的再生核，$g_1 \in \mathcal{H}_{K_1}$和 $g_2 \in \mathcal{H}_{K_2}$亦然。随后，使用核 K_1 和 K_2 的 KGD算法被应用于拟合训练样本。

需要说明的是，在所有模拟中，在 HSS 的 L 选定样本中，训练样本$D_{tr、L}$ 与验证样本 $D_{val、L}$ 的比例固定为 7:3。对于需要常数参数选择的其他偏差–方差分析方法，也采用相同比例。对于高阶，训练样本与验证样本的比例固定为 1:1。为确保公平比较，所有玩具模拟中参数选择方法的步长 β 均设为相同值： $\beta = 1$ 用于 $d = 1$， $\beta = 3$ 用于 $d = 3$。

4.1.1 Simulation 1: Feasibility and power of BSP

在本次模拟中，我们开展了两组实验。首先，我们固定数据˜规模 $|D| = 2000$，并观察两种不同范数如何随 BSP 的常数 C 变化。˜ ˆ 其次，我们考察了不同数据规模下最优 C （记为 C_{j^*}）和最优 t（记为ˆt^*）所在的范围。本次模拟有两个主要观察结果。

首先，在图 3 中， L_2 范数和 L_∞ 范数相对于 C都呈现出一个全局最小值 ˜ ˜，这表明通过选择 C来利用 BSP 为 KGD 配备相应方法是可行的。不同范数表现出的相似行为凸显了 BSP 的强大之处。请注意，在图 3 中，每个子图的两侧分别表示 C 过大或过小的情况，因此 BSP 确定的 ˆt^* 为 T。

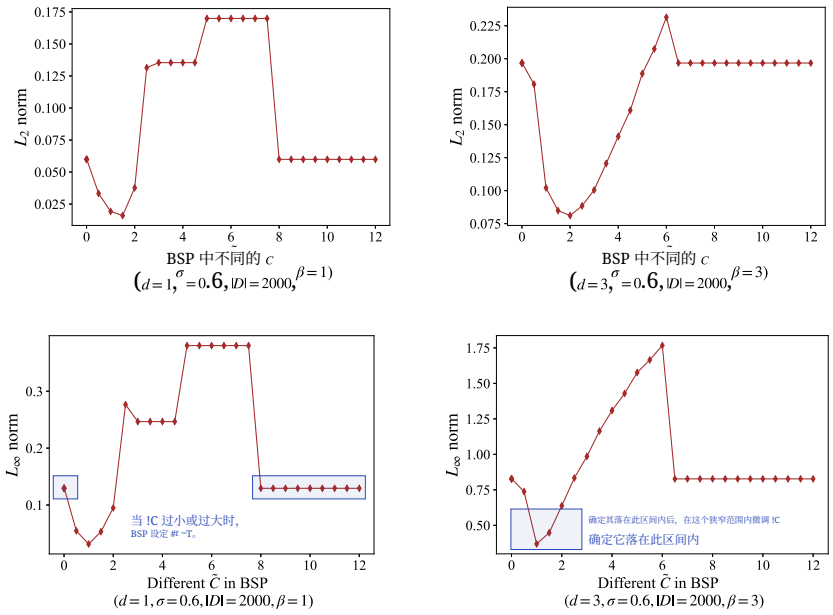

图 3： L_2 范数/L_∞ 范数与常数 $\tilde{C}$ 之间的关系。

其次，我们观察到，最优 C 落在相对较窄的范围内（$[0, 4]$），因此更容易确定。这表明了一种高效的参数选择方法：首先，使用对数尺度快速确定 C 的大致范围（即选择 $C \in \{0, 2^0, 2^1, 2^2, \dots\}$），然后在该范围内使用均匀尺度进一步细化选择（即选择 $C \in \{0.001, 0.002, \dots\}$）。这种方法称为对数均匀参数选择，并用于我们的实验；最终将 $\tilde{C}$ 的选择区间设为 2^{-10}（采用均匀尺度）。

需要强调的是，此参数选择过程不适合直接用于选择 $\hat{t}^*$，因为如图4所示，我们观察到$\hat{t}^*$ 的范围远大于 $[0,\ 800]$$\hat{}$ C_{j*}的 $[0,\ 4]$, ，并且如图1所示，泛化误差会随着 t逐渐变化。

4.1.2 SIMULATION 2: EFFECTIVENESS AND SUPERIOR PERFORMANCE OF HSS

在本次模拟中，我们展示了 HSS 的计算效率与预测性能，从而验证了定理 6 和推论 7。

图5比较了 BS、高阶和 HSS 在不同数据规模下的预测性能，其中 HSS 中的 L 固定为 $0.6|D|$。我们观察到，在 L_2 范数下，HSS 的性能与高阶相当，二者均接近 BS。然而，在 L_∞ 范数下，HSS 显著优于高阶，且更接近 BS 的性能。这表明，自适应停止规则 BSP 能够基于精心选取的常数 $\tilde{C}$确定高质量的步 $\hat{t}^*$，从而验证了我们的定理 6。这凸显了 HSS 的有效性与优势，尤其是其对不同范数的适应性，从而支持了推论 7。

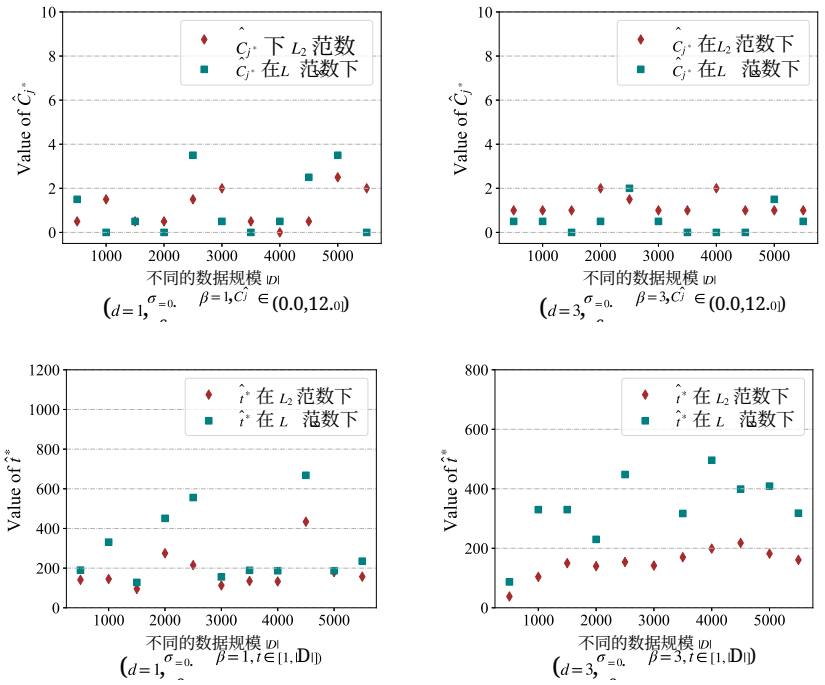

图4： BSP下$\hat{C}_{j*}$ 的范围以及BS下$\hat{t}^*$ 的范围（单次实验结果）。

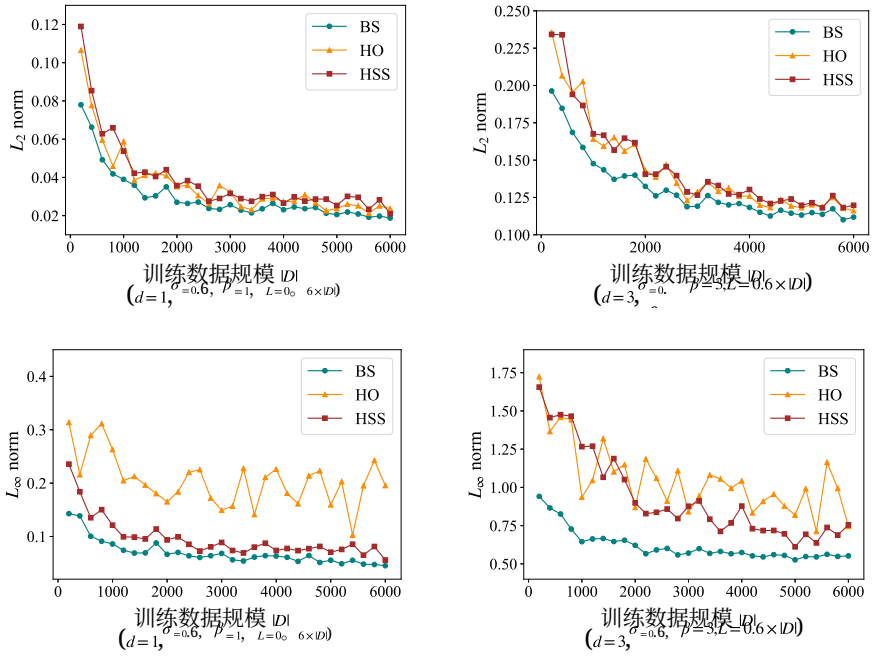

图5： BS、高阶和HSS的泛化性能（其中，HSS中的 L 固定为 $0.6|D|$）。

此外，我们从效率和准确率两方面比较了 HO、赤池信息准则、贝叶斯信息准则、HSS 及其他偏差–方差分析方法。文中还给出了采用 Lu 等人（2020年）的常数 C_{BP} 的 BP，以及采用 Raskutti 等人（2014年）的常数 C_{ESR} 的 ESR 的结果。由于 BP 和 LP 对内存的消耗很大，以下所有结果均在配备 NVIDIA RTX 4090（24GB）图形处理器、16 核 Intel Xeon Platinum 8352V 处理器（2.10GHz）和 120GB 随机存取存储器的系统上，使用 Python 3.7（Ubuntu 18.04）生成。比较仅涵盖 1000 和 1200 的数据规模，维度为 $d = 1$ 和 $d = 3$。每项结果均重复进行 10 次并取平均值。

表 2：实验 设置（适用于表 3 和表 4 中的所有偏差–方差分析方法）

设置	详情	设置	详情
备选参数数量	24	步长	$\beta = 1$ 对于 $d = 1$，$\beta = 3$ 对于 $d = 3$
参数选择区间	0.05	训练数据规模 $\lvert D\rvert$	1000, 1200
已选出最优参数	Yes	测试数据规模 $\lvert D'\rvert$	500
试验次数	10	准确性指标	L_2 范数，L_∞ 范数
最大迭代次数	$T = \lvert D\rvert$	效率指标	最大内存使用量（MMU）、时间（秒）

我们精心设计了实验，以确保比较公平，具体设置详见表 2。在准确率方面，我们使用了 L_2 范数和 L_∞ 范数；在效率方面，我们测量了每种算法执行期间的执行时间和最大内存使用量（MMU）。执行时间和 MMU 均单独记录。

表 3：数值性能 在 $|D| = 1000$ 时不同参数选择方法的数值性能

不同参数选择方法的数值性能								
方法	$d = 1$				$d = 3$			
	L_2 范数	L_∞ 范数	时间（秒）	MMU（GB）	L_2 范数	L_∞ 范数	时间（秒）	MMU（GB）
HO	0.0587	0.2630	1.03	0.43	0.1641	0.9370	15.98	0.54
AIC	0.0742	0.2407	7.26	0.66	0.1738	1.0100	110.76	0.80
BIC	0.0750	0.2995	8.85	0.65	0.2356	1.0573	109.75	0.81
BP	0.0643	0.1817	924.71	52.40	0.1680	1.0030	1081.37	53.38
Use C_{BP} in (Lu et al., 2020)	**0.0528**	**0.1299**	512.08	52.62	0.2095	1.7912	533.16	52.94
LP	0.0658	0.2394	1618.29	76.65	**0.1631**	0.9624	1543.62	75.79
ESR	0.0598	0.2604	4.88	0.92	0.1632	1.0001	93.09	1.03
Use C_{ESR} in Raskutti et al. (2014)	0.0660	**0.1599**	4.00	0.47	0.2085	**0.7784**	24.23	27.48
DP	**0.0566**	0.2788	4.32	0.49	**0.1617**	**0.9307**	87.01	0.62
HSS 使用 $L = \lvert D\rvert$	**0.0506**	**0.1216**	8.71	0.51	**0.1571**	**0.8633**	108.17	0.62

注：每项准确性指标的结果分别以红色、蓝色和绿色突出显示，依次表示最佳、次佳和第三佳表现。表4也适用此规则。

详细结果如表 3 和表4所示，其中 HSS 中的 L 固定为 $|D|$。就准确率而言，HSS 在 L_2 范数下的表现与高阶相近，但在 L_∞ 范数下显著更优，进一步证明了 HSS 的优越性能。就效率而言，尽管 HSS 的效率略低于高阶，但其表现仍处于可接受范围内。我们还观察到，反向传播的运行时间远高于 HSS，因为反向传播在每一步 t 都需要对每个常数进行逐项比较，而 HSS 仅执行两次连续迭代。在将使用选定常数 C 的 ESR˜和反向传播的结果与使用固定常数 C_{ESR} 和 C_{BP}的结果进行比较时，

表4: 当 $|D| = 1200$时不同参数选择方法的数值性能

不同参数选择方法的数值性能								
方法	$d = 1$				$d = 3$			
	L_2 范数	L_∞ 范数	时间（秒）	MMU（GB）	L_2 范数	L_∞ 范数	时间（秒）	MMU（GB）
HO	**0.0384**	0.2046	1.52	0.45	0.1594	1.0457	21.77	0.60
AIC	0.0439	0.1548	11.58	0.68	0.1503	1.0330	158.02	0.87
BIC	0.0485	0.2439	12.50	0.68	0.2169	**1.0292**	163.58	0.92
BP	0.0423	0.1346	1524.22	91.37	0.1513	1.0835	1788.85	91.48
Use C_{BP} in (Lu et al., 2020)	0.0456	**0.1164**	886.35	90.97	0.2050	1.7753	882.45	90.20
LP	0.0418	**0.1249**	1604.03*	76.32*	0.1531	1.0692	1750.27*	76.34*
ESR	**0.0391**	0.1906	7.42	0.95	**0.1494**	1.0864	133.70	1.06
Use C_{ESR} in Raskutti et al. (2014)	0.0566	0.1398	6.50	0.54	0.2062	**0.7983**	38.86	0.66
DP	0.0606	0.1635	6.90	0.51	**0.1503**	1.1473	123.31	0.65
HSS 与 $L = \|D\|$	**0.0393**	**0.1137**	12.26	0.51	**0.1492**	**0.8180**	158.00	0.64

注：当数据规模为 1200 时，线性规划计算因内存使用量过大（117 GB，占系统总内存 120 GB 的 97.5%）而无法运行。我们修改了代码，使其在执行过程中尽可能释放内存，从而使其能够正常运行。使用修改后的代码进行线性规划计算所得的结果以 $*$ 表示。

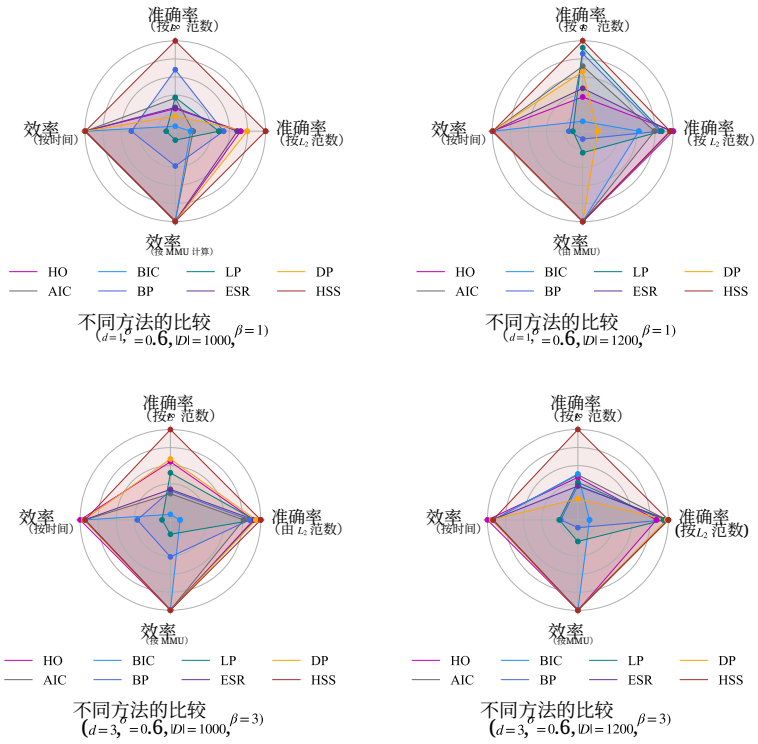

图6: 表 3 和.....中所列方法数值性能的示意图

4.

˜我们发现，在大多数情况下，使用固定 C 所得结果不如通过参数选择获得的结果。这表明，尽管 ESR 和反向传播展现出良好的理论˜性能，但其数值有效性在很大程度上取决于 C的质量。

为了清晰概述上述两张表中所有方法的比较结果，我们绘制了雷达图，如图6所示。我们调整了 MMU 和时间这两项指标，使其表示方式与准确率类似；在该图中，数值越接近边缘（或覆盖面积越大），

表示性能越好。根据图6，我们得出以下观察结果：(1) 在准确率和效率这四项指标上，HSS 的总体表现最佳。与其他算法相比，它在L_∞ 范数方面的优势尤为明显。(2) HO、AIC、BIC、ESR 和 DP 的效率表现良好，但准确率总体上不如 HSS。(3) BP 和 LP 在运行时间和 MMU 方面的表现都明显较差，主要是因为它们依赖于对每个给定常数进行逐项比较（见式（16）和（17））。

雷达图的结果表明，在信息熵方法、分裂方法和偏差–方差分析方法中，HSS 在效率和准确率两方面都表现出色。

4.1.3 Simulation 3: HSS overcoming covariate shift

该模拟表明，HSS 能有效解决协变量偏移问题，从而验证了推论 8，并为基于RKHS的非参数回归中的协变量偏移研究作出了贡献（Ma等，2023）。具体而言，Ma 等人（2023）关注核岭回归（KRR），并提出了一种称为截断重加权 KRR 估计量的变体，该估计量达到了极小极大速率最优性（Ma等，2023，定理 4）。相比之下，我们关注 KGD。

为简便起见，我们设定 $d = 1$，并假设 $\{x_i'\}_{i=1}^{|D'|}$ 是从 (超)立方体 $[0, b]$ 上的均匀分布中独立抽取的，其中 $b \in \{1.1, 1.2, \ldots, 1.5\}$。我们使用 Kullback-Leibler（KL）散度来量化分布差异。对于连续概率分布 P 和 Q，从 Q 到 P 的 KL 散度记为 $\mathbb{D}_{\mathrm{KL}}(P\|Q)$，定义为 $\mathbb{D}_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(\mathbf{x}) \ln\left(\frac{p(\mathbf{x})}{q}(\mathbf{x})\right) d\mathbf{x}$，其中 $p(\mathbf{x})$ 和 $q(\mathbf{x})$ 分别是 P 和 Q 的概率密度函数。为计算 KL 散度（下文记为 $\overline{\mathbb{D}_{\mathrm{KL}}}$），我们使用核密度估计（带宽为 0.05）获得概率密度函数，并使用高斯-勒让德求积进行积分。

图7展示了协变量偏移下的结果。图7(a)和图7(b)表明，HSS始终优于高阶（HO），其中“ΔL_2 范数 (%)”表示协变量偏移下的 L_2范数相对于无偏移时的 L_2 范数的变化，而“ΔL_∞ 范数 (%)”含义类似。图7(c)和图7(d)通过箱线图展示了HSS和HO在10次试验中的详细结果，可以看出，随着 $\mathbb{D}_{\mathrm{KL}}$ 变化，HSS的波动较小。这些结果证实了HSS处理协变量偏移的稳健性，从而支持推论 8 中所述的、在 $\|\cdot\|_\infty$度量下具有理论最优性的泛化误差界保证。

我们进一步探究了 HSS 在协变量偏移下使用不同 L时的性能，如图 8 所示。虚线表示当 $L = 0.5|D|$时 HSS 的性能。我们发现，在协变量偏移下，使用较小的 L 通常比使用较大的 L 略有预测性能优势，尤其是在 L_∞ 范数下。这可能是因为使用较少的训练数据来选择 L 有助于避免过拟合，从而提升协变量偏移下的性能。

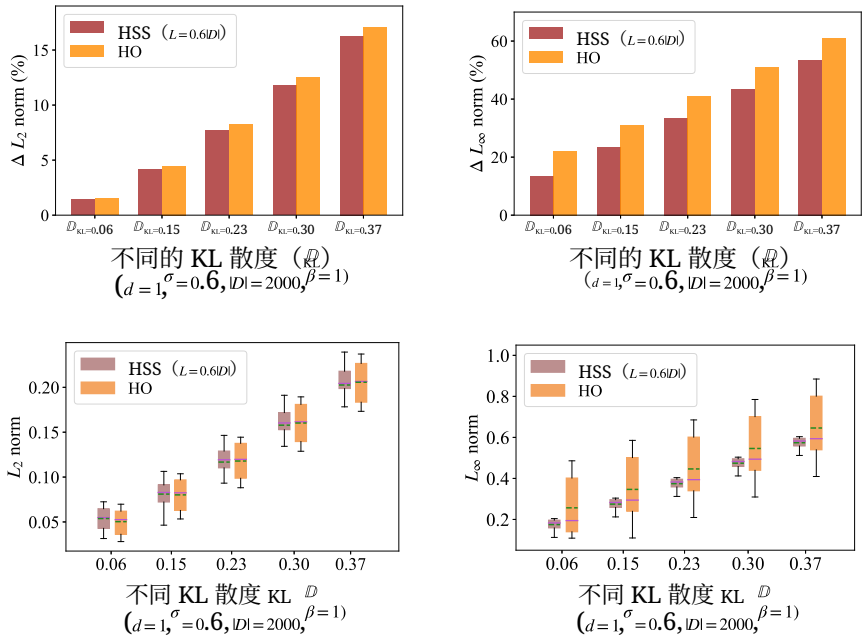

图7： 协变量偏移下的结果。

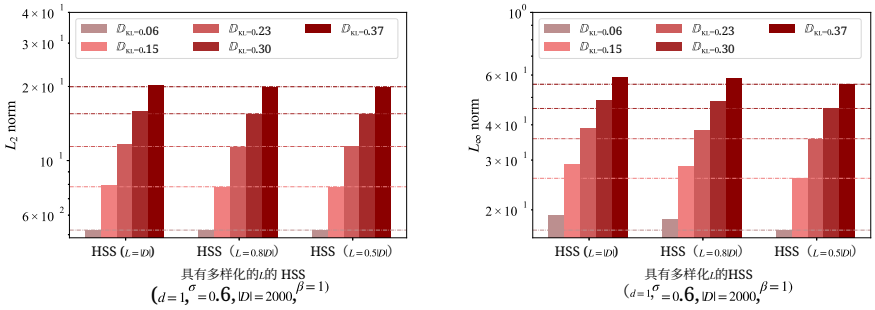

图8： 协变量偏移下的结果。

4.2 真实数据示例

最后，我们在 L_2 范数下，使用两个真实世界数据集评估所提出的 HSS。这些数据集涉及地球表面不同纬度、经度和海拔处的磁场信息，具体为磁场总强度数据集和磁偏角数据集。它们对导航、定位、地质勘探等多种应用至关重要。这两个数据集采集自网站 https://geomag.bgs.ac.uk/data_service/models_compass/ igrf_calc.html，其地磁场信息由第 13 代国际地磁参考场（IGRF-13）提供（Alken 等，2021），依据卫星和地面观测站记录的观测数据生成。实验中，我们仅使用 2024年8月15日的纬度 ϕ、经度 θ、海拔 h、相应的总强度（ nT ）和磁偏角（度）属性。

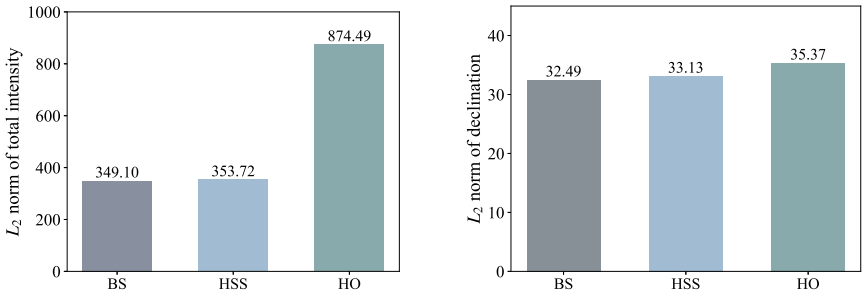

图 9： BS、HO 和 HSS 对磁场总强度数据和磁偏角数据的泛化性能。对于 HSS，总强度数据中的 L 设为 $0.8|D|$，磁偏角数据中的 L 设为 $0.9|D|$ 。

为生成训练样本，我们首先根据超立方体 $[-1,1]^d$ 上的均匀分布抽取 2000 个样本，其中 $d=3$；然后采用与 (Liu 等，2024) 相同的数据处理和收集策略，从上述网站收集总强度数据和磁偏角数据。我们从以下位置收集 2664 个样本作为测试样本： ϕ 和 θ 每隔五度取样， h 固定为 0 km；并将这些样本以米勒圆柱投影可视化。与 Liu 等（2024）类似，我们向总强度数据的训练样本添加标准差为 $\sigma=500$ 的截断高斯噪声，并向磁偏角数据的训练样本添加标准差为$\sigma=20$ 的截断高斯噪声，从而得到含噪数据。我们对 HSS 中的常数 C 采用与玩具模拟中相同的˜参数选择。为公平比较 HSS、BS 和 HO，将总强度数据的步长 β 设为 45，将磁偏角数据的步长设为 20。每项实验均进行五次，并取平均值。

我们在 HSS 中为磁总强度数据设置了 $L=0.8\times 2000$ ，为磁偏角数据设置了 $L=0.9\times 2000$ 。图9显示，HSS 的表现几乎与 BS 相同，且优于 HO（尤其是在总强度数据上），这证明了 HSS 的优势。图10展示了与总强度相关的全球地图可视化结果，包括 IGRF-13 提供的总强度数据（ Alken 等，2021 ），以及由 HO 和 HSS 得到的总强度拟合数据。我们在每幅地图中用浅白色矩形标出了两个区域，以突出它们之间的差异。与真实值相比，HSS 给出的预测更接近 IGRF-13 提供的结果。

5. 进一步讨论

偏差–方差分析方法的半自适应特性以及信息熵方法的次优性，限制了它们在参数选择中的有效性。在此背景下，分割方法——尤其是交叉验证和留出法——在实践中仍占主导地位。然而，由于需要牺牲一部分样本来创建验证集，分割方法在实践中往往表现不佳。为避免舍弃任何样本，我们提出 HSS 方法，将偏差–方差分析方法与分割方法的优势相结合，并针对所有正则性指数 $r\in[1/2,\infty)$、容量指数 $s\in(0,1]$，以及不同的误差度量（$\|\cdot\|_\rho$、$\|\cdot\|_D$ 和 $\|\cdot\|_K$）实现最优泛化误差界。

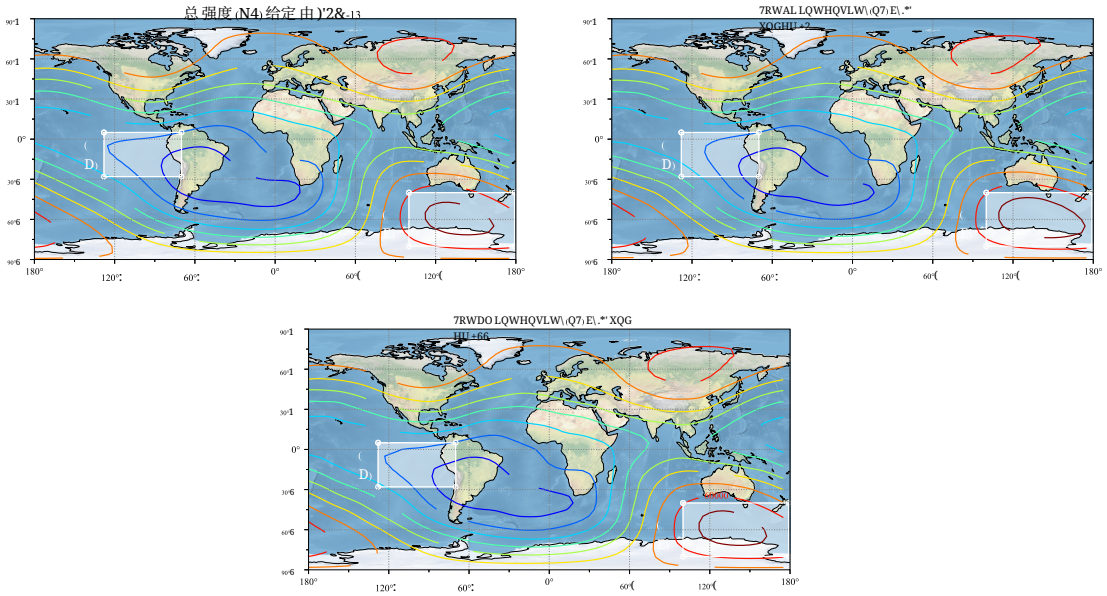

图10： 2024年8月15日，WGS84椭球体表面由 IGRF-13 给出的磁总强度分布图（上图），以及 KGD 在 HO（中图）和 KGD 在 HSS（下图）下的预测结果。所用投影为米勒圆柱投影。

值得注意的是，由于 HSS 无需依赖整个数据集来选择常数，未来的研究可以探索其在分布式学习系统中的应用，其中每个本地代理都根据自己的本地数据选择常数。基于本文的理论结果，将 HSS 与分布式学习相结合，有望推动保护隐私的分布式基于核的梯度下降算法的发展，同时保持较高的预测准确度。

此外，鉴于针对球面数据的核方法已有大量研究（例如我们在真实数据实验中使用的磁总强度数据），且许多现实世界的球面数据集都是通过确定性采样获取的（例如卫星确定性采样），针对这类球面数据专门为 KGD 设计一种完全自适应的参数选择策略，将是未来一项值得研究的方向。

致谢

本工作部分受到国家自然科学基金委员会（项目编号62276209）的资助。

参考文献

赤池弘次。重新审视统计模型识别。*IEEE*自动控制汇刊，19(6):716–723，1974年。

Patrick Alken、Erwan Th´ebault、Ciar´an D Beggan、Hagay Amit、J Aubert、J Baerenzung、TN Bondar、WJ Brown、S Califf、A Chambodut 等。国际地磁参考场：第十三代。《地球、行星与空间》，73:1–25，2021年。Frank Bauer、Sergei Pereverzev 和 Lorenzo Rosasco。学习理论中的正则化算法。*J. Complex.*，23(1):52–72，2007。Yoshua Bengio、Ian Goodfellow 和 Aaron Courville。《深度学习》，第1卷。麻省理工学院出版社，2017年。Gilles Blanchard 和 Nicole Kr¨amer。随机设计回归中核共轭梯度的收敛率。*Anal. Appl.*，14(06):763–794，2016。Gilles Blanchard、Peter Math´e 和 Nicole M¨ucke。监督学习中的 Lepskii 原理。*arXiv*预印本 *arXiv:1905.10764*，2019年。Andrea Caponnetto 和 Ernesto De Vito。正则化最小二乘算法的最优速率。*Found. Comput. Math.*，7(3):331–368，2007。Andrea Caponnetto 和 Yuan Yao。基于交叉验证的学习理论中正则化算子自适应方法。*Anal. Appl.*，8(02):161–183，2010。Alain Celisse 和 Martin Wahl。分析核化谱滤波学习算法的偏差原则。*J. Mach. Learn. Res.*，22:1–76，2021。Olivier Chapelle、Vladimir Vapnik、Olivier Bousquet 和 Sayan Mukherjee。支持向量机多参数的选择。*Mach. Learn.*，46:131–159，2002。Gerda Claeskens、Christophe Croux 和 Johan Van Kerckhoven。支持向量机变量选择的信息准则。*J. Mach. Learn. Res.*，9:541–558，2008。Ernesto De Vito、Sergei Pereverzyev 和 Lorenzo Rosasco。利用平衡原理的自适应核方法。*Found. Comput. Math.*，10(4):455–479，2010。Sergey Demyanov、James Bailey、Kotagiri Ramamohanarao 和 Christopher Leckie。基于赤池信息准则和贝叶斯信息准则的径向基函数核支持向量机参数估计方法。发表于亚洲机器学习会议，第97–112页，2012。Marius-Constantin Dinu、Markus Holzleitner、Maximilian Beck、Hoan Duc Nguyen、Andrea Huber、Hamid Eghbal-zadeh、Bernhard A. Moser、Sergei Pereverzyev、Sepp Hochreiter 和 Werner Zellinger。通过聚合解决无监督领域自适应中的参数选择问题。发表于第十一届国际学习表征会议，2023。Heinz Werner Engl、Martin Hanke 和 Andreas Neubauer。《逆问题正则化》，第375卷。施普林格科学与商业媒体，1996年。Simon Fischer 和 Ingo Steinwart。正则化最小二乘算法的 Sobolev 范数学习率。*J. Mach. Learn. Res.*，21(205):1–38，2020年。

Junichi Fujii、Masatoshi Fujii、Takayuki Furuta和Ritsuo Nakamoto。等价于海因茨不等式的范数不等式。美国数学会会刊，118(3):827–830，1993年。L Lo Gerfo、Lorenzo Rosasco、Francesca Odone、E De Vito和Alessandro Verri。用于监督学习的谱算法。*Neural Comput.*，20(7):1873–1897，2008年。Elke R Gizewski、Lukas Mayer、Bernhard A Moser、Duc Hoan Nguyen、Sergiy Pereverzyev Jr、Sergei V Pereverzyev、Natalia Shepeleva和Werner Zellinger。关于RKHS中无监督领域自适应的一种正则化。《应用计算调和分析》，57:201–227，2022年。Zheng-Chu Guo、Shao-Bo Lin和Ding-Xuan Zhou。分布式谱算法的学习理论。《反问题》，33(7):074009，2017年。L´aszl´o Gy¨orfi、Michael Kohler、Adam Krzy˙zak和Harro Walk。非参数回归的分布无关理论，第1卷。施普林格，2002年。Omar Hagrass、Bharath Sriperumbudur和Bing Li。谱正则化核两样本检验。《统计年刊》，52(3):1076–1101，2024年。Peter Hall和JS Marron。论非参数回归中的方差估计。《生物统计学》。，77(2):415–419，1990年。Ting Hu、Qiang Wu和Ding-Xuan Zhou。面向最小误差熵原理的分布式核梯度下降算法。《应用计算调和分析》，49(1):229–256，2020年。Clifford M. Hurvich、Jeffrey S. Simonoff和Chih-Ling Tsai。使用改进的赤池信息准则进行非参数回归中的平滑参数选择。《英国皇家统计学会杂志 *B*辑：统计方法学》，60(2):271–293，1998年。Raytcho D Lazarov、Shuai Lu和Sergei V Pereverzev。论数值分析若干问题中的平衡原理。《数值数学》，106:659–689，2007。Shao-Bo Lin。核岭回归的自适应参数选择。《应用计算调和分析》，第101671页，2024年。Shao-Bo Lin 和 Ding-Xuan Zhou。分布式基于核的梯度下降算法。《构造逼近》，47(2):249–276，2018a。Shao-Bo Lin 和 Ding-Xuan Zhou。核偏最小二乘的最优学习速率。《傅里叶分析与应用杂志》，24:908–933，2018b。Shao-Bo Lin、Xin Guo和Ding-Xuan Zhou。基于正则化最小二乘的分布式学习。*J. Mach. Learn. Res.*，18(1):3202–3232，2017年。Shao-Bo Lin、Yunwen Lei和Ding-Xuan Zhou。提升核岭回归：最优学习速率与提前停止。*J. Mach. Learn. Res.*，20(1):1738–1773，2019年。Xiaotong Liu、Jinxin Wang、Di Wang和Shao-Bo Lin。用于球面上核插值的加权谱滤波器：含噪数据预测准确度的估计。《*SIAM*影像科学杂志》，17(2):951–983，2024年。

Shuai Lu、Peter Math´e和Sergei V Pereverzev。一般正则化方案下监督学习中的平衡原理。《应用计算调和分析》，48(1):123–148，2020年。Cong Ma、Reese Pathak和Martin J Wainwright。基于RKHS的非参数回归中协变量偏移的最优处理。《统计年刊》，51(2):738–761，2023年。Drton Mathias和Plummer Martyn。奇异模型的贝叶斯信息准则。《英国皇家统计学会杂志 *B*辑：统计方法学》，79(2):323–380，2017年。Ryuei Nishii。多元回归中变量选择准则的渐近性质。《统计年刊》，第758–765页，1984年。Garvesh Raskutti、Martin J Wainwright和Bin Yu。经由凸规划得到的核类上稀疏加性模型的极小极大最优速率。《机器学习研究杂志》，13(1):389–427，2012年。Garvesh Raskutti、Martin J Wainwright和Bin Yu。提前停止与非参数回归：一种最优的数据依赖停止规则。*J. Mach. Learn. Res.*，15(1):335–366，2014年。Philip T Reiss和R Todd Ogden。一类半参数线性模型的平滑参数选择。《英国皇家统计学会杂志 *B*辑：统计方法学》，71(2):505–523，2009年。Robert Schaback和Holger Wendland。核技术：从机器学习到无网格方法。《数值学报》，15:543–639，2006年。Gideon Schwarz。估计模型的维数。《统计年刊》，第461–464页，1978。Steve Smale和Ding-Xuan Zhou。经由积分算子及其近似得到的学习理论估计。《构造逼近》，26(2):153–172，2007。Ingo Steinwart和Andreas Christmann。支持向量机。施普林格科学与商业媒体，2008年。Yuting Wei、Fanny Yang和Martin J Wainwright。核提升算法的提前停止：基于局部复杂度的一般分析。《*IEEE*信息论汇刊》。，65(10):6685–6703，2019年。Zongmin Wu。紧支撑正定径向函数。《计算数学进展》，4:283–292，1995年。Ganggang Xu、Zuofeng Shang和Guang Cheng。面向分治核岭回归的分布式广义交叉验证及其渐近最优性。《计算与图形统计学杂志》，28(4):891–908，2019年。Yuan Yao、Lorenzo Rosasco和Andrea Caponnetto。论梯度下降学习中的提前停止。《构造逼近》，26(2):289–315，2007。Tong Zhang。核学习的有效维数与泛化。《神经信息处理系统进展》，15，2002年。

Tong Zhang。核方法的留一法界。*Neural Comput.*，15(6):1397–1437，2003年。

Yuchen Zhang、John Duchi 和 Martin Wainwright。分而治之的核岭回归：一种具有极小极大最优速率的分布式算法。*J. Mach. Learn. Res.*，16(1):3299–3340，2015年。

Ap附录 A：算子表示与算子集中性

我们分析中使用的主要工具是 (Smale and Zhou, 2007; Lin 等人, 2017; Guo 等人, 2017) 提出的积分算子方法。令 $S_D: \mathcal{H}_K \to \mathbb{R}^{|D|}$ 采样算子（Smale and Zhou, 2007）定义为

$$S_D f := (f(x))_{(x,y)\in D}.$$

其缩放伴随算子 $S_D^\top : \mathbb{R}^{|D|} \to \mathcal{H}_K$ 定义为

$$S_D^\top \mathbf{c} := \frac{1}{|D|}\sum_{i=1}^{|D|} c_i K_{x_i}, \qquad \mathbf{c} := (c_1, c_2, \ldots, c_{|D|})^\top \in \mathbb{R}^{|D|}.$$

定义 $L_{K,D}: \mathcal{H}_K \to \mathcal{H}_K$ 积分算子 L_K 的经验版本为

$$L_{K,D} f := \frac{1}{|D|}\sum_{(x,y)\in D} f(x) K_x.$$

对于任意 $f \in \mathcal{H}_K$，容易验证

$$\|f\|_D = \|L_{K,D}^{1/2} f\|_K, \qquad \text{and} \quad \|f\|_\rho = \|L_K^{1/2} f\|_K. \tag{A1}$$

梯度下降算法 (3) 可改写为

$$f_{t+1,\beta,D} = f_{t,\beta,D} - \beta\left(L_{K,D} f_{t,\beta,D} - S_D^\top y_D\right) = (I - \beta L_{K,D}) f_{t,\beta,D} + \beta S_D^\top y_D. \tag{A2}$$

直接计算（Yao 等人，2007；Lin 和 Zhou，2018a）可得

$$f_{t,\beta,D} = \sum_{k=0}^{t-1} \beta \pi'_{k+1,\beta}(L_{K,D}) S_D^\top y_D = g_{t,\beta}(L_{K,D}) S_D^\top y_D, \tag{A3}$$

其中 $\pi'_{k+1,\beta}(u) := \Pi_{\ell=k+1}^{t-1}(1-\beta u) = (1-\beta u)^{t-k-1}$, $\pi'_{t,\beta} := 1$ 和

$$g_{t,\beta}(u) = \sum_{k=0}^{t-1} \beta \pi'_{k+1,\beta}(u) = \frac{1-(1-\beta u)^t}{u}, \qquad \forall u > 0. \tag{A4}$$

由 (A3) 给出的 KGD 算子表示具有重要意义，在整个 KGD 分析过程中都会用到。以下引理在逆问题（Engl 等人，1996，第 6 章）和非参数回归（Bauer 等人，2007；Yao 等人，2007；Lin 和 Zhou，2018a）中广为人知，它揭示了这一表示的一些重要性质。

引理 A.1 令 $g_{t,\beta}$ 由式（*A4*）定义。我们有

$$\|L_{K,D} g_{t,\beta}(L_{K,D})\| \le 1, \qquad \frac{1}{t}\|g_{t,\beta}(L_{K,D})\| \le \beta, \tag{A5}$$

and

$$\|L_{K,D}^v (I - L_{K,D} g_{t,\beta}(L_{K,D}))\| \le \left(\frac{v}{t\beta}\right)^v, \qquad v \ge 0. \tag{A6}$$

基于算子表示，我们使用算子差来描述偏差、方差和泛化。下面的引理结合了若干标准集中不等式（Caponnetto 和 De Vito，2007；Blanchard 和 Kr¨amer，2016；Guo 等人，2017年；Blanchard 等人，2019年）（另见 Lin 等人，2019年，引理22）。

引理 A.2 设 $\delta \in (0,1)$，且 D 是一个数据集，其样本根据 ρ独立抽取。在条件 *(15)* 下，以至少 $1-\delta$ 的置信度，

$$\begin{aligned}
\mathcal{S}_{D,t} &\le C_1^* \left(\frac{\log \max\{1, \mathcal{N}(t^{-1})\}}{t^{-1}|D|} + \sqrt{\frac{\log \max\{1, \mathcal{N}(t^{-1})\}}{t^{-1}|D|}} \right), \\
\mathcal{R}_D &\le \frac{4\kappa^2}{\sqrt{|D|}} \log \frac{8}{\delta}, \\
\mathcal{P}_{D,t} &\le 2(\kappa M + \gamma) \mathcal{A}_{D,\lambda} \log\big(8/\delta\big),
\end{aligned}$$

and

$$\begin{aligned}
&(1 + 4\eta_{\delta/4})^{-1} \sqrt{\max\{\mathcal{N}(t^{-1}), 1\}} \le \sqrt{\max\{\mathcal{N}_D(t^{-1}), 1\}} \\
&\quad \le (1 + 4\sqrt{\eta_{\delta/4}} \vee \eta_{\delta/4}^2) \sqrt{\max\{\mathcal{N}(t^{-1}), 1\}}
\end{aligned}$$

同时成立，其中 $C_1^* :=$ 最大值$\{(\kappa^2+1)/3, 2\sqrt{\kappa^2+1}\}$,

$$\mathcal{S}_{D,t} := \|(L_K + t^{-1}I)^{-1/2}(L_K - L_{K,D})(L_K + t^{-1}I)^{-1/2}\|, \tag{A7}$$

$$\mathcal{R}_D := \|L_K - L_{K,D}\|_{HS}, \tag{A8}$$

$$\mathcal{P}_{D,t} := \left\| (L_K + t^{-1}I)^{-1/2}(L_K f_\rho - S_D^\top y_D) \right\|_K, \tag{A9}$$

$\eta_\delta :=$ 2对数$(16/\delta)\sqrt{t}/\sqrt{|D|}$，且 $\|A\|_{HS}$ 是希尔伯特–施密特算子 A的希尔伯特–施密特范数。

根据上述引理，我们可以推导出$\mathcal{Q}_{D,t}$的一个上界。

引理 A. **3** 对于任意 $\delta \in (0,\ 1)$，若 $C_1^* \mathcal{U}_{D,\ t,\ \delta} \le 1/2$，则以 $1-\delta$ 的置信度，有

$$\mathcal{Q}_{D,t} \le \sqrt{2}.$$

证明 直接计算可得

$$\begin{aligned}
&(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2} \\
&= (L_K + \lambda I)^{1/2}\big[(L_{K,D} + \lambda I)^{-1} - (L_K + \lambda I)^{-1}\big](L_K + \lambda I)^{1/2} + I \\
&= I + (L_K + \lambda I)^{-1/2}(L_K - L_{K,D})(L_K + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}.
\end{aligned}$$

由（A7）可得

$$\begin{aligned}
&\|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}\| \\
&\quad \le 1 + \mathcal{S}_{D,t} \|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}\|.
\end{aligned}$$

剩下唯一需要做的是给出对 t 的限制，使得 $\mathcal{S}_{D,\,t}<1$。为此，回顾引理 A.2，可得以 $1-\delta$ 的置信度，有

$$\mathcal{S}_{D,t} \;\leq\; C_1^*\left(\frac{\log\max\{1,\mathcal{N}(t^{-1})\}}{t^{-1}|D|}+\sqrt{\frac{\log\max\{1,\mathcal{N}(t^{-1})\}}{t^{-1}|D|}}\right)\leq C_1^*\mathcal{U}_{D,t,\delta}.$$

因此，对于任意满足条件的 t

$$C_1*\mathcal{U}_{D,t,\delta}\leq 1/2,$$

我们有 $\mathcal{Q}_{D,\,t}\leq\sqrt{2}$。引理 A.3 得证。 ■

由引理 A.2 和引理 A.3，我们可以建立算子差异与 $\mathcal{W}_{D,\,t}$ 之间的关系，这与（Lin 等人，2019年，引理 23）中的关系类似

引理 A.4 对于任意 $0<\delta<1$ 和 $1\leq t\leq T$，在条件 *(15)* 下，以 $1-\delta$ 的置信度，有

$$\mathcal{P}_{D,t}\mathcal{Q}_{D,t}\leq 2\sqrt{2}(\kappa M+\gamma)\mathcal{A}_{D,\lambda}\log\frac{8}{\delta}\leq\sqrt{2}\mathcal{W}_{D,t}\log^2\frac{16}{\delta}. \tag{A10}$$

证明 第一个不等式可直接由引理 A.2 和引理 A.3 得出。接下来我们证明 (A10) 中的第二个不等式。由引理 A.2 和 (30) 可知，以 $1-\delta$ 的置信度，

$$\mathcal{A}_{D,t}\leq\frac{\sqrt{t}}{|D|}+\frac{\sqrt{\max\{\mathcal{N}_D(t^{-1}),1\}}(1+8\sqrt{t/|D|})}{\sqrt{|D|}}\log\frac{16}{\delta}.$$

于是，由引理 A.2 可得，以 $1-\delta$ 的置信度，

$$\mathcal{P}_{D,t}\leq 2(\kappa M+\gamma)\left(\frac{\sqrt{t}}{|D|}+\frac{\sqrt{\max\{\mathcal{N}_D(t^{-1}),1\}}(1+8\sqrt{t/|D|})}{\sqrt{|D|}}\right)\log^2\frac{16}{\delta}.$$

因此，对于任意 $1\leq t\leq T$，根据引理 A.3，以置信度 $1-\delta$，有

$$\mathcal{P}_{D,t}\mathcal{Q}_{D,t}\leq 2\sqrt{2}(\kappa M+\gamma)\log^2\frac{16}{\delta}\left(\frac{\sqrt{t}}{|D|}+\frac{\sqrt{\max\{\mathcal{N}_D(t^{-1}),1\}}(1+8\sqrt{t/|D|})}{\sqrt{|D|}}\right).$$

结合式 (6)，引理 A.4 的证明即告完成。 ■

附录 B：命题和推论的证明

本节重点证明命题。

证明 [命题的证明 2] 根据式 (22) 和 (A3)，我们有

$$f^{\diamond}_{t,\beta,D}=g_{t,\beta}(L_{K,D})L_{K,D}f_\rho. \tag{B11}$$

于是，由 (26)、(16)（其中 $r\geq 1/2$ ）和 (A1) 可得

$$\begin{aligned}\mathcal{B}_{t,\beta,D} &\leq 2\|(L_{K,D}+t^{-1}I)^{1/2}(g_{t,\beta}(L_{K,D})L_{K,D}-I)f_\rho\|_K\\ &\leq 2\|(L_{K,D}+t^{-1}I)^{1/2}(g_{t,\beta}(L_{K,D})L_{K,D}-I)L_K^{r-1/2}\|\|h_\rho\|_\rho.\end{aligned}$$

如果 $\frac{1}{2} \le r \le 1$，由引理 A.1 和 Cordes 不等式（Fujii 等人，1993）可得

$$\|A^\alpha B^\alpha\| \le \|AB\|^\alpha, \qquad 0 \le \alpha \le 1 \tag{B12}$$

对于所有满足以下条件的对称正定正算子 A和 B

$$\begin{aligned}
&\|(L_{K,D}+t^{-1})^{1/2}(g_{t,\beta}(L_{K,D})L_{K,D}-I)L_K^{r-1/2}\| \\
\le\ &\|(L_{K,D}+t^{-1})^{1/2}(g_{t,\beta}(L_{K,D})L_{K,D}-I)(L_{K,D}+t^{-1}I)^{r-1/2}\|\mathcal{Q}_{D,t}^{2r-1} \\
\le\ &2^r\mathcal{Q}_{D,t}^{2r-1}\big(\|L_{K,D}^r(g_{t,\beta}(L_{K,D})L_{K,D}-I)\|+t^{-r}\|g_{t,\beta}(L_{K,D})L_{K,D}-I\|\big) \\
\le\ &2^r\mathcal{Q}_{D,t}^{2r-1}((r/\beta)^r+1)t^{-r}.
\end{aligned}$$

如果 $r > 1$，再次由引理 A.1 可得界 $\|L_{K,D}\| \le \kappa^2$, $\|L_K\| \le \kappa^2$，以及以下 Lipschitz 不等式（Blanchard 和 Kr¨amer，2016）

$$\begin{aligned}
&\|L_{K,D}^{r-1/2}-L_K^{r-1/2}\| \le \|L_{K,D}^{r-1/2}-L_K^{r-1/2}\|_{HS} \\
\le\ &\max\{1,(r-1/2)\kappa^{2r-3}\}\|L_{K,D}-L_K\|_{HS}^{\min\{1,r-1/2\}}
\end{aligned} \tag{B13}$$

that

$$\begin{aligned}
&\|(L_{K,D}+t^{-1})^{1/2}(g_{t,\beta}(L_{K,D})L_{K,D}-I)L_K^{r-1/2}\| \\
\le\ &\|(L_{K,D}+t^{-1})^{1/2}L_{K,D}^{r-1/2}(g_{t,\beta}(L_{K,D})L_{K,D}-I)\| \\
&+\|(L_{K,D}+t^{-1})^{1/2}(g_{t,\beta}(L_{K,D})L_{K,D}-I)\|\|L_K^{r-1/2}-L_{K,D}^{r-1/2}\| \\
\le\ &\left((r/\beta)^r+((r-1/2)/\beta)^{r-1/2}\right)t^{-r}+t^{-1/2}((1/(2\beta))^{1/2}+1) \\
&\times\max\{1,(r-1/2)\kappa^{2r-3}\}\mathcal{R}_D^{\min\{1,r-1/2\}}.
\end{aligned}$$

将上述两个估计代入 (B12)，可得

$$\mathcal{B}_{t,\beta,D} \le C_1'\begin{cases} \mathcal{Q}_{D,t}^{2r-1}t^{-r}, & \text{if } 1/2 \le r \le 1, \\ t^{-r}+t^{-1/2}\mathcal{R}_D^{\min\{1,r-1/2\}}, & \text{if } r > 1. \end{cases}$$

然后，由引理 A.3 可知，以 $1-\delta$ 的置信度，有 $\mathcal{Q}_{D,t}^{2r-1} \le 2^{(2r-1)/2}$, $\forall\, t \le T$ 且 $\mathcal{R}_D \le \frac{4\kappa^2}{\sqrt{|D|}}\log\frac{8}{\delta}$. 这证明了(28)。$\mathcal{V}_{t,\beta,D}$为界定方差，由 (27)、(56)、(48) 和引理 A.1 可得

$$\begin{aligned}
&\mathcal{V}_{t,\beta,D} \le 2\|(L_{K,D}+t^{-1}I)^{1/2}g_{t,\beta}(L_{K,D})(L_{K,D}f_\rho-S_D^\top y_D)\|_K \\
\le\ &2\|g_{t,\beta}(L_{K,D})(L_{K,D}+t^{-1}I)\|\mathcal{Q}_{D,t}\mathcal{P}_{D,t} \\
\le\ &2(1+\beta)\mathcal{Q}_{D,t}\mathcal{P}_{D,t}.
\end{aligned}$$

那么，由引理 A.3 和引理 A.4 可知，以置信度 $1-\delta$，有

$$\mathcal{V}_{t,\beta,D} \le 2\sqrt{2}(\kappa M+\gamma)(1+\beta)\mathcal{A}_{D,\lambda}\log\frac{8}{\delta} \le 2(1+\beta)\mathcal{W}_{D,t}\log^2\frac{16}{\delta}.$$

命题 2 的证明至此完成。 ■

To p为证明推论 3，我们需要以下引理，该引理由 Zhang 等人（2015）推导得出。

引理 B.1 如果存在 $L \in \mathbb{N}$，使得对所有 $\ell \geq L$ 都有 $\sigma_\ell = 0$，那么对于任意 $t \in \mathbb{N}$，有 $\mathcal{N}(t^{-1}) \leq L$。如果存在一个 $s \in (0,1]$，使得 $\sigma_\ell \leq \ell^{-1/s}$，那么对于某个绝对常数 $C_0 > 0$，有 $\mathcal{N}(t^{-1}) \leq C_0 t^s$。如果存在 c_1、$c_2 \geq 0$，使得对所有 $\ell = 1$，都有 $\sigma_\ell \leq c_1 e^{-c_2 \ell^2}$，2，.....，那么 $\mathcal{N}(t^{-1}) \leq C_1 \sqrt{\text{对数}}$ t。

接着，我们利用上述引理和命题 2 来证明推论 3。

证明 [推论 3] 的证明结合引理 B.1 和 (30)，可得对任意 $t = 1, \ldots, T$

$$\mathcal{A}_{D,t} \leq \begin{cases} \frac{2\sqrt{L}}{\sqrt{|D|}}, & \text{if } \sigma_\ell = 0, \ell \geq L+1, \\ \frac{C_0 t^{s/2}}{\sqrt{|D|}} & \text{if } \sigma_\ell \leq c_0 \ell^{-1/s}, \\ \frac{C_1 \log^{1/4}|D|}{\sqrt{|D|}}, & \text{if } \sigma_\ell \leq c_1 e^{-c_2 \ell^2}. \end{cases} \tag{B14}$$

因此，由引理 1 和命题 2 可得

$$\mathcal{V}_{t,\beta,D} \leq C_1' \log\frac{8}{\delta} \begin{cases} \frac{\sqrt{L}}{\sqrt{|D|}}, & \text{if } \sigma_\ell = 0, \ell \geq L+1, \\ \frac{t^{s/2}}{\sqrt{|D|}} & \text{if } \sigma_\ell \leq c_0 \ell^{-1/s}, \\ \frac{\log^{1/4}|D|}{\sqrt{|D|}}, & \text{if } \sigma_\ell \leq c_1 e^{-c_2 \ell^2}, \end{cases}$$

其中 $C_1' := 2\sqrt{2}(\kappa M + \gamma)(1+\beta) \max\{2, C_0, C_1\}$. 由 (28) 和 (31) 可得

$$\mathcal{B}_{t,\beta,D} \leq C_2' \log^2\frac{16}{\delta} \begin{cases} (|D|/L)^{-1/2}, & \text{if } \sigma_\ell = 0, \ell \geq L+1, \\ |D|^{-r/(2r+s)} & \text{if } \sigma_\ell \leq c_0 \ell^{-1/s}, \\ (|D|/\sqrt{\log|D|})^{-1/2}, & \text{if } \sigma_\ell \leq c_1 e^{-c_2 \ell^2}, \end{cases}$$

其中 $C_2' := \max\{2^{r-1/2}, 1+4\kappa^2\}$。结合上述两个估计，由引理~1 和 (31) 可知，(32) 以置信度 $1-\delta$ 且 $C := C_1' + C_2'$ 成立。至此，推论 3 的证明完成。

■

证明 [命题 4] 的证明由 (A2) 和 (A3)，我们有

$$\begin{aligned} & \|f_{t+1,\beta,D} - f_{t,\beta,D}\|_D + t^{-1/2}\|f_{t+1,\beta,D} - f_{t,\beta,D}\|_K \\ \leq\ & 2\beta\|(L_{K,D}+t^{-1}I)^{1/2}(L_{K,D}g_{t,\beta}(L_{K,D})-I)S_D^\top y_D\|_K \\ \leq\ & 2\beta\|(L_{K,D}+t^{-1}I)^{1/2}(L_{K,D}g_{t,\beta}(L_{K,D})-I)(L_{K,D}f_\rho - S_D^\top y_D)\|_K \\ & +2\beta\|(L_{K,D}+t^{-1}I)^{1/2}(L_{K,D}g_{t,\beta}(L_{K,D})-I)L_{K,D}f_\rho\|_K. \end{aligned} \tag{B15}$$

但由 (A6) 可得

$$\begin{aligned} & \|(L_{K,D}+t^{-1}I)^{1/2}(L_{K,D}g_{t,\beta}(L_{K,D})-I)(L_{K,D}f_\rho - S_D^\top y_D)\|_K \\ \leq\ & \mathcal{P}_{D,t}\mathcal{Q}_{D,t}\|(L_{K,D}+t^{-1}I)(L_{K,D}g_{t,\beta}(L_{K,D})-I)\| \\ \leq\ & \mathcal{P}_{D,t}\mathcal{Q}_{D,t}\left(\|L_{K,D}(L_{K,D}g_{t,\beta}(L_{K,D})-I)\| + t^{-1}\|(L_{K,D}g_{t,\beta}(L_{K,D})-I)\|\right) \\ \leq\ & (1+1/\beta)t^{-1}\mathcal{P}_{D,t}\mathcal{Q}_{D,t}. \end{aligned} \tag{B16}$$

若 (16) 在 $\frac{1}{2} \le r \le 1$ 下成立，则由 (A6) 和 (B12) 可得

$$\begin{aligned}
& \|(L_{K,D}+t^{-1}I)^{1/2}(L_{K,D}g_{t,\beta}(L_{K,D})-I)L_{K,D}f_\rho\|_K \\
\le\ & \|(L_{K,D}+t^{-1}I)^{1/2}L_{K,D}(L_{K,D}g_{t,\beta}(L_{K,D})-I)L_K^{r-1/2}\|\|h_\rho\|_\rho \\
\le\ & \|h_\rho\|_\rho \mathcal{Q}_{D,t}^{2r-1}\|L_{K,D}(L_{K,D}g_{t,\beta}(L_{K,D})-I)(L_{K,D}+t^{-1}I)^r\| \\
\le\ & 2^r\|h_\rho\|_\rho \mathcal{Q}_{D,t}^{2r-1}\left(\|L_{K,D}^{r+1}(L_{K,D}g_{t,\beta}(L_{K,D})-I)\| + t^{-r}\|L_{K,D}(L_{K,D}g_{t,\beta}(L_{K,D})-I)\|\right) \\
\le\ & 2^r\|h_\rho\|_\rho \mathcal{Q}_{D,t}^{2r-1}\left(\left(\frac{r+1}{\beta}\right)^{r+1}+\frac{1}{\beta}\right)t^{-r-1}. \qquad \text{(B17)}
\end{aligned}$$

若 (16) 在 $r > 1$ 下成立，则由 (A6)、(B13) 和 (25) 可得

$$\begin{aligned}
& \|(L_{K,D}+t^{-1}I)^{1/2}(L_{K,D}g_{t,\beta}(L_{K,D})-I)L_{K,D}f_\rho\|_K \\
\le\ & \|(L_{K,D}+t^{-1}I)^{1/2}L_{K,D}(L_{K,D}g_{t,\beta}(L_{K,D})-I)L_K^{r-1/2}\|\|h_\rho\|_\rho \\
\le\ & \|(L_{K,D}+t^{-1}I)^{1/2}L_{K,D}(L_{K,D}g_{t,\beta}(L_{K,D})-I)L_{K,D}^{r-1/2}\|\|h_\rho\|_\rho \\
& +\|(L_{K,D}+t^{-1}I)^{1/2}L_{K,D}(L_{K,D}g_{t,\beta}(L_{K,D})-I)\times(L_K^{r-1/2}-L_{K,D}^{r-1/2})\|\|h_\rho\|_\rho \\
\le\ & \|h_\rho\|_\rho\left[\left(\frac{r+1}{\beta}\right)^{r+1}+\left(\frac{r+1/2}{\beta}\right)^{r+1/2}\right]t^{-r-1} \\
& +\|(L_{K,D}+t^{-1}I)^{1/2}L_{K,D}(L_{K,D}g_{t,\beta}(L_{K,D})-I)\|\|L_K^{r-1/2}-L_{K,D}^{r-1/2}\|_{HS}\|h_\rho\|_\rho \\
\le\ & \|h_\rho\|_\rho\left[\left(\frac{r+1}{\beta}\right)^{r+1}+\left(\frac{r+1/2}{\beta}\right)^{r+1/2}\right]t^{-r-1} \\
& +\max\{(r-1/2)\kappa^{2r-3},1\}\left[\left(\frac{3}{2\beta}\right)^{3/2}+\frac{1}{\beta}\right]t^{-3/2}\mathcal{R}_D^{\min\{1,r-1/2\}}. \qquad \text{(B18)}
\end{aligned}$$

将 (B18)、(B17) 和 (B16) 代入 (B15)，得到

$$\begin{aligned}
& \|f_{t+1,\beta,D}-f_{t,\beta,D}\|_D \le (2+2\beta)t^{-1}\mathcal{P}_{D,t}\mathcal{Q}_{D,t} \\
& +c_1\begin{cases} \mathcal{Q}_{D,t}^{2r-1}(t)^{-r-1}, & \text{if } \frac{1}{2}\le r\le 1, \\ t^{-r-1}+t^{-3/2}\mathcal{R}_D^{\min\{1,r-1/2\}}, & \text{if } r>1. \end{cases} \qquad \text{(B19)}
\end{aligned}$$

此外，若 (15) 成立且 $t \le T$，则由引理 A.3 和 A.4 可直接得出 (33) 以置信度 $1-\delta$ 成立。至此，命题 4 的证明完成。 ■

证明 [命题 5] 的证明由命题 4 可得

$$\mathcal{W}_{D,\hat{t}} \le C_2'\begin{cases} 2^{r-1/2}\hat{t}^{-r}, & \text{if } \frac{1}{2}\le r\le 1, \\ \hat{t}^{-r}+4\kappa^2\hat{t}^{-1/2}|D|^{-\min\{1/2,r/2-1/4\}}, & \text{if } r>1. \end{cases}$$

因此，由（35）可得 $\hat{t} \le t^*$。进而，根据引理 B.1 和引理 A.2，对于任意 $t \le T$，以 $1-\delta$ 的置信度，有

$$\begin{aligned}
& \frac{\sqrt{t}}{|D|}+\frac{\sqrt{\max\{\mathcal{N}_D(t^{-1}),1\}}(1+8\sqrt{t/|D|})}{\sqrt{|D|}} \\
\le\ & \left(\frac{\sqrt{t}}{|D|}+\frac{(1+4(1+t/|D|))h(t)(1+8\sqrt{t/|D|})}{\sqrt{|D|}}\right)\log^2\frac{16}{\delta},
\end{aligned}$$

其中

$$h(t) := \begin{cases} \sqrt{L}, & \text{if } \sigma_\ell = 0, \ell \geq L+1, \\ C_0 t^{s/2}, & \text{if } \sigma_\ell \leq c_0 \ell^{-1/s}, \\ \sqrt{C_1}\log^{1/4} t, & \text{if } \sigma_\ell \leq c_1' e^{-c_2 \ell^2}. \end{cases} \tag{B20}$$

因此，由 (6) 可知，以置信度 $1-\delta$，有

$$\mathcal{W}_{D,t} \leq c_2' \left(\frac{\sqrt{t}}{|D|} + \frac{81 h(t)}{\sqrt{|D|}} \right) \log^2 \frac{16}{\delta} =: \mathcal{T}_{D,t,\delta}.$$

其中 $c_2' := 4\sqrt{2}(\kappa M + \gamma)$。定义 $t^{**} := t^{**}_{D,\beta}$ 为满足下式的最大整数

$$\mathcal{T}_{D,t,\delta} \leq C_2' \begin{cases} 2^{r-1/2} t^{-r}, & \text{if } \frac{1}{2} \leq r \leq 1, \\ t^{-r} + 4\kappa^2 t^{-1/2} |D|^{-\min\{1/2, r/2-1/4\}}, & \text{if } r > 1. \end{cases} \tag{B21}$$

于是，由 (35) 可得

$$t^{**} \leq t^*.$$

显然，$\mathcal{T}_{D,t,\delta}$ 关于 t 单调不减，但 (B21) 的右端关于 t 单调不增。因此，对于任意 $t \leq t^{**}$，(B21) 均成立。令

$$t_0 = c_\delta \begin{cases} (|D|/L)^{\frac{1}{2r}}, & \text{if } \sigma_\ell = 0, \ell \geq L+1, \\ |D|^{\frac{1}{2r+s}}, & \text{if } \sigma_\ell \leq c_0 \ell^{-1/s}, \\ (|D|/\sqrt{\log|D|})^{\frac{1}{2r}}, & \text{if } \sigma_\ell \leq c_1 e^{-c_2 \ell^2}, \end{cases}$$

其中 $0 < c_\delta \leq 1$ 依赖于 δ，其具体取值将在下文确定，我们有

$$\mathcal{T}_{D,t_0,\delta} \leq c_2' H(|D|) \log^2 \frac{16}{\delta}$$

and

$$c_3' c_\delta^{-r} H(|D|) \leq \begin{cases} 2^{r-1/2} t_0^{-r}, & \text{if } \frac{1}{2} \leq r \leq 1, \\ t_0^{-r} + 4\kappa^2 t_0^{-1/2} |D|^{-\min\{1/2, r/2-1/4\}}, & \text{if } r > 1, \end{cases}$$

其中

$$H(|D|) := \begin{cases} (|D|/L)^{-1/2}, & \text{if } \sigma_\ell = 0, \ell \geq L+1, \\ |D|^{-r/(2r+s)} & \text{if } \sigma_\ell \leq c_0 \ell^{-1/s}, \\ (|D|/\sqrt{\log|D|})^{-1/2}, & \text{if } \sigma_\ell \leq c_1 e^{-c_2 \ell^2}, \end{cases} \tag{B22}$$

c_2' 和 c_3' 是仅依赖于 c_1'、 κ 和 c_0的常数， c_1， c_2。容易推导出存在 c_δ 使得（B21）成立。以 $\sigma_\ell \leq c_0 \ell^{-1/2}$ 为例，如果我们设定

$$c_\delta = \min\{1, c_3'/c_2'\}^{\frac{2}{2r+s}} \left(\log \frac{16}{\delta} \right)^{-\frac{4}{2r+s}} \leq 1,$$

then

$$c_2' |D|^{-\frac{r}{2r+s}} c_\delta^{s/2} \log^2 \frac{16}{\delta} \leq c_3' c_\delta^{-r} |D|^{-\frac{r}{2r+s}},$$

这意味着对于 t_0，(B21)成立。其余两种情况也可类似推导。为简洁起见，我们省略详细证明。因此，我们有 $t^* \geq t^{**} \geq t_0$。这

证明了(38)。随后，由命题 2、引理 1、 t^* 的定义以及(38)可得

$$\begin{aligned}
&\max\{\|f_{D,t^*}-f_\rho\|_\rho, \|f_{D,t^*}-f_\rho\|_D, (t^*)^{1/2}\|f_{D,t^*}-f_\rho\|_K\} \\
\leq\ & 2(1+\beta)\mathcal{W}_{D,t^*}\log^2\frac{16}{\delta} \\
+\ & C_1'\log^2\frac{16}{\delta}\begin{cases} 2^{r-1/2}t^{-r}, & \text{if } \frac{1}{2}\leq r\leq 1, \\ (t^*)^{-r}+4\kappa^2(t^*)^{-1/2}|D|^{-\min\{1/2,r/2-1/4\}}, & \text{if } r>1 \end{cases} \\
\leq\ & C_3'\log^2\frac{16}{\delta}\begin{cases} 2^{r-1/2}t_0^{-r}, & \text{if } \frac{1}{2}\leq r\leq 1, \\ t_0^{-r}+4\kappa^2 t_0^{-1/2}|D|^{-\min\{1/2,r/2-1/4\}}, & \text{if } r>1 \end{cases} \\
\leq\ & \tilde{C}'\log^2\frac{16}{\delta}\begin{cases} (|D|/L)^{-1/2}, & \text{if } \sigma_\ell=0, \ell\geq L+1, \\ |D|^{-r/(2r+s)} & \text{if } \sigma_\ell\leq c_0\ell^{-1/s}, \\ (|D|/\sqrt{\log|D|})^{-1/2}, & \text{if } \sigma_\ell\leq c_1 e^{-c_2\ell^2}, \end{cases}
\end{aligned}$$

其中 C_3' 和 C' 是与 $\delta|D|$ 或 t无关的常数。至此，命题5的证明完成。 ■

附录 C：定理 6 的证明

借助命题 2、命题 4 和命题5，我们可以如下证明定理 6。**证明** [定理 6] 的证明我们将证明分为两种情况：1) $\hat{t}<t_0$；2) $\hat{t}\geq t_0$，其中t_0 由式 (37) 指定。对于第一种情况，即 $\hat{t}<t_0$，由三角不等式可得

$$\|f_{\hat{t},\beta,D}-f_\rho\|_*\leq\|f_{\hat{t},\beta,D}-f_{t_0,\beta,D}\|_*+\|f_{t_0,\beta,D}-f_\rho\|_*, \tag{C23}$$

其中 $\|\cdot\|_*$ 表示 $\|\cdot\|_\rho t_0^{-1/2}\|\cdot\|_K$ 或 $\|\cdot\|_D$。但推论 3 表明

$$\begin{aligned}
&\max(\|f_{t_0,\beta,D}-f_\rho\|_\rho, \|f_{t_0,\beta,D}-f_\rho\|_D, t_0^{-\frac{1}{2}}\|f_{t_0,\beta,D}-f_\rho\|_K) \\
&\leq \tilde{C}\log^2\frac{16}{\delta}\begin{cases} (|D|/L)^{-1/2}, & \text{if } \sigma_\ell=0, \ell\geq L+1, \\ |D|^{-r/(2r+s)} & \text{if } \sigma_\ell\leq c_0\ell^{-1/s}, \\ (|D|/\sqrt{\log|D|})^{-1/2}, & \text{if } \sigma_\ell\leq c_1 e^{-c_2\ell^2}, \end{cases}
\end{aligned} \tag{C24}$$

以置信度 $1-\delta$成立。因此，只需对$\|f_{\hat{t}\beta D}-f_{t_0\beta D}\|_\rho$进行界定。再次利用三角不等式，可得

$$\|f_{\hat{t},\beta,D}-f_{t_0,\beta,D}\|_*\leq\sum_{k=\hat{t}}^{t_0-1}\|f_{k+1,\beta,D}-f_{k,\beta,D}\|_*.$$

对于任意 $k=\hat{t},\ldots,t_0-1$，引理 1 表明，以 $1-\delta$ 的置信度，有

$$\|f_{k+1,\beta,D}-f_{k,\beta,D}\|_*\leq 2(1+\beta)\bar{c}_1 k^{-1}\mathcal{W}_{D,k}\log^4\frac{16}{\delta}.$$

由于引理 B.1 和引理 A.2 得出，以 $1-\delta$ 的置信度，

$$\begin{aligned}&\frac{\sqrt{t}}{|D|}+\frac{\sqrt{\max\{\mathcal{N}_D(t^{-1}),1\}}(1+8\sqrt{t/|D|})}{\sqrt{|D|}}\\&\leq\left(\frac{\sqrt{t}}{|D|}+\frac{(1+4(1+t/|D|))h(t)(1+8\sqrt{t/|D|})}{\sqrt{|D|}}\right)\log^2\frac{16}{\delta},\end{aligned}$$

其中 $h(t)$由 (B20) 给出，因此，由 (6) 可知，以 $1-\delta$ 的置信度，有

$$\mathcal{W}_{D,t}\leq c_4'\left(\frac{\sqrt{t}}{|D|}+\frac{(1+4(1+t/|D|))h(t)(1+8\sqrt{t/|D|})}{\sqrt{|D|}}\right) \tag{C25}$$

其中 $c_4':=4\sqrt{2}(\kappa M+\gamma)$。因此，对于任意 $k\leq t_0\leq\tilde{c}|D|$，以 $1-\delta$ 的置信度，由 (B20) 可得

$$\mathcal{W}_{D,k}\leq\bar{c}_2\frac{h(t)}{\sqrt{|D|}}\left(1+\frac{k^{1/2}(h(k))^{-1}}{\sqrt{|D|}}\right)\leq 2\bar{c}_2\frac{h(t)}{\sqrt{|D|}},$$

其中 $\bar{c}_2:=c_4'(1+\tilde{c})(\sqrt{\tilde{c}+1}+(5+4\tilde{c})(C_0+1+\sqrt{C_1})(1+8\tilde{c}))^2$。若 $\sigma_\ell\leq c_0\ell^{-1/s}$，则对于任意 $k=\hat{t},\ \ldots,\ t_0-1,\ r\geq 1/2$ 可得

$$\begin{aligned}\|f_{t_0,\beta,D}-f_{\hat{t},\beta,D}\|_*&\leq 8\bar{c}_1\bar{c}_2(1+\beta)\log^4\frac{16}{\delta}\sum_{k=\hat{t}}^{t_0-1}k^{-1}\frac{h(k)}{\sqrt{|D|}}\\&\leq 8\bar{c}_1\bar{c}_2(1+\beta)(2s+1)\frac{t_0^{s/2}}{\sqrt{|D|}}\log^4\frac{16}{\delta}\\&\leq\bar{c}_3|D|^{-r/(2r+s)}\log^4\frac{16}{\delta}.\end{aligned}$$

其中 $\bar{c}_3:=4\bar{c}_1\bar{c}_2\tilde{c}^{s/2}(1+\beta)(2s+1)(1+\tilde{c}^{(1+s)/2})$。若 $\sigma_\ell=0,\ \ell\geq L+1$，结合 $h(k)=L$ 采用相同方法可得

$$\|f_{t_0,\beta,D}-f_{\hat{t},\beta,D}\|_*\leq\bar{c}_3\frac{\sqrt{L}\log|D|}{|D|}\log^4\frac{16}{\delta}.$$

类似地，对于 $\sigma_\ell\leq c_1e^{-c_2\ell^2}$，我们得到

$$\|f_{t_0,\beta,D}-f_{\hat{t},\beta,D}\|_*\leq\bar{c}_3\frac{\log|D|}{|D|}\log^4\frac{16}{\delta}.$$

将上述三个估计和 (C24) 代入 (C23)，我们以置信度 $1-\delta$ 得到

$$\|f_{\hat{t},\beta,D}-f_\rho\|_*\leq\bar{c}_4\log^4\frac{16}{\delta}\begin{cases}\frac{\sqrt{L}\log|D|}{\sqrt{|D|}}, & \text{if } \sigma_\ell=0,\ell\geq L+1,\\ |D|^{-r/(2r+s)} & \text{if } \sigma_\ell\leq c_0\ell^{-1/s},\\ \frac{\log|D|}{\sqrt{|D|}}, & \text{if } \sigma_\ell\leq c_1e^{-c_2\ell^2}\end{cases}$$

在 $\bar{c}_4 = \bar{c}_3 + \tilde{C}.$ 的条件下，现在我们转向第二种情况：$\hat{} \ t \geq t_0$。在这种情况下，根据 $\hat{}t$、 t^* 的定义以及命题5，我们得到

$$
\begin{aligned}
& \|f_{\hat{t},\beta,D} - f_\rho\|_\rho \\
\leq \ & 2(1+\beta)\mathcal{W}_{D,\hat{t}} \log^2 \frac{16}{\delta} + \log^2 \frac{16}{\delta} \begin{cases} 2^{r-1/2}\hat{t}^{-r}, & \text{if } \frac{1}{2} \leq r \leq 1, \\ \hat{t}^{-r} + 4\kappa^2 \hat{t}^{-1/2} |D|^{-\min\{1/2, r/2-1/4\}}, & \text{if } r > 1 \end{cases} \\
\leq \ & \bar{c}_8 \log^2 \frac{16}{\delta} \begin{cases} 2^{r-1/2}\hat{t}^{-r}, & \text{if } \frac{1}{2} \leq r \leq 1, \\ \hat{t}^{-r} + 4\kappa^2 \hat{t}^{-1/2} |D|^{-\min\{1/2, r/2-1/4\}}, & \text{if } r > 1 \end{cases} \\
\leq \ & \bar{c}_8 \log^2 \frac{16}{\delta} \begin{cases} 2^{r-1/2}t_0^{-r}, & \text{if } \frac{1}{2} \leq r \leq 1, \\ t_0^{-r} + 4\kappa^2 t_0^{-1/2} |D|^{-\min\{1/2, r/2-1/4\}}, & \text{if } r > 1 \end{cases} \\
\leq \ & \bar{c}_9 \log^2 \frac{16}{\delta} \begin{cases} (|D|/L)^{-1/2}, & \text{if } \sigma_\ell = 0, \ell \geq L+1, \\ |D|^{-r/(2r+s)} & \text{if } \sigma_\ell \leq c_0 \ell^{-1/s}, \\ (|D|/\sqrt{\log |D|})^{-1/2}, & \text{if } \sigma_\ell \leq c_1 e^{-c_2 \ell^2}, \end{cases}
\end{aligned}
$$

其中， $\bar{c}_8, \bar{c}_9$ 是与 $|D|, \delta$ 或 t无关的常数。$\|f_{\hat{t},\beta,D} - f_\rho\|_D$ 和$\|f_{\hat{t},\beta,D} - f_\rho\|_K$ 的界可以用相同的方法推导。为简洁起见，我们省略细节。定理 6 的证明至此完成。

■