一种分布无关的截断核岭回归估计量及相关谱分析

Asma Ben Saber[a] 和 Abderrazek Karoui[a1]

[a] 突尼斯迦太基大学比塞大科学学院数学系 .

arXiv:2301.07172v1 [math.ST] 17 Jan 2023

摘要— 众所周知，核岭回归（KRR）是一种广受欢迎的非参数回归估计方法。然而，面对规模为 $n \gg 1$,的大型数据集时，基于 KRR 的估计量存在计算负荷过大的缺点。近年来，人们提出了一些可扩展的 KRR 方法，旨在降低 KRR 的计算复杂度，同时保持其出色的收敛速率。本文研究一种用于非参数回归的新型可扩展 KRR 方法。所提出的截断核岭回归（TKRR）方法简单易行。它通过将与 Mercer 核 $\mathbb{K}$, 相关的完整 $n \times n$ 随机核或格拉姆矩阵 B_n, 替换为其主要 $n \times N$ 子矩阵 A_N, 来实现，其中通常 $N \ll n$. 。与一些现有的可扩展 KRR 估计量不同，TKRR 无需预处理步骤，例如奇异值分解（SVD）或利用某个随机矩阵进行草图处理。此外，我们证明，TKRR 适用于遵循未知概率律的 d–维随机抽样数据。为此，我们对与某一概率测度相关的紧核积分算子进行谱分析，该测度不同于通常概率测度。随后，我们将这一衰减估计推广至相关随机格拉姆矩阵的迹尾部衰减。我们预计，这一谱分析也有助于研究其他机器学习主题，例如基于核的无监督领域学习问题。本文特别致力于制定规则，以最优地选择相关的截断阶数 N 以及正则化参数的取值 $\lambda > 0$. 。所提出的规则基于与核 $\mathbb{K}$. 相关的正积分算子之谱的行为及其衰减率。这些参数的最优取值确保就经验风险误差而言，TKRR 与完整 KRR 估计量具有相同的最优收敛速率。最后，我们提供了一些数值模拟，以展示所提出的 TKRR 估计量的性能。

*202*数学主题分类： 主要分类 62G08、62G05；次要分类 62G07、62C2 0.

关键词： 非参数回归、缩放核岭回归、谱分析、随机格拉姆矩阵、再生核希尔伯特空间、特征值。

1 引言

对于整数 $d \geq 1$, 设 $\boldsymbol{X} \subset \mathbb{R}^d$ 为一个紧集。那么，我们考虑通常的 d–维非参数回归问题，其定义如下：

$$Y_i = f^*(X_i) + \varepsilon_i, \quad 1 \leq i \leq n. \tag{1}$$

这里， $\boldsymbol{Y} = [Y_i]_{1 \leq i \leq n}^T$ 是实值随机响应向量， f^* : $\boldsymbol{X} \to \mathbb{R}$ 是待估计的未知d–元回归函数。 $X_i \in \boldsymbol{X}$ 是 n 协变量。 ε_i 是方差为 $\mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$. 的 n 独立同分布的中心化随机变量。此外，我们假设 ε_i 相互独立

[1]通讯作者：AbderrazekKaroui (abderrazek.karoui@fsb.ucar.tn) 本研究部分得到 DGRST 研究资助 LR21ES10 和 PHC-Utique 研究项目 20G1503 的支持。

与 X_i. 独立。在文献中，有多种方案可为回归函数 f^*. 提供不同的估计量。核岭回归（KRR）是其中一种流行方案；在假设该函数可由其在相关再生核希尔伯特空间（RKHS）上的投影很好地近似的前提下，它能够为 f^* 提供相当好的估计。

近年来，人们越来越关注如何为核岭回归算法开发可扩展方法，其主要目标是：大幅降低核岭回归的计算时间复杂度，同时保持较高的收敛速率。对于现有的可扩展核岭回归估计量，我们简要介绍以下四种改进型核岭回归估计量。第一种估计量由 [2] 给出，简要描述如下。假设协变量 x_i 是确定性的，且回归函数 f^* 属于与半正定核 $\mathbb{K}(\cdot</style id='9'>, \cdot)</style id='12'>$定义在 $\boldsymbol{X} \times \boldsymbol{X}$. 上的再生核希尔伯特空间 $\mathcal{H},$ 。于是，作者提出了一种近似方案：不使用需要求逆正则化后的 $n \times n$ 核格拉姆矩阵的完整核岭回归估计量，而是用格拉姆矩阵的最佳秩 $r-$ 近似来替代该矩阵，其中 $1 \leq r \leq n.$ 为某个整数。具体做法是对主格拉姆矩阵进行奇异值分解，然后仅保留对角矩阵中主要的显著特征值 r，并将其余特征值 $n-r$ 置为零。第二种可扩展核岭回归估计量是 [25], 中给出的基于草图的估计量，另见 [3]。更具体地说，给定 n 个样本和一个正整数 $m < n,$ ，草图的思路是用该核矩阵的 $m-$ 维随机草图来替代完整核矩阵。投影维数 m 和草图矩阵的选择方式应确保核岭回归估计量达到最优收敛速率。特别是，草图维数 m 必须与核矩阵或格拉姆矩阵的统计维数成正比（相差对数因子）$\mathbf{K}$。回顾一下，对于给定的正则化参数 $\lambda > 0,$ ，核矩阵 $\mathbf{K}$ 的统计维数最常用的定义为

$$d_{\mathbf{K}}(\lambda) = \mathbf{Tr}\big((\mathbf{K} + \lambda I_n)^{-1}\mathbf{K}\big), \tag{2}$$

其中，$\mathbf{Tr}(A)$ 表示矩阵 $A.$ 的迹。第三种可扩展核岭回归估计量见 [20], 称为 FALKON。该算法利用随机投影降低传统核岭回归的计算复杂度和内存需求。具体而言，给定样本 $\{x_i, y_i\}_{i=1}^n$，仅使用从中均匀抽取的、包含 $M \ll n$ 个训练点的子集 $\{\tilde{x}_1 \cdots \tilde{x}_M\} \subseteq \{x_1 \cdots x_n\}$ 来构造估计量

$$\widehat{f}_{\lambda,M}(x) = \sum_{i=1}^{M} \tilde{\alpha}_i \mathbb{K}(x_i, \tilde{x}_i).$$

通常， $M = O(\sqrt{n})$ 足以使 FALKON 算法达到完整 KRR 的最优统计精度。第四种可扩展核岭回归方法，也是最早提出的方法，是由 [17] 提出的随机傅里叶特征（RFF）。RFF 要求正定核 $\mathbb{K}(\cdot, \cdot)$ 是卷积核或 Bochner 型平稳核。得益于这类核的结构，可以通过蒙特卡洛近似高效地近似该核，其中涉及 s 个傅里叶特征，其中 $s \ll n.$ 因此，大规模核矩阵 K 被替换为一个包含 s 个傅里叶特征的 $s \times s$ 半正定矩阵。近期， [28] 提出了一种将 RFF 扩展到更一般核函数的方法。下一节将更详细地介绍 RFF 方法。

需要指出的是，我们提出的 TKRR 与 RFF 有一些共同点。尽管如此，TKRR 并不要求正定核具有卷积型结构。它是通过用其 $n \times N$主子矩阵 $A_N = \left[\frac{1}{n}\mathbb{K}(X_i, X_j)\right]_{1 \leq i \leq n}$, 替换 $n \times n$ 随机核矩阵 $B_n = \left[\frac{1}{n}\mathbb{K}(X_i, X_j)\right]_{1 \leq i, j \leq n}$ 来实现的，其中 $N \ll n.$ 此外， n 协变量 X_i 是独立同分布的，且根据未知概率律 $\rho(\cdot).$ 抽取。我们假设测度 $d\rho$ 相对于与 Mercer 核相关联的通常概率测度 dP,绝对连续，

核 $\mathbb{K}(\cdot,\cdot)$. 更准确地说，我们的 TKRR 定义如下，

$$\widehat{f}_{N,\lambda}(x)=\sum_{j=1}^{N}\widehat{\omega}_j\frac{1}{n}\mathbb{K}(X_j,x),\qquad \widehat{\boldsymbol{\omega}}=\left[\widehat{\omega}_j\right]_{1\le j\le N}^{T}=\left(A_N^*A_N+\lambda I_N\right)^{-1}A_N^*\boldsymbol{Y}. \tag{3}$$

特别地，我们证明，截断阶数最优值的选择取决于希尔伯特-施密特积分算子 $T_{\mathbb{K}}$ 的谱的行为，该算子定义在 $L^2(\boldsymbol{X},\ dP)$。 $T_{\mathbb{K}}$ 的谱衰减率，以及这种衰减成立所对应的秩，对于选择最优截断阶数至关重要 我们假设，对于定义在 上的概率测度，算子 $T_{\mathbb{K}}$ 的谱性质是已知的。本文的第一个主要结果证明，在对未知采样测度 $d\rho(\cdot)$ 作出某些假设的条件下，完整随机核矩阵的特征值，以及截断 格拉姆矩阵的奇异值序列，与定义在 $(,\ dP)$ 上的算子 $T_{\mathbb{K}}$ 的特征值具有相似的衰减率。本文的第二个主要结果是，我们的 TKRR 估计量在求解非参数（NP）回归问题时的经验风险误差 L^2-。特别地，我们证明，所提出的 TKRR 估计量有三大优势：其经验风险误差 L^2- 与完整核岭回归算法的风险 L^2- 相似，速度远快于原有的基于核岭回归的算法，并且适用于一整类未知的采样概率分布。这使得我们提出的分布无关的截断核岭回归估计量非常适合用于驱动数据分析。

本文结构安排如下。第 2 节介绍经典核岭回归估计量以及基于随机傅里叶特征的可扩展核岭回归估计量的数学预备知识。随后，我们描述所提出的 TKRR 估计量。第 3 节证明一个颇为实用的界：对于与未知概率测度相关联的紧核积分算子，其特征值可由该算子与其通常概率测度相关联时的特征值加以界定。随后，我们利用这一结果估计与从未知概率密度中抽取的采样集相关联的随机核矩阵的特征值衰减率。第 4 节给出 TKRR 估计量的经验风险误差。随后，利用第 3 节的谱分析，在相关核积分算子的谱具有指数衰减或多项式衰减率的假设下，给出 TKRR 估计量的收敛速率。第 5 节介绍 Sinc 核和高斯核积分算子的一些谱性质。最后，我们通过数值模拟展示所提出的 TKRR 估计量的理论性质和性能。

2 再生核希尔伯特空间及相关核岭回归估计量的预备知识

本段首先介绍传统核岭回归估计量的一些预备知识。随后，我们介绍可扩展的随机傅里叶特征估计量（RFF）。本文提出的 TKRR 与 RFF 背后的某些思想相同。不过，这两种可扩展核岭回归估计量存在根本区别。RFF 基于对 Bochner 型卷积核的蒙特卡洛近似，而 TKRR 则基于对相当一般的核随机矩阵进行截断。最后，我们介绍本文提出的 TKRR 估计量。

问题 (1) 的核岭回归估计量简述如下。首先回顾一下，定义在 $\boldsymbol{X}\times\boldsymbol{X}$ 上的实值核 $\mathbb{K}(\cdot,\cdot)$ 若连续且半正定，则称为 Mercer 核。为简便起见，后文中可假设 $\mathbb{K}(\cdot,\cdot)$ 是正定的。此外，设 $d\rho$ 是支撑集为 $\boldsymbol{X}\subseteq\mathbb{R}^d$ 的概率测度，并设 $(\varphi_n)_{n\ge1},(\lambda_n)_{n\ge1}$ 为相应希尔伯特–施密特算子的正交归一特征函数及对应特征值 $T_{\mathbb{K}}$.

即

$$T_{\mathbb{K}}\varphi_n(x) = \int_{\boldsymbol{X}} \mathbb{K}(x,y)\varphi_n(y)\, d\rho(y) = \lambda_n\, \varphi_n(x), \quad x \in \boldsymbol{X}.$$

于是，根据 Mercer 定理，我们有： $\mathbb{K}(x,\ y) = \sum_{n=0}^{\infty} \lambda_n\, \varphi_n(x)\varphi_n(y)$，其中 $x,\ y \in \boldsymbol{X}$。该级数在紧集 $\boldsymbol{X} \times \boldsymbol{X}$. 上均匀收敛。此外，相应的再生核希尔伯特空间 $\mathcal{H}$ 定义如下：

$$\mathcal{H} = \left\{ f \in L^2(\boldsymbol{X}, d\rho),\ f = \sum_{n\geq 0} a_n(f)\varphi_n,\ \|f\|_{\mathcal{H}}^2 = \sum_{n\geq 1} \frac{|a_n(f)|^2}{\lambda_n} < +\infty \right\}. \tag{4}$$

给定正则化参数 $\lambda > 0,$ 核岭回归算法通过求解以下最小化问题来找到一个解 $f_\lambda \in \mathcal{H}$

$$\widehat{f}_K = \arg\min_{f\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \Big(f(X_i) - Y_i\Big)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}. \tag{5}$$

众所周知，例如参见 [22], ，式 (5) 的一个解由估计量给出 r

$$\widehat{f}_K(x) = \sum_{i=1}^{n} \widehat{c}_i \mathbb{K}(X_i, x), \tag{6}$$

$\widehat{\ }$其中，展开系数向量 $\mathbf{C} = (\widehat{c}_i)_{1\leq i\leq n}$ 是以下方程组的一个解

$$\left[\Big[\mathbb{K}(x_i,x_j)\Big]_{1\leq i,j\leq n} + n\lambda I_n \right] \widehat{\mathbf{C}} = G_\lambda\, \widehat{\mathbf{C}} = \mathbf{Y}, \quad \mathbf{Y} = (Y_i)_{1\leq i\leq n}^T. \tag{7}$$

这里， I_n 是单位矩阵，$\frac{1}{n}[\mathbb{K}(x_i,\ x_j)]_{1\leq i,\ j\leq n}$ 是与核 $\mathbb{K}(\cdot,\ \cdot)$。相关的随机格拉姆矩阵。该结果是著名表示定理的推论。这个定理特别有用，因为即使与最小化问题 (5) 相关的再生核希尔伯特空间是无限维的，解 (6) 也始终位于某个有限维空间中。此外，在由式 (7) 给出的正则化随机格拉姆矩阵 G_λ 可逆的条件下，展开系数向量$\widehat{\ }\mathbf{C}$ 由下式给出

$$\widehat{\mathbf{C}} = G_\lambda^{-1}\mathbf{Y} = \left[\Big[\mathbb{K}(x_i,x_j)\Big]_{1\leq i,j\leq n} + n\lambda I_n \right]^{-1} \mathbf{Y}, \quad \mathbf{Y} = (Y_i)_{1\leq i\leq n}^T. \tag{8}$$

如前所述，传统核岭回归估计量的优势在于收敛速率快。然而，其主要缺点是计算负荷大或时间复杂度高。为克服这一缺点，近期文献中提出了一些可扩展核岭回归估计量。这类核岭回归方法的先驱是 [17] 中提出的随机傅里叶特征（RFF）估计量。后者采用 Bochner 型卷积正定核。也就是说，这些核函数具有如下形式： $\mathbb{K}(x,\ y) = \kappa(x-y)\ x,\ y \in \boldsymbol{X}$，其中

$$\kappa(x-y) = \int_{\mathbb{R}^d} e^{-2i\pi\omega^T\cdot(x-y)}\, dP(\omega) = \int_{\mathbb{R}^d} e^{-2i\pi\omega^T\cdot x} e^{2i\pi\omega^T\cdot y}\, dP(\omega).$$

这里， $dP(\omega)$ 是一个概率测度。通常，核是实值的。因此，我们只需考虑该核的实部。RFF 的基本思想是用以下蒙特卡洛近似替代前述核

$$\widetilde{\mathbb{K}}(x,y) = \frac{1}{s}\sum_{j=1}^{s} Z_{\omega_j}^*(x) Z_{\omega_j}(y) = Z_{\boldsymbol{\omega}}^*(x) Z_{\boldsymbol{\omega}}(y), \quad Z_{\boldsymbol{\omega}}(x) = \frac{1}{\sqrt{s}}\Big(\cos(2\pi\omega_1^T\cdot x), \ldots, \cos(2\pi\omega_s^T\cdot x)\Big)^T.$$

这里， ω_i 是独立同分布的，并按照 $P(\cdot)$ 抽取， s 是傅里叶特征的数量，通常假设 $s \ll n.$ 。 $n \times n$ 核矩阵$\mathbf{K}$ 及其蒙特卡洛核矩阵$\widetilde{\ }$近似$\mathbf{K}$ 定义如下

$$\mathbf{K} = \Big[\kappa(X_i - X_j)\Big]_{1\le i,j\le n}, \quad \widetilde{\mathbf{K}} = \boldsymbol{Z}\boldsymbol{Z}^*, \quad \boldsymbol{Z} = \Big[Z_{\boldsymbol{\omega}}(X_1), \ldots, Z_{\boldsymbol{\omega}}(X_n)\Big]^T \in \mathbb{R}^{n\times s}. \tag{9}$$

注意，由于 $\mathbb{E}\big(\widetilde{\ }\mathbf{K}\big) = \mathbf{K}$，因此 $\widetilde{\ }\mathbf{K}$ 是核矩阵 $\mathbf{K}$ 的无偏近似。RFF 估计器定义为

$$\widetilde{f}_{s,\lambda}(x) = \sum_{j=1}^{s} \widetilde{\alpha}_j Z_{\omega_j}(x) = Z_{\boldsymbol{\omega}}^T \widetilde{\boldsymbol{\alpha}}, \qquad \widetilde{\boldsymbol{\alpha}} = \big(\boldsymbol{Z}^*\boldsymbol{Z} + s\lambda I_s\big)^{-1}\boldsymbol{Z}^*\boldsymbol{Y}. \tag{10}$$

也就是说， $\widetilde{\boldsymbol{\alpha}} \in \mathbb{R}^s$ 是该最小化问题的解

$$\widetilde{\boldsymbol{\alpha}} = \underset{\alpha\in\mathbb{R}^s}{\arg\min} \frac{1}{n}\big\|\boldsymbol{Y} - \boldsymbol{Z}\alpha\big\|_{\ell_{2,n}}^2 + s\lambda\big\|\alpha\big\|_{\ell_{2,n}}^2. \tag{11}$$

这里，$\|\cdot\|_{\ell_{2,\ n}}$ 是 $\mathbb{R}^n.$ 的通常欧几里得范数。更多详情请参见例如 [4]。近期，[28], 提出并研究了 RFF 的一种扩展。该方法将 RFF 方法扩展到如下形式的核

$$\mathbb{K}(x,y) = \int_{\mathcal{V}} z(\omega,x)z(\omega,y)\,dP(\omega), \tag{12}$$

其中 $z:\ \mathcal{V}\times\boldsymbol{X}\to\mathbb{R}$ 是一个连续有界函数。

接下来，我们将介绍如何推导所提出的 TKRR 估计量。对于两个正整数 $N \le n$，令 $S_X = \{X_1, \ldots, X_n\}$ 为一个随机采样集，其服从概率密度函数为 $P(\cdot)$，且支撑集为 $\boldsymbol{X}.$ 的概率律。设 $\mathbb{K}(\cdot,\ \cdot)$ 是一个连续且正定的 Mercer 核，使得定义在 $L^2(\boldsymbol{X},\ dP)$ 上的积分算子 $T_{\mathbb{K}}$ 满足

$$T_{\mathbb{K}}(f)(x) = \int_{\boldsymbol{X}} \mathbb{K}(x,y)f(y)dP(y) \tag{13}$$

是一个希尔伯特–施密特算子，其正特征值构成可数集，并按递减顺序排列

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge \cdots \ge 0.$$

特征值 λ_n 与一组标准正交的特征函数 $\{\varphi_n,\ n\in\mathbb{N}\}.$ 相关联。此外，在积分算子 $T_{\mathbb{K}}$ 为单射的假设下， φ_n 构成希尔伯特空间 $L^2(\boldsymbol{X},\ dP)$ 。回顾一下，Mercer 核具有如下谱分解

$$\mathbb{K}(x,y) = \sum_{m=1}^{\infty} \lambda_m\varphi_m(x)\varphi_m(y), \qquad \forall\, x,y\in\boldsymbol{X}. \tag{14}$$

下文中，我们采用以下记号:

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n), \quad A_N^*A_N = K_N^2, \quad \|\boldsymbol{u}\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2, \quad \boldsymbol{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n. \tag{15}$$

对于 $\boldsymbol{Y}\in\mathbb{R}^n,\ \ K_1\in\mathbb{R}^{n\times N},\ \ K_2\in\mathbb{R}^{N\times N}$ 两个正定矩阵以及 $\lambda>0$，我们旨在求得

$$\underset{\boldsymbol{\omega}\in\mathbb{R}^N}{\arg\min}\|K_1\boldsymbol{\omega} - \boldsymbol{Y}\|_n^2 + \lambda\omega^T\cdot K_2\omega = \underset{\boldsymbol{\omega}\in\mathbb{R}^N}{\arg\min} J_\lambda(\boldsymbol{\omega}) \tag{16}$$

利用 K_1 的线性性以及 $<\cdot,\ \cdot>$, $\mathbb{R}^n$ 的通常内积，并结合 $K_2^* = K_2$，这一事实，可得

$$\begin{aligned}\nabla J_\lambda(\boldsymbol{\omega})\cdot\boldsymbol{h} &= \frac{2}{n}<K_1\boldsymbol{\omega}-\boldsymbol{Y},K_1\boldsymbol{h}>+2\lambda<K_2\boldsymbol{\omega},\boldsymbol{h}>=\frac{2}{n}<K_1^*K_1\boldsymbol{\omega}-K_1^*\boldsymbol{Y},\boldsymbol{h}>+2\lambda<K_2\boldsymbol{\omega},\boldsymbol{h}>\\ &= 2<\frac{1}{n}K_1^*K_1\boldsymbol{\omega}-\frac{1}{n}K_1^*\boldsymbol{Y}+\lambda K_2\boldsymbol{\omega},\boldsymbol{h}>\end{aligned}$$

因此，当 $\left(\frac{1}{n}K_1^*K_1+\lambda K_2\right)\boldsymbol{\omega}=\frac{1}{n}K_1^*\boldsymbol{Y}$. 时， $\nabla J_\lambda(\boldsymbol{\omega})=0$。也就是说

$$\boldsymbol{\omega}=\left(K_1^*K_1+n\lambda K_2\right)^{-1}K_1^*\boldsymbol{Y}. \tag{17}$$

TKRR 估计量描述如下。给定紧集 $\boldsymbol{X}\subset\mathbb{R}^d$，定义在 $\boldsymbol{X}\times\boldsymbol{X}$ 上的正定 Mercer 核 $\mathbb{K}$，以及一组 n 个独立同分布采样点，这些采样点服从支撑集位于 $\boldsymbol{X}$，上的未知概率密度；我们考虑一个整数 $1\le N\le n$ 和矩形随机 Gram 矩阵

$$K_1=A_N=\left[\frac{1}{n}\mathbb{K}(X_i,X_j)\right]_{1\le i\le n,1\le j\le N} \tag{18}$$

注意， $K_1^*K_1\in\mathbb{R}^{N\times N}$ 是半正定的，并且我们可以控制 $K_1^*K_1$ 的特征值 $\lambda_i\left(K_1^*K_1\right)$。我们的 TKRR 估计量记作 $f_{N,\lambda}$，它由使用前述格拉姆矩阵 $K_1=A_N$ 和矩阵 $K_2=\frac{1}{n}I_N$. 的最小化问题（16）的解给出。因此，通过在正则化参数 $\lambda>0$，取适当值时使用（17）估计量 $f_{N,\lambda}$ 可由以下显式公式给出

$$\widehat{f}_{N,\lambda}(x)=\sum_{j=1}^{N}\widehat{\omega}_j\frac{1}{n}\mathbb{K}(X_j,x),\qquad \widehat{\boldsymbol{\omega}}=\left(A_N^*A_N+\lambda I_N\right)^{-1}A_N^*\boldsymbol{Y},\quad \boldsymbol{Y}=\left[Y_i\right]_{1\le i\le n}^T. \tag{19}$$

3 分布无关 Gram 矩阵的谱衰减

在本段中，我们首先研究一个重要问题：如何估计与相当一般且未知的采样概率律相关的随机格拉姆矩阵的衰减率。随后，我们将这一衰减率推广到矩形随机 Gram 矩阵的奇异值。最后得到的衰减率估计将用于截断 KRR 的经验风险误差。更重要的是，格拉姆矩阵特征值的衰减率问题，在许多用于机器学习和数据科学应用的基于核的算法中都特别有用。例如，在基于核的无监督领域学习问题中，源概率测度通常不同于目标概率测度。后者用于估计由输入预测输出 $y\in\mathbf{Y}\subset\mathbb{R}$ 的期望风险。 $\boldsymbol{x}\in\mathbb{R}^d$. 关于这一问题的更多详情，请参阅 [14] 及其中的参考文献。

通常，在使用核岭回归算法解决学习问题时，相关的格拉姆矩阵通常由一组遵循未知概率测度 $d\rho$ 的 d–维随机采样向量构造而成；该测度通常不同于相应核积分算子所关联的通常概率测度 dP，。例如，在一维情形下，对于正参数 $c>0$，Sinc核 $\mathbb{K}_c(x,\ y)=\sin(c(x-y))\,\pi(x-y)$ ' 定义在 $I=[-1,1]^2$ 上，与 I上的均匀测度相关联，该测度由 $dP(x)=\frac{1}{2}\mathbf{1}_{[-1,1]}(x)\,dx$. 给出。此外，对于两个正

实数 ξ， $c>0$，高斯核 $\mathbb{K}_G(x,\ y)=e^{-\xi|x-y|^2}$ 与其通常的高斯测度 $dP_c(x)=\sqrt{\dfrac{\pi}{2c}}e^{-2cx^2}$相关联。关于这两个核算子谱及其特征值的超指数衰减率的更多细节，读者可分别参阅 [8] 和 [18], 。

众所周知，格拉姆矩阵的谱性质与相应积分算子的谱性质相似，参见例如 [21, 8]。因此，我们首先关注与同一个连续 Mercer 核 $\mathbb{K}(\cdot,\cdot)$相关联、但分别作用于不同希尔伯特空间 $L^2(\boldsymbol{X},dP)$ 和 $L^2(\boldsymbol{X},d\rho)$ 上的两个自伴希尔伯特–施密特算子 $T_{\mathbb{K}}$ and$T_{\mathbb{K},\rho}$, 的谱之间的比较。这里， dP是最初考虑的概率测度， $d\rho$ 是支撑在紧集 $\boldsymbol{X}\subset\mathbb{R}^d$. 上的未知概率测度。此外，我们假设以下两个条件:

$\mathbf{H_0}$：算子 $T_{\mathbb{K}}$: $L^2(\boldsymbol{X},dP)\ \to L^2(\boldsymbol{X},dP)$ 是一一映射。

$\mathbf{H_1}$： $d\rho$ 关于 dP 绝对连续，且 $\dfrac{d\rho}{dP}$ 在 $\boldsymbol{X}$.上有界

令

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_n\geq\cdots\geq 0 \text{ and } \lambda_{1,\rho}\geq\lambda_{2,\rho}\geq\cdots\geq\lambda_{n,\rho}\geq\cdots\geq 0$$

表示 $T_{\mathbb{K}}$ and $T_{\mathbb{K},\rho}$, 的特征值，并按降序排列。此外，令 $\{\varphi_n,\,n\in\mathbb{N}\}$ 表示 $T_{\mathbb{K}}$. 的特征函数集合。注意，根据假设 $\mathbf{H_0}$ ，并结合核 $\mathbb{K}(\cdot,\cdot)$ 是连续 Mercer 核这一假设， φ_n 构成希尔伯特空间 $L^2(\boldsymbol{X},dP)$ 的标准正交基。在实际应用中，回归估计量需要处理 $d-$ 维数据样本。这些样本通常假设为独立同分布，并服从一个概率密度函数未知的概率律 $\rho(\cdot)$。给定前述连续 Mercer 核 $\mathbb{K}(\cdot,\cdot)$，及其相关联的自伴希尔伯特算子 $T_{\mathbb{K}}$（其谱性质已知），我们考虑以下一般情形：$n\times n$ 随机格拉姆矩阵

$$B_n=\frac{1}{n}\Big[\mathbb{K}(X_i,X_j)\Big]_{1\leq i,j\leq n},\tag{20}$$

其中， X_i 是根据 $\rho(\cdot)$抽取的独立同分布随机样本。以下定理为未知特征值$\lambda_{k,\ \rho}$ 以及随机矩阵 B_n. 的迹尾部期望值提供了有趣且实用的上界

定理 1. 在前述记号和假设$\mathbf{H_0}$ 与$\mathbf{H_1}$，下，我们有

$$\lambda_{m,\rho}\leq\sum_{j\geq m}\lambda_j\|\varphi_j\|_\rho^2,\quad\forall\, m\geq 1.\tag{21}$$

此外，对于任意整数 $1\leq k\leq n$，我们有

$$\mathcal{T}_k(B_n)=\mathbb{E}_Z\Big[\sum_{j\geq k}\lambda_j(B_n)\Big]\leq\sum_{j\geq k}\lambda_{j,\rho}\lesssim\lambda_k\|\varphi_k\|_\rho^2,\qquad whenever\ \sum_{i\geq k}\frac{(i-k+1)\lambda_i\|\varphi_i\|_\rho^2}{\lambda_k\|\varphi_k\|_\rho^2}<+\infty.$$

其中， $\|\varphi_j\|_\rho^2=\displaystyle\int_{\boldsymbol{X}}|\varphi_j^2(y)|\,d\rho(y)$。 (22)

证明： 我们首先注意到，假设 $\mathbf{H_1}$ 以及著名且强大的拉东–尼科迪姆定理保证了测度导数 $\dfrac{d\rho}{dP}$. 的存在。此外，对于任意函数 $f\in L^1(\boldsymbol{X},d\rho)$，我们有变量替换型公式，

$$\|f\|_\rho=\int_{\boldsymbol{X}}|f(y)|d\rho(y)=\int_{\boldsymbol{X}}|f(y)|\frac{d\rho(y)}{dP(y)}\,dP(y)=\int_{\boldsymbol{X}}|f(y)|\frac{\rho(y)}{P(y)}\,dP(y).$$

特别地，由于 Mercer 核 $\mathbb{K}(\cdot,\cdot)$ 连续，因此 $T_{\mathbb{K}}$ 的特征函数 φ_m 在紧集 $\boldsymbol{X}$. 上连续。因此，它们有界且属于 $L^2(\boldsymbol{X}, dP) \cap L^2(\boldsymbol{X}, d\rho)$。也就是说，对于任意整数 $m \geq 1.$， $\|\varphi_m\|_\rho$ 均有定义。此外，容易看出，函数 ψ_k: $x \to \varphi_k(x)\sqrt{\frac{P}{\rho}}(x)$ (x) 属于 $L^2(\boldsymbol{X}, d\rho)$。接下来，对于整数 $m \geq 1$, 考虑 $L^2(\boldsymbol{X}, d\rho)$中的两个$m-$维子空间，它们定义为

$$S_m = \mathrm{Span}\Big\{\varphi_k(x)\sqrt{\frac{P(x)}{\rho(x)}};\, 1 \leq k \leq m\Big\}, \quad V_m = \mathrm{Span}\Big\{\varphi_k(x);\, 1 \leq k \leq m\Big\}.$$

φ_k 在 $L^2(\boldsymbol{X}, d\rho)$中的线性无关性，是 $L^2(\boldsymbol{X}, dP)$中其正交性以及条件 $\mathrm{H_1}$. 的结果。更确切地说，若 $\sum_{k=1}^{m} a_k\varphi_k(x) = 0$，则 $\sum_{k=1}^{m} a_k\varphi_k(x)P(x)\,\rho(x) = 0.$ 因此，对于任意 $1 \leq j \leq m$, 有 $a_j =< \sum_{k=1}^{m} a_k\varphi_k(x)\, P(x)\,\rho(x)\ \varphi_j >_\rho = 0.$ 也就是说， φ_j 在 $L^2(\boldsymbol{X}, d\rho)$中也线性无关。接下来，我们证明 $S_m^\perp$ 和 $V_m^\perp$，即 $L^2(\boldsymbol{X}, d\rho)$中 S_m 和 V_m 的正交补，由下式给出

$$S_m^\perp = \mathrm{Span}\Big\{\varphi_k(x)\sqrt{\frac{P(x)}{\rho(x)}};\, k \geq m+1\Big\}, \quad V_m^\perp = \mathrm{Span}\Big\{\varphi_k(x)\frac{P(x)}{\rho(x)};\, k \geq m+1\Big\}.$$

显然，由变量替换型公式可得

$$\int_{\boldsymbol{X}} \varphi_j(x)\varphi_k(x)\frac{P(x)}{\rho(x)}d\rho(x) = \int_{\boldsymbol{X}} \varphi_j(x)\varphi_k(x)dP(x) = 0, \quad \forall j \leq m,\, k \geq m+1.$$

也就是说， $S_m^\perp$ 和 $V_m^\perp$ 中的任意元素分别与 S_m 和 V_m中的每个元素正交。接下来，假设存在 $\psi \in L^2(\boldsymbol{X}, d\rho)$与 S_m, 正交，则我们验证 $\psi \in S_m^\perp$。由于函数 $\psi(\cdot)\sqrt{\frac{\rho}{P}}(\cdot)$ $(\cdot)$属于 $L^2(\boldsymbol{X}, dP)$,, 因此对 $1 \leq k \leq m$,写出

$$0 \quad = \quad \int_{\boldsymbol{X}} \varphi_k(x)\sqrt{\frac{P(x)}{\rho(x)}}\psi(x)\,\rho(x)dx = \int_{\boldsymbol{X}} \varphi_k(x)\psi(x)\sqrt{\frac{\rho(x)}{P(x)}}\,P(x)dx,$$

由此可得， $\psi(\cdot)\sqrt{\frac{\rho}{P}}(\cdot)$ $(\cdot)$ 属于正交子空间 $E_m^\perp = \mathrm{Span}\Big\{\varphi_k(x)\ \ 1 \leq k \leq m\Big\}^\perp$.此外，由于 $\{\varphi_j,\, j \geq 1\}$ 是 $L^2(\boldsymbol{X}, dP)$ 的标准正交基，因此 $E_m^\perp = \mathrm{Span}\{\varphi_k(x)\ \ k \geq m+1\Big\}.$ 也就是说 $\psi(\cdot) \in S_m^\perp.$ 同理，我们可以验证，若 $\psi \in L^2(\boldsymbol{X}, d\rho)$ 与 V_m 正交，则 $\psi \in \mathrm{Span}\Big\{\varphi_k(x)\ {}^{P}(x)\,\rho(x);\ k \geq m+1\Big\}.$这直接源于以下事实:

$$0 = \int_{\boldsymbol{X}} \varphi_j(x)\psi(x)d\rho(x) = \int_{\boldsymbol{X}} \varphi_j(x)\sqrt{\frac{P(x)}{\rho(x)}}\psi(x)\sqrt{\frac{\rho(x)}{P(x)}}d\rho(x), \quad 1 \leq j \leq m \tag{23}$$

以及假设 $\mathbf{H_1}$ 的第二个陈述，这意味着函数 $\psi(\cdot)\sqrt{\frac{\rho}{P}}(\cdot)$ $(\cdot) \in L^2(\boldsymbol{X},\ d\rho)$。因此，由 (23) 可得， $\psi(\cdot)\sqrt{\frac{\rho}{P}}(\cdot)$ $(\cdot)$ $\in S_n^\perp$这意味着 $\psi(\cdot) \in V_m^\perp.$

接下来，我们利用作用于希尔伯特空间 $\mathcal{H}.$ 上的自伴紧算子 T 的正特征值极小极大刻画，这些特征值按降序排列。更确切地说，

$$\lambda_m = \min_{U_{m-1}} \max_{f \in U_{m-1}^\perp, \|f\|_{\mathcal{H}} \leq 1} < Tf, f >_{\mathcal{H}} .$$

这里， U_m 遍历 $\mathcal{H}$ 的 $m-$维子空间，且 $<\cdot,\ \cdot>_{\mathcal{H}}$ 是 $\mathcal{H}.$的内积。特别地，对于特殊情形 $\mathcal{H}=L^2$（$\boldsymbol{X}$，$d\rho$）和 $U_{m-1}=V_{m-1}$，，可得

$$\lambda_{m,\rho}\leq\max_{f\in V_{m-1}^{\perp},\|f\|_\rho\leq 1}<T_{\mathbb{K},\rho}f,f>_\rho.$$

由于 $f\in V_{m-1}^{\perp}$，因此 $f(x)=\sum_{k\geq m}\alpha_k\varphi_k(x)\ \overline{P(x)\ \rho(x)}$ 因此，利用 Mercer 核 $\mathbb{K}(\cdot,\ \cdot)$，的谱分解以及 φ_k在 $L^2(\boldsymbol{X},\ dP)$，中的正交归一性，可得对于$f\in V_{m-1}^{\perp}$,

$$\begin{aligned}
<T_{\mathbb{K},\rho}f,f> &= \int_{\boldsymbol{X}}\left(\int_{\boldsymbol{X}}\sum_{n\geq 1}\lambda_n\varphi_n(x)\varphi_n(y)\cdot\sum_{k\geq m}\alpha_k\varphi_k(y)\frac{P(y)}{\rho(y)}d\rho(y)\right)\sum_{j\geq m}\alpha_j\varphi_j(x)\frac{P(x)}{\rho(x)}d\rho(x)\\
&= \int_{\boldsymbol{X}}\left(\sum_{n\geq 1,k\geq m}\varphi_n(x)\int_{\boldsymbol{X}}\lambda_n\cdot\alpha_k\,\varphi_n(y)\varphi_k(y)dP(y)\right)\sum_{j\geq m}\alpha_j\varphi_j(x)\frac{P(x)}{\rho(x)}d\rho(x)\\
&= \int_{\boldsymbol{X}}\sum_{k\geq m}\lambda_k\cdot\alpha_k\,\varphi_k(x)\sum_{j\geq m}\alpha_j\varphi_j(x)dP(x)\\
&= \sum_{k\geq m}\lambda_k\cdot\alpha_k^2.
\end{aligned}$$

另一方面，根据 f 的前述展开以及柯西-施瓦茨不等式,

$$|\alpha_k|^2=|<f,\varphi_k>_\rho|^2\leq\|f\|_\rho^2\,\|\varphi_k\|_\rho^2.$$

因此,

$$\max_{f\in V_{m-1}^{\perp},\|f\|_\rho\leq 1}<T_{\mathbb{K},\rho}f,f>_\rho\leq\sum_{k\geq n}\lambda_k\cdot\|\varphi_k\|_\rho^2.$$

至此，(21) 的证明完成。为证明 (22)，第一个不等式见 [21], ，另见[8]。为证明第二个不等式，我们利用 (21) 得到

$$\mathbb{E}_Z\Big[\sum_{j\geq k}\lambda_j(B_n)\Big]\leq\sum_{j\geq k}\sum_{i\geq j}\lambda_i\|\varphi_i\|_\rho^2\leq\sum_{i\geq k}(i-k+1)\lambda_i\|\varphi_i\|_\rho^2\lesssim\lambda_k\|\varphi_k\|_\rho^2.$$

为验证前述第三个不等式，只需写出

$$\sum_{i\geq k}(i-k+1)\lambda_i\|\varphi_i\|_\rho^2=\lambda_k\|\varphi_k\|_\rho^2\cdot\sum_{i\geq k}\frac{(i-k+1)\lambda_i\|\varphi_i\|_\rho^2}{\lambda_k\|\varphi_k\|_\rho^2}\lesssim\lambda_k\|\varphi_k\|_\rho^2,$$

当 $\sum_{i\geq k}(i-k+1)\dfrac{\lambda_i\|\varphi_i\|_\rho^2}{\lambda_k\|\varphi_k\|_\rho^2}<+\infty.$时。类似地，可得前述第二个不等式，即 $\mathbb{E}_Z\Big[\sum_{j\geq k}\lambda_j(B_n)\Big]$.的上界

□

我们应该指出，核岭回归算法中使用的许多 Mercer 核的特征值集合都呈指数级衰减至零。也就是说， $\lambda_m\lesssim e^{-bm}$，对于某个 $b>0.$，其对应特征函数集合的幅值$\|\varphi_m\|_\infty$ φ_m 有一个 $O(m^{a/2})$ 量级的上界，对于某个$a\geq 0.$。对于未知概率测度 $d\rho$，我们假设它相对于原测度 dP绝对连续，并根据（21），对于任意 $m\geq 0$,

$$\lambda_{m+1,\rho}\lesssim\sum_{k\geq m+1}k^ae^{-bk}\lesssim\int_m^\infty t^ae^{-bt}\,dx=\frac{1}{b^{a+1}}\int_{bm}^\infty u^ae^{-u}\,du=\frac{1}{b^{a+1}}\Gamma(a+1,bm).\tag{24}$$

这里， $\Gamma(a+1,\ bm)$ 是上不完全伽马函数。众所周知，对于 $x > a-1$，我们有 $\Gamma(a,\ x) \le \dfrac{x^a e^{-x}}{x+1-a}$. 因此，给定 Mercer 核且其概率测度最终未知的此类积分算子的特征值 $d\rho$，具有以下上界，

$$\lambda_{m+1,\rho} \lesssim \frac{(bm)^{a+1}e^{-bm}}{bm-a} \lesssim m^a e^{-bm}.$$

此外，在 此情形下，对于由式 (20) 给出的格拉姆矩阵 B_n，根据式 (22)，其期望
其 tr 的 d 值
尾部有界于

$$\mathbb{E}_Z\Big[\sum_{j\ge m}\lambda_j(B_n)\Big] \lesssim m^a e^{-bm}. \tag{25}$$

接下来，对于正整数 $N \le m$，我们将式 (22) 中格拉姆矩阵 B_n 的特征值衰减率估计推广至 A_N的奇异值情形，即 B_n. 的主矩形 $m \times N$子矩阵。也就是说

$$A_N = \frac{1}{n}\Big[\mathbb{K}(X_i,X_j)\Big]_{1\le i\le n,\,1\le j\le N}. \tag{26}$$

我们回顾一下， X_i 是服从某个未知概率律的独立同分布随机样本，其概率密度函数为 $\rho(\cdot)$。为此，我们使用文献 [23] 中给出的关于子矩阵奇异值的以下交错不等式。

交错 奇异值不等式 [23]: 设 B 为一个奇异值为 s

$$\sigma_1(B) \ge \sigma_2(B) \ge \cdots \ge \sigma_{\min(m,n)}(B).$$

设 D 为 B的一个子矩阵，其奇异值为

$$\mu_1(D) \ge \mu_2(D) \ge \cdots \ge \mu_{\min(p,q)}(D).$$

于是，我们有

$$\sigma_i(B) \ge \mu_i(D) \ge \sigma_{i+(m-p)+(n-q)}(B), \qquad \forall\, 1 \le i \le \min(p+q-m, p+q-n). \tag{27}$$

利用式 (27) 中的第一个不等式和定理 2 中的不等式 (22)，可得以下推论。

推论 1. 设 B_n 和 A_N 分别为由(20) 和(26)给出的格拉姆矩阵及其主子矩阵。在定理 *2* 的假设和前述记号下，若 $\kappa_1 =$ 上确界 $\mathbb{K}(x,x),_{x\in \boldsymbol{X}}$
那么，对于任意 $1 \le m \le N$，我们有

$$\mathcal{T}_M = \mathbb{E}_Z\Big[\sum_{j\ge m}\mu_j(A_N)\Big] \lesssim \lambda_m\|\varphi_m\|_\rho^2 \tag{28}$$

and

$$\mathbb{E}_Z\Big[\sum_{j\ge m}\mu_j^2(A_N)\Big] \lesssim \frac{\kappa_1}{m}\lambda_m\|\varphi_m\|_\rho^2. \tag{29}$$

证明： 利用交错性质 (27)，并且由于格拉姆矩阵 B_n 是半正定的，即 $\sigma_i(B_n) = \lambda_i(B_n)$，因此我们得到

$$\mu_i(A_N) \le \lambda_i(B_n), \qquad \forall\, 1 \le i \le N. \tag{30}$$

利用前一个不等式和 (22)，可得

$$\mathbb{E}_Z\Big[\sum_{j\geq m}\lambda_j(B_n)\Big]\leq\sum_{j\geq m}\lambda_{j,\rho}\lesssim\lambda_m\|\varphi_m\|_\rho^2.$$

为验证 (29)，注意到由于 $\lambda_1(B_n)\geq\lambda_2(B_n)\geq\cdots\geq\lambda_n(B_n)\geq 0$，且由于

$$\mathbf{Tr}(B_n)=\sum_{i=1}^n\lambda_i(B_n)=\frac{1}{n}\sum_{i=1}^n\mathbb{K}(Z_i,Z_i)\leq\kappa_1,$$

则 $m\cdot\lambda_m(B_n)\leq\kappa_1$. 因此，利用 (30)，可得 $\mu_m(A_N)\leq\frac{\kappa_1}{m}$。由此可得

$$\mathbb{E}_Z\Big[\sum_{j\geq m}\mu_j^2(A_N)\Big]\leq\mathbb{E}_Z\Big[\lambda_m(B_n)\sum_{j\geq m}\mu_j(A_N)\Big]\leq\frac{\kappa_1}{m}\mathbb{E}_Z\Big[\sum_{j\geq m}\lambda_j(B_n)\Big]\lesssim\frac{\kappa_1}{m}\lambda_m\|\varphi_m\|_\rho^2.$$

4 TKRR估计量的经验风险误差分析与收敛速率

本段研究由 (19) 给出的 TKRR估计量 $f_{N,\lambda}$，在通常假设下的经验风险误差：真实回归函数 f^* 属于 $\mathcal{H}$，即与正定 Mercer 核 $\mathbb{K}(\cdot,\cdot)$。回顾一下，该经验风险误差定义为

$$\mathcal{R}(\widehat{f}_{N,\lambda})=\mathbb{E}\Big(\big\|\big(\widehat{f}_{N,\lambda}(X_i)-f^*(X_i)\big)_i\big\|_n^2\Big)=\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n\big(\widehat{f}_{N,\lambda}(X_i)-f^*(X_i)\big)^2\Big).\tag{31}$$

这里，$\mathbb{E}=\mathbb{E}_{X\times\varepsilon}=\mathbb{E}_X\mathbb{E}_\varepsilon$，分别表示关于独立随机变量 X_i 和 ε_i的期望之积。回顾一下，给定的独立同分布随机采样点 X_i 是根据 $\rho(\cdot)$抽取的。我们考虑再生核希尔伯特空间 $\mathcal{H}$的如下有限维子空间，其定义为

$$\mathcal{H}_N=\mathrm{Span}\Big\{\frac{1}{n}\mathbb{K}(X_j,\cdot),1\leq j\leq N\Big\},\qquad\mathcal{F}_N=\mathrm{Span}\Big\{\varphi_j(\cdot),1\leq j\leq N\Big\}.\tag{32}$$

我们假设 $\mathcal{H}$ 是 $L^2(\boldsymbol{X},dP)$，的子空间，且真实回归函数 $f^*\in\mathcal{H}$.。也就是说，$_\infty f^*=\sum a_j(f^*)\varphi_j$，$a_j(f^*)=<f^*,\varphi_j>_{\mathcal{H}}$.。此外，我们考虑 $_{j=1}f^*$ 在子空间 $\mathcal{H}_N$ 和 $\mathcal{F}_N$，上的如下正交投影：

$$\widetilde{\pi}_Nf^*(\cdot)=\sum_{j=1}^N\widetilde{\omega}_j\frac{1}{n}\mathbb{K}(X_j,\cdot),\qquad\pi_Nf^*(\cdot)=\sum_{j=1}^Na_j(f^*)\varphi_j(\cdot).\tag{33}$$

回顾一下，$\widetilde{\pi}_Nf^*$ 和 π_Nf^* 分别是由 $\mathcal{H}_N$ 和 $\mathcal{F}_N$，中的元素对 f^* 在 $L^2(dP)$– 范数下所作的最佳逼近。下文中，我们采用以下两个技术假设：
$\mathbf{H_2}$：随机投影矩阵 $\Big[\varphi_j(X_i)\Big]_{1\leq i\leq n,\ 1\leq j\leq N}$ 是满秩的。接下来，我们考
虑截断投影核，

$$\mathbb{K}_N(x,y)=\sum_{m=1}^N\lambda_m\varphi_m(x)\varphi_m(y),\quad x,y\in\boldsymbol{X}.\tag{34}$$

容易看出，在条件 $\mathbf{H_2}$,下，投影 $\pi_N f^*$具有以下第二种等价展开形式

$$\pi_N f^*(x) = \sum_{j=1}^{N} \omega'_j \frac{1}{n} \mathbb{K}_N(X_j, x), \quad x \in \boldsymbol{X}. \tag{35}$$

最后，为了研究我们的 TKRR估计量的经验风险误差，我们还需要对由 (33) 和 (35) 给出的投影 $\widetilde{\pi}_N f^*$ 和 $\pi_N f^*$的系数作出以下假设:

$\mathbf{H_3}$: $\max\left(\left\|\widetilde{\boldsymbol{\omega}}\right\|_n^2, \left\|\boldsymbol{\omega}'\right\|_n^2\right) \leq 1, \quad$ where $\quad \left\|\widetilde{\boldsymbol{\omega}}\right\|_n^2 = \frac{1}{n}\sum_{j=1}^{N} \widetilde{\omega}_j^2, \quad \left\|\boldsymbol{\omega}'\right\|_n^2 = \frac{1}{n}\sum_{j=1}^{N} \omega'^2_j.$

注意，将上界 1 替换为任意常数$C > 0.$，即可放宽前述假设。以下技术引理将在证明我们的 TKRR 经验风险误差时用到。它表明，当 $f^* \in \mathcal{H}.$时， $\widetilde{\pi}_N f^*$ 可由 $\pi_N f^*$很好地近似

引理 1。　在前述记号和假设$\mathbf{H_1}$, $\mathbf{H_2}$ 以及$\mathbf{H_3}$, 下，令 $f \in \mathcal{H}$，则有

$$\mathbb{E}\Big(\left\|\left((\pi_N f - \widetilde{\pi}_N f)(X_i)\right)_i\right\|_n^2\Big) \lesssim \frac{1}{N}\sum_{k\geq N+1} \lambda_k \|\varphi_k\|_\infty^2 + \lambda_{N+1}\|f\|_{\mathcal{H}}^2. \tag{36}$$

证明: 众所周知，若 $f \in \mathcal{H}$，则有

$$\begin{aligned} \|f - \pi_N f\|_P^2 &= \sum_{k\geq N+1} |a_k(f)|^2 = \sum_{k\geq N+1} \lambda_k \frac{|a_k(f)|^2}{\lambda_k} \\ &\leq \lambda_{N+1} \sum_{k\geq N+1} \frac{|a_k(f)|^2}{\lambda_k} \leq \lambda_{N+1}\|f\|_{\mathcal{H}}^2. \end{aligned}$$

这里， $a_k(f) =< f, \varphi_k >_P$，其中 $< \cdot, \cdot >_P$ 是 $L^2(\boldsymbol{X}, dP)$ 的通常内积。此外，利用$\mathbf{H_1}$, , 对于任意 $g \in \mathcal{H}$,

$$\|g\|_\rho^2 = \int_{\boldsymbol{X}} |g(x)|^2 \, d\rho(x) \lesssim \int_{\boldsymbol{X}} |g(x)|^2 \, dP(x) = \|g\|_{L^2(dP)}^2. \tag{37}$$

另一方面，对于 $1 \leq i \leq n$，我们有

$$\mathbb{E}\Big(\big((\pi_N f - \widetilde{\pi}_N f)(X_i)\big)^2\Big) = \int_{\boldsymbol{X}} |\widetilde{\pi}_N(f)(x) - \pi_N f(x)|^2 \, d\rho(x) \lesssim \|\widetilde{\pi}_N(f) - \pi_N(f)\|_{L^2(dP)}^2.$$

此外，由于 $\widetilde{\pi}_N$ 是一个投影算子，因此我们有

$$\left\|\widetilde{\pi}_N\big(f - \pi_N f\big)\right\|_{L^2(dP)} \leq \|f - \pi_N f\|_{L^2(dP)} \leq \sqrt{\lambda_{N+1}}\|f\|_{\mathcal{H}}. \tag{38}$$

此外，由于 $\widetilde{\pi}_N(\pi_N f)$ 是 $\pi_N f$在子空间 $\mathcal{H}_N$中的元素所得到的最佳逼近，并且由于非截断核展开 $\sum_{j=1}^{N} \omega'_j \frac{1}{n}\mathbb{K}(X_j, x) \in \mathcal{H}_N$，因此我们有 _

$$\left\|\widetilde{\pi}_N(\pi_N f) - \pi_N f\right\|_{L^2(dP)}^2 \leq \frac{1}{n^2}\Big\|\sum_{j=1}^{N} \omega'_j \sum_{k\geq N+1} \lambda_k \varphi_k(X_j)\varphi_k(\cdot)\Big\|_{L^2(dP)}^2 \leq \Big(\frac{1}{n}\sum_{k\geq N+1} \lambda_k \|\varphi_k\|_\infty^2\Big)\|\boldsymbol{\omega}'\|_n^2,$$

最后，由于

$$\|\widetilde{\pi}_N f - \pi_N f\|_{L^2(dP)} \leq \|\widetilde{\pi}_N(f - \pi_N f)\|_{L^2(dP)} + \|\widetilde{\pi}_N(\pi_N f) - \pi_N f\|_{L^2(dP)},$$

因此，利用前面的不等式，即可得到所需结果（36）。

下面的定理给出了我们的 TKRR估计量的经验风险误差上界。

定理 2. 令 n、$N>0$ 为两个正整数，且 $\lambda>0$. 在沿用前述记号并满足假设 $\mathbf{H_1}$, $\mathbf{H_2}$ 和 $\mathbf{H_3}$, 的条件下，对于 $f^*\in\mathcal{H}$,

$$\mathbb{E}\Big(\big\|(\widehat{f}_{N,\lambda}-f^*)(X_i)\big\|_n^2\Big) \leq \frac{\lambda}{2}+2\frac{\sigma^2}{n}\sum_{j=1}^{N}\mathbb{E}_X\Big[\Big(\frac{\mu_j^2}{\mu_j^2+\lambda}\Big)^2\Big] + C_1\Big(\lambda_{N+1}\|f^*\|_{\mathcal{H}}^2+\frac{1}{N}\sum_{k\geq N+1}\lambda_k\|\varphi_k\|_\infty^2\Big). \quad (39)$$

这里，$C_1>0$ 是一致正常数。

证明： 令 $\widetilde{\pi}_N$ 和 π_N 为由 (33) 定义的投影算子，则有

$$f^*(X_j)=\widetilde{\pi}_N f^*(X_j)+\Big(f^*(X_j)-\pi_N f^*(X_j)\Big)+\Big(\pi_N f^*(X_j)-\widetilde{\pi}_N f^*(X_j)\Big),\quad 1\leq j\leq n. \quad (40)$$

因此，由 (19) 可得

$$\Big[\widehat{f}_{N,\lambda}(X_i)\Big]_{1\leq i\leq n}^T=A_N\big(A_N^*A_N+\lambda I_N\big)^{-1}A_N^*\Big[f^*(X_j)+\varepsilon_j\Big]_{1\leq j\leq n}^T. \quad (41)$$

结合前面的两个恒等式，可得 $F_N=\big(\widetilde{\pi}_N f^*(X_i)\big)_{1\leq i\leq n}$,

$$\Big[\widehat{f}_{N,\lambda}(X_i)\Big]_{1\leq i\leq n}^T=A_N\big(A_N^*A_N+\lambda I_N\big)^{-1}A_N^*\Big[\big(\widetilde{\pi}_N f^*(X_j)+\varepsilon_j\big)+ +\big(f^*(X_j)-\pi_N f^*(X_j)\big)+\big(\pi_N f^*(X_j)-\widetilde{\pi}_N f^*(X_j)\big)\Big]_{1\leq j\leq n}^T. \quad (42)$$

令 $\|A\|_2$ 表示矩阵 $2-$ 范数。利用（40）和（42），以及 $\big\|A_N\big(A_N^*A_N+\lambda I_N\big)^{-1}A_N^*\big\|_2\leq 1$，这一事实，可得

$$\begin{aligned}\big\|\big(\widehat{f}_{N,\lambda}(X_i)-f^*(X_i)\big)_i\big\|_n &\leq \big\|A_N(K_N^2+\lambda I_N)^{-1}A_N^*(F_N+\boldsymbol{\varepsilon})^T-F_N^T\|_n \\ &+ 2\Big(\big\|\big(f^*(X_i)-\pi_N f(X_i)\big)_i\big\|_n+\big\|\big(\pi_N f^*(X_i)-\widetilde{\pi}_N f(X_i)\big)_i\big\|_n\Big) \quad (43)\end{aligned}$$

由于 $f^*\in\mathcal{H}$，因此按照引理 1 的证明，我们有

$$\begin{aligned}\mathbb{E}\left(\big\|\big(f^*(X_i)-\pi_N f(X_i)\big)_i\big\|_n^2\right) &= \frac{1}{n}\sum_{i=1}^n\int_{\boldsymbol{X}}|f^*(x)-\pi_N f^*(x)|^2d\rho(x) \\ &\lesssim \int_{\boldsymbol{X}}|f^*(x)-\pi_N f^*(x)|^2dP(x)=\|f^*-\pi_N f^*\|_{L^2(dP)}^2 \\ &\lesssim \lambda_{N+1}\|f^*\|_{\mathcal{H}}^2. \quad (44)\end{aligned}$$

此外，由引理 1 可得

$$\mathbb{E}\Big(\big\|\big((\pi_N f^*-\widetilde{\pi}_N f^*)(X_i)\big)_i\big\|_n^2\Big)\lesssim\frac{1}{N}\sum_{k\geq N+1}\lambda_k\|\varphi_k\|_\infty^2+\lambda_{N+1}\|f^*\|_{\mathcal{H}}^2. \quad (45)$$

接下来，我们估计经验风险误差中的以下主要量:

$$R_{n,N} = \mathbb{E}\left(\left\|\left(A_N\left(A_N^*A_N + \lambda I_N\right)^{-1}A_N^*\left[\left(\widetilde{\pi}_N f^*(X_j) + \varepsilon_j\right)_j\right] - \widetilde{\pi}_N f^*(X_i)\right)_i\right\|_n^2\right).$$

通过采用一些与 [2], 主定理证明中类似的技巧，但在当前的随机采样情形下，可以得到 $A_N = U\Sigma_{n,N}V^*$的如下奇异值分解，其中 U, V是正交矩阵，而$\Sigma_{n,N}$是一个 $n \times N$矩形对角矩阵。因此，我们有

$$K_N^2 = A_N^*A_N = \left(U\Sigma_{n,N}V^*\right)^*\left(U\Sigma_{n,N}V^*\right) = V\Sigma_{n,N}^*\Sigma_{n,N}V^*.$$

令$D_N = \Sigma_{n,N}^*\Sigma_{n,N}$. 则,

$$(K_N^2 + \lambda I_N)^{-1} = \big(V(D_N + \lambda I_N)\big)V^*)^{-1} = V(D_N + \lambda I_N)^{-1}V^*$$

and

$$\begin{aligned} A_N(K_N^2 + \lambda I_N)^{-1}A_N^* &= U\Sigma_{n,N}V^*V(D_N + \lambda I_N)^{-1}V^*V\Sigma_{n,N}^*U^* \\ &= U\Sigma_{n,N}(D_N + \lambda I_N)^{-1}\Sigma_{n,N}^*U^*. \end{aligned}$$

从而

$$\begin{aligned} \left\|A_N(K_N^2 + \lambda I_N)^{-1}A_N^*(F_N + \boldsymbol{\varepsilon})^T - F_N^T\|_n\right. &\leq \|U\Sigma_{n,N}(D_N + \lambda I_N)^{-1}\Sigma_{n,N}^*U^*F_N^T - F_N^T\|_n \\ &\quad + \|U\Sigma_{n,N}(D_N + \lambda I_N)^{-1}\Sigma_{n,N}^*U^*\boldsymbol{\varepsilon}^T\|_n. \end{aligned} \tag{46}$$

接下来，令$\|\cdot\|_{\ell_2,n}$ 和 $\|\cdot\|_2$ 分别表示 $\mathbb{R}^n$ 及一个矩形矩阵的通常 2–范数。
特别地，对于~~N~~个 $n \times N$ 矩阵 A， $\|A\|_2 = \sigma_1(A)$， A. 的最大奇异值。此外，由于 $\widetilde{\pi}_N f^*(X_i) = \sum_{j=1}^{n} \widetilde{\omega}_j \mathbb{K}(X_i, X_j)$ $1 \leq i \leq n$,那么我们有

$$F_N^T = \left(\widetilde{\pi}_N f^*(X_i)\right)_{1\leq i\leq n}^T = A_N\,\widetilde{\boldsymbol{\omega}} = U\Sigma_{n,N}V^*\,\widetilde{\boldsymbol{\omega}}, \qquad \widetilde{\boldsymbol{\omega}} = (\widetilde{\omega}_1, \ldots, \widetilde{\omega}_N)^T.$$

此外，由于 2– 范数在乘以正交矩阵后保持不变，因此可得

$$\begin{aligned} \left\|U\Sigma_{n,N}(D_N + \lambda I_N)^{-1}\Sigma_{n,N}^*U^*F_N^T - F_N^T\right\|_n^2 &= \frac{1}{n}\left\|U\big(\Sigma_{n,N}(D_N + \lambda I_N)^{-1}\Sigma_{n,N}^* - I_n\big)\Sigma_{n,N}V^*\widetilde{\boldsymbol{\omega}}\right\|_{\ell_2,n}^2 \\ &= \frac{1}{n}\left\|\big(\Sigma_{n,N}(D_N + \lambda I_N)^{-1}\Sigma_{n,N}^* - I_n\big)\Sigma_{n,N}V^*\widetilde{\boldsymbol{\omega}}\right\|_{\ell_2,n}^2 \\ &\leq \left\|\big(\Sigma_{n,N}(D_N + \lambda I_N)^{-1}\Sigma_{n,N}^* - I_n\big)\Sigma_{n,N}\right\|_2^2 \cdot \frac{1}{n}\|V^*\widetilde{\boldsymbol{\omega}}\|_{\ell_2,n}^2 \\ &\leq \left\|\big(\Sigma_{n,N}(D_N + \lambda I_N)^{-1}\Sigma_{n,N}^* - I_n\big)\Sigma_{n,N}\right\|_2^2 \cdot \|\widetilde{\boldsymbol{\omega}}\|_n^2. \end{aligned} \tag{47}$$

令 $\phi_n^\lambda = \Sigma_{n,\ N}(D_N + \lambda I_N)^{-1}\Sigma_{n,\ N}^* - I_n$，则有

$$\left\|\big(\Sigma_{n,N}(D_N + \lambda I_N)^{-1}\Sigma_{n,N}^* - I_n\big)\Sigma_{n,N}\right\|_2^2 = \left|\sigma_1\big(\phi_n^\lambda\Sigma_{n,N}\big)\right|^2. \tag{48}$$

另一方面，我们有 $\Sigma_{n,N} = \left[s_{i,j}\right]_{1\leq i\leq n}$， $s_{i,j} = \mu_j\delta_{i,j}$， 其中 $\mu_j = \sqrt{\lambda_j(A_N^*A_N)}$ 是$_{1\leq j\leq N}$ ~~A_N.~~ 的奇异值。因此，我们有

$$\phi_{n,N}^\lambda\Sigma_{n,N} = \left[\gamma_{i,j}\right]_{\substack{1\leq i\leq n\\ 1\leq j\leq N}}, \quad \gamma_{i,j} = \Big(\frac{\mu_j^2}{\mu_j^2 + \lambda} - 1\Big)\mu_j\delta_{i,j} = \frac{-\lambda\mu_j}{\mu_j^2 + \lambda}\delta_{i,j}.$$

Hence, $|\sigma_1(\phi_n^\lambda \Sigma_{n,N})|^2 = \max\limits_{1\le j\le N}\big(\frac{\lambda}{\mu_j^2+\lambda}\big)^2\mu_j^2$. Since $\sup\limits_{x\ge 0}\frac{\lambda x}{x^2+\lambda} = \frac{\sqrt{\lambda}}{2}$, then $\max\limits_{1\le j\le N}\Big(\frac{\lambda}{\mu_j^2+\lambda}\Big)^2\mu_j^2 = \frac{\lambda}{4}$. That is

$$\big|\sigma_1(\phi_n^\lambda \Sigma_{n,N})\big|^2 \le \frac{\lambda}{4}. \tag{49}$$

因此，利用 (47)–(49) 以及假设 $\mathbf{H_3}$, ，可得

$$\mathbb{E}\left(\big\|U\Sigma_{n,N}(D_N+\lambda I_N)^{-1}\Sigma_{n,N}^* U^* F_N^T - F_N^T\big\|_n^2\right) \le \frac{\lambda}{4}. \tag{50}$$

接下来，为了界定（46）右侧第二项的期望，我们首先注意到，若 $B_n = \big[b_{ij}\big]_{1\le i,\ j\le n}$是一个实系数 $n\times n$矩阵，则其弗罗贝尼乌斯范数 $\|B_n\|_F$为

$$\|B_n\|_F^2 = \sum_{i,j=1}^n (b_{ij})^2 = \mathbf{Tr}(B_n^* B_n) = \sum_{i=1}^n \sigma_i^2(B_n),$$

其中， $\sigma_i(B_n)$ 是 B_n. 的奇异值。此外，弗罗贝尼乌斯范数在左乘或右乘正交矩阵时保持不变。我们考虑如下特殊取值:

$$B_n = U\Sigma_{n,N}(D_N+\lambda I_N)^{-1}\Sigma_{n,N}^* U^*. \tag{51}$$

在这种情况下， B_n 的秩至多为 N，并且

$$\|B_n\|_F^2 = \sum_{i,j=1}^n b_{ij}^2 = \sum_{j=1}^N \Big(\frac{\mu_j^2}{\mu_j^2+\lambda}\Big)^2.$$

此外，利用 ε_i 独立同分布且与 X_i 相互独立这一事实，其中 $\mathbb{E}_\varepsilon(\varepsilon_i) = 0$ 且$\mathbb{E}_\varepsilon(\varepsilon_i^2) = \sigma^2$，可以很容易地验证

$$\begin{aligned}\frac{1}{n}\mathbb{E}_{X\times\varepsilon}\left(\big\|B_n\boldsymbol{\varepsilon}^T\big\|_2^2\right) &= \frac{1}{n}\mathbb{E}_X\mathbb{E}_\varepsilon\sum_{i=1}^n\left(\sum_{j=1}^n b_{ij}\varepsilon_j\right)^2 \\ &= \frac{1}{n}\mathbb{E}_X\left[\sum_{i,j=1}^n b_{ij}^2\mathbb{E}_\varepsilon(\varepsilon_j^2)\right] = \frac{\sigma^2}{n}\mathbb{E}_X\left[\sum_{j=1}^N\Big(\frac{\mu_j^2}{\mu_j^2+\lambda}\Big)^2\right].\end{aligned} \tag{52}$$

将（43）两边平方，并利用不等式（$a+b$）$^2 \le 2$（a^2+b^2），其中 a, $b\in\mathbb{R}$，然后取期望，可得

$$\begin{aligned}\mathbb{E}\left(\big\|\big(\widetilde{f}_{N,\lambda}(X_i) - f^*(X_i)\big)_i\big\|_n^2\right) \le\ & 2\,\mathbb{E}\left(\big\|A_N(K_N^2+\lambda I_N)^{-1}A_N^* F_N^T - F_N^T\|_n^2\right) \\ & +2\,\mathbb{E}\left(\big\|A_N(K_N^2+\lambda I_N)^{-1}A_N^*\boldsymbol{\varepsilon}^T\|_n^2\right) + 4\,\mathbb{E}\left(\big\|\big(f^*(X_i)-\pi_N f^*(X_i)\big)_i\big\|_n^2\right) \\ & +4\,\mathbb{E}\left(\big\|\big(\pi_N f^*(X_i)-\widetilde{\pi}_N f^*(X_i)\big)_i\big\|_n^2\right).\end{aligned} \tag{53}$$

最后，将前述不等式与式（44）、（45）、（46）、（50）和（52）结合，即可得到所需结果（39）。

注 1. 尽管我们的 *TKRR* 估计量的风险误差(39) 与 *RFF* 估计器的风险误差有一些相似之处（例如见 *[4],* ），但这两种风险及其证明存在根本区别。更具体地说，风险误差(39) 涉及积分算子 $T_{\mathbb{K}}$ 的特征值，以及该算子的加权尾迹。此外，它适用于相当一般的随机核矩阵，且随机采样集未知。 *RFF* 估计器的经验风险误差涉及核矩阵$\mathbf{K}$的特征值。 *[2]*提出的谱截断核岭回归估计器的经验风险误差也有类似的说明。

注 2. 条件 $f^* \in \mathcal{H}$ 可以放宽。我们可以假设再生核希尔伯特空间 $\mathcal{H}$ 是 $L^2(\boldsymbol{X}, dP)$ 的一个闭子空间，并且回归函数 f^* 属于 $L^2(\boldsymbol{X}, dP)$的另一个子空间 $\mathcal{F}$。令 $\pi_{\mathcal{H}}$ 为 $\mathcal{H}$ 上的正交投影算子，并令 $g^* = \pi_{\mathcal{H}} f^* \in \mathcal{H}.$ 。于是，通过写作$f^* = g^* + (f^* - g^*)$并采用前一定理的相同证明，可得到 $f^* \in \mathcal{F}$的如下经验风险误差，

$$\begin{aligned}\mathbb{E}\Big(\big\|(\widehat{f}_{N,\lambda} - f^*)(X_i)\big\|_n^2\Big) \quad &\leq \quad \frac{\lambda}{2} + 2\frac{\sigma^2}{n}\sum_{j=1}^{N}\mathbb{E}_X\Big[\Big(\frac{\mu_j^2}{\mu_j^2+\lambda}\Big)^2\Big] \\ &+ \quad C_1\left(\lambda_{N+1}\|\pi_H f^*\|_{\mathcal{H}}^2 + \frac{1}{N}\sum_{k\geq N+1}\lambda_k\|\varphi_k\|_\infty^2 + \|f^* - \pi_{\mathcal{H}} f^*\|_{L^2(dP)}^2\right).\end{aligned}$$

在先前的一般设定中，经验风险误差取决于额外的投影误差 $\|f^* - \pi_{\mathcal{H}} f^*\|_{L^2\,(dP)}$，因此必须对其进行估计。

接下来，我们利用风险误差（39）研究与所提出的 TKRR 相关的两个主要问题：如何选择合适的正则化参数 λ，以及 TKRR 估计量的收敛速率。注意，根据（39），前述两个问题本质上取决于特征值序列（λ_k）$_{k\geq 1\cdot}$ 的性质。具体而言， λ_k 的衰减率以及 $N_0 \in \mathbb{N}$ 最小秩（仅取决于核 $\mathbb{K}$），即该衰减率开始成立时对应的最小秩，在推导 TKRR 估计量的最优收敛速率时都起着重要作用。为此，我们考虑两种常见的谱类型：具有多项式衰减率的谱和具有指数衰减率的谱。更确切地说，我们假设对于某个实数$a \geq 0$，积分算子 $T_{\mathbb{K}}$ 的特征函数 φ_k 满足以下一致有界条件

$$\|\varphi_k\|_\infty \lesssim k^{a/2}, \qquad k \geq 1. \tag{54}$$

注意，对于某些正实数 $s > a \geq 0$ 和 $b > 0$， $T_{\mathbb{K}}$的特征值的指数衰减率和多项式衰减率分别为

$$\lambda_k \lesssim e^{-bk}, \quad \forall\, k \geq N_b \tag{55}$$

and

$$\lambda_k \lesssim k^{-2s}, \quad \forall\, k \geq N_s, \tag{56}$$

分别地。这里， N_b 和 N_s 是这些衰减率开始成立时对应的秩。特别地，由于在 (39) 中有 $\frac{\sigma^2}{n}\sum_{j=1}^{N}\mathbb{E}_X\left[\Big(\frac{\mu_j^2}{\mu_j^2+\lambda}\Big)^2\right] \leq \sigma^2\frac{N}{n}$，因此，对于指数衰减率的情形，利用 (54) 和 (55) 可将不等式 (39) 改写如下

$$\mathbb{E}\Big(\big\|(\widehat{f}_{N,\lambda} - f^*)(X_i)\big\|_n^2\Big) \leq \frac{\lambda}{2} + 2\sigma^2\frac{N}{n} + C_2\, N^\eta \lambda_N, \quad \eta = \max(0, a-1). \tag{57}$$

因此，关于数据集大小 n，参数 λ， N 的最优取值为

$$\lambda = O\Big(\sigma^2\frac{N}{n}\Big) = O\Big(N^\eta e^{-bN}\Big), \quad N \geq N_b.$$

直接计算可知，当以下条件成立时，结论成立

$$N = \max\left(N_b, \frac{1}{b}\log\Big(\frac{n}{\sigma^2}\Big)\right), \quad \lambda = O\left(\sigma^2\min\Big(\frac{N_b}{n}, \frac{1}{b\,n}\log\big(\frac{n}{\sigma^2}\big)\Big)\right). \tag{58}$$

以类似的方式，当 $T_{\mathbb{K}}$ 的谱具有多项式衰减率时，利用 (56)，参数 N， λ 的最优取值为

$$N = \max\left(N_s, \Big(\frac{n}{\sigma^2}\Big)^{1/(2s+\gamma)}\right), \quad \lambda = O\left(\min\Big(\frac{N_s}{n}, \Big(\frac{\sigma^2}{n}\Big)^{1-1/(2s+\gamma)}\right), \quad \gamma = \min(1, 2-a). \tag{59}$$

我们刚刚证明了以下命题，该命题给出了当 $f^* \in \mathcal{H}$ 且特征值以指数或多项式速率衰减时，我们的TKRR估计量的最优收敛速率 λ_k.

命题1。 在定理 *2* 的假设下，在假设 (55) 成立时，*TKRR* $\widehat{\ }$估计量$f_{N,\,\lambda}$ 的最优收敛速率为

$$\mathbb{E}\Big(\big\|\big(\widehat{f}_{N,\lambda}-f^*\big)(X_i)\big\|_n^2\Big)=O\left(\sigma^2\max\Big(\frac{N_b}{n},\frac{1}{b\,n}\log\big(\frac{n}{\sigma^2}\big)\Big)\right),\quad N\geq N_b. \tag{60}$$

此外，在假设 (56)成立时，我们有

$$\mathbb{E}\Big(\big\|\big(\widehat{f}_{N,\lambda}-f^*\big)(X_i)\big\|_n^2\Big)=O\left(\max\Big(\sigma^2\frac{N_s}{n},\Big(\frac{n}{\sigma^2}\Big)^{-\frac{(2s+\gamma-1)}{(2s+\gamma)}}\Big)\right),\quad \gamma=\min(1,2-a),\quad N\geq N_s. \tag{61}$$

注 3。 利用正紧积分算子 $T_{\mathbb{K}}$在 $\epsilon>0$ 水平下的一致自由度这一概念，可以推广并改进前一个命题中的结果 (60) 。因此，该一致自由度记为 $d_\infty(T_{\mathbb{K}},\epsilon)$ ，并定义为

$$d_\infty(T_{\mathbb{K}},\epsilon)=\min\{k\in\mathbb{N};\,\lambda_k\leq\epsilon\}=N_{T_{\mathbb{K}}}(\epsilon)=\#\{\lambda_j;\,\lambda_j>\epsilon\}+1. \tag{62}$$

在这种情况下，出现在 (57) 右侧的精细截断阶数 N可简单地表示为

$$N=N(\epsilon_n)=d_\infty(T_{\mathbb{K}},\epsilon_n)\quad with\quad \epsilon_n=\frac{\sigma^2}{n}\big(d_\infty(T_{\mathbb{K}},\epsilon_n)\big)^{1-\eta},\quad \eta=\max(0,a-1). \tag{63}$$

除了前一个命题所揭示的良好收敛性质外，与完整核岭回归估计量相比，我们的 TKRR 估计量还具有计算负荷低这一理想特性。

注 4。 由 (18) 和 (19)可知，*TKRR*估计量的时间复杂度为 $O(nN^2)$，而完整核岭回归估计量的时间复杂度为 $O(n^3)$。由于通常有 $N\ll n$，因此 *TKRR* 比经典核岭回归快得多。另一方面，我们的 *TKRR* 与文献 *[2]* 中给出的谱截断核岭回归具有相同的时间复杂度。不过，我们的 *TKRR* 无需进行任何奇异值分解，并且能够处理从相当一般的未知多元概率分布中抽取的随机抽样数据集。此外，所提出的 *TKRR* 与文献 *[25]* 中给出的草图核岭回归组合估计量相比也具有竞争力。对于稠密草图，该估计量的时间复杂度为 $O(n^2N^2)$ ；对于傅里叶草图和哈达玛草图，其时间复杂度为 $O(n^2\log(N))$。

我们的 TKRR 的另一项重要特性是其最优收敛速率与完整核岭回归估计量的最优收敛速率相同。更确切地说，以下命题表明，完整核岭回归估计量具有相同的收敛速率 (60) 和 (61)。注意，完整核岭回归估计量 $\widehat{f}_n^{\lambda}$ 由 (6)–(8) 给出。

命题 2. 在定理 *1* 和定理 *2* 的假设以及条件 (55) 和(56)下，完整核岭回归估计量 $\widehat{f}_n^{\lambda}$ 的最优收敛速率分别与收敛速率(60)和 (61)相同。

证明： 要推导完整核岭回归估计量 (6)–(8) 的经验风险误差，可以对定理 2 的证明稍作改编。更确切地说，无需考虑由 (42) 给出的正则化截断矩阵 A_N，的伪逆，只需考虑由 (7) 给出的正则化完整格拉姆矩阵 $A_n+\lambda I_n$，的逆。此外，在这种情况下，得益于表示定理

以及假设 $f^* \in \mathcal{H}$ 并且按照 [2], 中的做法，我们可以假设 $f^*(X_i) = \widetilde{\pi}_n f^*(X_i)$ 对于$1 \le i \le n$. 在这种情况下， $\widehat{f}_n^{\lambda}$ 的经验风险误差为

$$\mathbb{E}\Big(\big\| (\widehat{f}_n^{\lambda} - f^*)(X_i) \big\|_n^2 \Big) \le \frac{\lambda}{2} + 2\frac{\sigma^2}{n} \sum_{j=1}^{n} \mathbb{E}_X \Big[\Big(\frac{\lambda_j(A_n)}{\lambda_j(A_n) + \lambda} \Big)^2 \Big] \tag{64}$$

由于 $0 < \dfrac{\lambda_j(A_n)}{\lambda_j(A_n) + \lambda} < 1$，因此对于任意整数 $1 \le N < n$,
我们有

$$\begin{aligned} \mathbb{E}\Big(\big\| (\widehat{f}_n^{\lambda} - f^*)(X_i) \big\|_n^2 \Big) &\le \frac{\lambda}{2} + 2\sigma^2 \frac{N}{n} + 2\frac{\sigma^2}{n} \sum_{j=N+1}^{n} \mathbb{E}_X \Big(\frac{\lambda_j(A_n)}{\lambda_j(A_n) + \lambda} \Big) \\ &\le \frac{\lambda}{2} + 2\sigma^2 \frac{N}{n} + 2\frac{\sigma^2}{n\lambda} \sum_{j=N+1}^{n} \mathbb{E}_X \big(\lambda_j(A_n) \big). \end{aligned} \tag{65}$$

在积分算子 $T_{\mathbb{K}}$ 的特征值序列呈指数衰减这一假设下，依据式（55），并利用前述不等式和估计（22），对于 $N_b \le N < n$,

$$\mathbb{E}\Big(\big\| (\widehat{f}_n^{\lambda} - f^*)(X_i) \big\|_n^2 \Big) \lesssim \frac{\lambda}{2} + 2\sigma^2 \frac{N}{n} + 2\frac{\sigma^2}{n\lambda} \lambda_{N+1} \|\varphi_{N+1}\|_{\infty}^2. \tag{66}$$

因此， λ 和 N 的最优值满足以下估计

$$\lambda = O\Big(\sigma^2 \frac{N}{n} \Big) = O\Big(\frac{\sigma^2}{n} N^{\eta} e^{-bN} \Big), \quad \eta = \max(0, a-1), \quad N \ge N_b. \tag{67}$$

结合（66）和（67）可知，在假设（55）下，我们的 TKRR 估计量与完整的核岭回归具有相同的收敛速率。类似地，容易验证，在假设（56）下情况也相同。具体细节留给读者。

5 示例与数值模拟

本节首先阐述第 3 节和第 4 节的结果。我们考虑 Sinc 核和高斯核的两个示例。这些 Mercer 核常用于基于核岭回归的 NP 回归估计量。随后，我们给出数值模拟，以展示所提出的 TKRR 估计量的理论性质。

5.1 Sinc 核与高斯核示例

5.1.1 Sinc 核情形

众所周知，关于 Sinc 核算子的谱分析，以及与该核相关的各种数理统计和信号处理应用，已有丰富的文献，参见例如 [11]。下面简要介绍该核及其相关的积分算子。设正实数 $c > 0$ 称为带宽，则一维 Sinc 核定义为

$$\mathbb{K}_c(x, y) = \frac{\sin(c(x-y))}{\pi(x-y)}, \quad x, y \in I = [-1, 1]. \tag{68}$$

由于 $\mathbb{K}_c(x, y) = \kappa(x-y)$，其中 $\kappa(\cdot)$ 是关于$I = [-1, 1]$上的均匀概率测度的傅里叶变换，因此根据 Bochner 定理， $\mathbb{K}_c(\cdot, \cdot)$ 是一个正定核。众所周知，与之关联的再生核希尔伯特空间是带宽为 $c.$ 的带限函数空间。也就是说，

它是由 $L^2(\mathbb{R})$ 中傅里叶变换的支撑集包含于 $[-c,c]$. 的函数构成的子空间。与 Sinc核 相关的通常概率测度是均匀测度 $d\mu(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)dx$. 我们回顾一下，如果 $\mathcal{Q}_c$ 是定义在 $L^2(I,d\mu)$ 上、核为 $\mathbb{K}_c(x,y)$ 的希尔伯特–施密特算子，那么我们有

$$\mathcal{Q}_c(\varphi_{m,c})(x) = \int_{-1}^{1} \frac{\sin(c(x-y))}{\pi(x-y)}\varphi_{m,c}(y)dy = \lambda_m(c)\varphi_{m,c}(x), \quad x \in [-1,1].$$

这里，$\varphi_{m,c}$, $\lambda_m(c)$ 是第 $m-$ 个特征函数及其对应的特征值，属于 $\mathcal{Q}_c$. 。$\varphi_{m,c}$ 称为长椭球波函数（PSWFs）。自 20 世纪 60 年代 D. Slepian 与其合作者 H. Landau 和 H. Pollak 开创该领域以来，文献中已对长椭球波函数的丰富性质及其特征值的行为进行了广泛研究，更多详情请参见 [11]。自早期关于长椭球波函数的研究以来，人们已知 $\lambda_m(c)$ 序列具有渐近超指数衰减率。近期， [7],表明 $\lambda_m(c)$ 的最优渐近超指数衰减率为

$$\lambda_m(c) \lesssim \exp\Big(-2m\log\Big(\frac{4m}{ec}\Big)\Big), \quad m \gg 1.$$

此外， [9], 给出了 $\lambda_m(c)$ 的以下非渐近衰减率,

$$\lambda_m(c) \le \exp\Big(-(2m+1)\log\Big(\frac{2(m+1)}{ec}\Big)\Big), \quad m \ge \frac{ec}{2} = N_c. \tag{69}$$

因此，对于 $d=1$ 以及 Sinc 核的情形，可以在命题1的经验风险误差（60）中取 $\underline{b}=2$ 和 $N_b = \frac{ec}{2}$。也就是说

$$\mathbb{E}\Big(\big\|(\widehat{f}_{N,\lambda} - f^*)(X_i)\big\|_n^2\Big) = O\left(\sigma^2 \max\Big(\frac{ec}{2n}, \frac{1}{2\,n}\log\big(\frac{n}{\sigma^2}\big)\Big)\right), \quad N \ge \frac{ec}{2}. \tag{70}$$

注意，当带宽 c, 取较大值时，衰减秩 $N_c = \frac{ec}{2}$ 是悲观的。事实上，根据 H. Landau 的研究， $\lambda_n(c)$ 的衰减区域始于 $N_{1,c} = \frac{2c}{\pi} + O\big(\log(c)\big)$. 更确切地说，已有研究表明，对于 Sinc 核，层级 $\epsilon > 0$ 下的一致自由度为

$$d_\infty(T_{\mathbb{K}_c}, \epsilon) = \frac{2c}{\pi} + \frac{1}{\pi^2}\log\Big(\frac{1-\epsilon}{\epsilon}\Big)\log\Big(\frac{2c}{\pi}\Big) + o\big(\log c\big), \quad c \gg 1.$$

此外，文献 [9] 还表明

$$\sup_{x\in[-1,1]} |\varphi_{m,c}(x)|^2 \lesssim \big(m(m+1)+c^2\big)^{\frac{1}{2}} \lesssim m, \quad \forall\, m \ge \Big[\frac{2c}{\pi}\Big]. \tag{71}$$

也就是说，对于满足 $a=1$ 的 Sinc 核，利用式 (63) 以及前述一致自由度，可以得到一个更精细的截断阶数

$$N = N(\epsilon_n) = \left\lceil \frac{2c}{\pi} + \Big(\frac{1}{\pi^2}\log\big(\frac{1}{\epsilon_n}\big) + 1\Big)\log\Big(\frac{2c}{\pi}\Big)\right\rceil, \tag{72}$$

其中， ϵ_n 是该方程的一个解

$$\epsilon_n = \frac{\sigma^2}{n}\left(\frac{2c}{\pi} + \Big(\frac{1}{\pi^2}\log\Big(\frac{1}{\epsilon_n}\Big) + 1\Big)\log\Big(\frac{2c}{\pi}\Big)\right). \tag{73}$$

这里，$\lceil x \rceil$ 表示大于或等于 $x.$ 的最小整数。最后，对于维度 $d \geq 2$，可以使用张量积 Sinc 核

$$\mathbb{K}_c^d(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{d} \mathbb{K}_c(x_i - y_i), \quad \boldsymbol{x} = \big(x_i\big)_{i=1}^d, \quad \boldsymbol{y} = \big(y_i\big)_{i=1}^d \in I^d.$$

与 $\mathbb{K}_c^d(\boldsymbol{x}, \boldsymbol{y})$ 相关的 $d-$ 变量特征函数及其对应特征值，由 $\varphi_{k,c}(x_i)$ 及其对应特征值 $\lambda_k(c)$ 的张量积给出。。这使我们能够推导出特征值的衰减率估计，以及与由 Sinc 核构成的 $d-$ 维张量积相关的一致自由度。具体细节留给读者。

5.1.2 高斯核情形

与 Sinc 核一样，我们首先回顾一维高斯核 $\mathbb{K}_G(x, y)$ $x, y \in \mathbb{R}$ 及其一些相关谱性质。然后，我们展示如何将这些性质推广到 $d-$ 维情形。众所周知，高斯核是数理统计和机器学习应用中最常用的核函数之一。该核定义如下。考虑两个正实数 $\xi, c > 0$，则 $\mathbb{K}_G$ 是一个正定卷积核，其对应的通常概率测度为 $dP_c(x)$，定义为

$$\mathbb{K}_G(x, y) = e^{-\xi(x-y)^2}, \qquad dP_c(x) = \sqrt{\frac{\pi}{2c}} e^{-2cx^2} \quad x, y \in \mathbb{R}. \tag{74}$$

由 [29], 积分算子的特征值 $T_{\mathbb{K}_G}$ 可得

$$\lambda_k = \sqrt{\frac{\pi}{\xi + c + \sqrt{c^2 + 2c\xi}}} \left(\frac{\xi}{c + \xi + \sqrt{c^2 + 2c\xi}} \right)^{k-1}, \quad k \geq 1. \tag{75}$$

相关特征函数 ϕ_k， $k \geq 1$ 是众所周知的膨胀厄米函数，表示为

$$\varphi_k(x, \gamma) = \frac{1}{\sqrt{2\gamma}} \phi_k(\sqrt{2\gamma} x) = \frac{1}{\sqrt{2\gamma}} \alpha_{k-1} H_{k-1}\big(\sqrt{2\gamma} x\big) e^{-\gamma x^2}, \quad \gamma = \sqrt{c^2 + 2c\xi}, \quad k \geq 1. \tag{76}$$

这里，$\alpha_k = \dfrac{2^{-k/2}}{\pi^{1/4}\sqrt{k!}}$ 和 $H_k(x) = (-1)^k e^{x^2} \dfrac{d^k}{dx^k}\Big(e^{-x^2}\Big)$， $k \geq 0$ 是厄米多项式。此外，众所周知，归一化厄米函数 ϕ_k 满足以下一致有界性，

$$\sup_{x \in \mathbb{R}} |\phi_k(x)| \leq 1, \quad k \geq 1. \tag{77}$$

因此，利用 (75)–(77) 以及带有 $\eta = 0,$ 的经验风险误差 (57)，可以根据参数$\xi,$ $c > 0$，并按照规则 (58)（其中 $b = \dfrac{c + \xi + \sqrt{c^2 + 2c\xi}}{\xi}$.），得到截断阶数 N 和正则化参数 $\lambda > 0,$ 的最优取值

最后，对于一般的 $d-$维情形，高斯核及其对应的概率测度定义如下:

$$\mathbb{K}_G^d(\boldsymbol{x}, \boldsymbol{y}) = e^{-\xi \|\boldsymbol{x} - \boldsymbol{y}\|_2^2}, \quad dP_c(\boldsymbol{x}) = \left(\frac{\pi}{2c}\right)^{d/2} e^{-2c\|\boldsymbol{x}^2\|_2^2}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

相应的 $d-$元特征函数及其对应的特征值由一元 $\varphi_k(x_i, \gamma)$ 及其对应的 $\lambda_k.$ 的张量积给出

5.2 数值模拟

本段中，我们给出一些数值模拟，以展示本文的不同结果，尤其是定理 1 和定理 2 的结果，以及后者的推论。我们仅讨论一维情形，因为本文的目的是阐释理论结果。以下三个示例给出了这些数值模拟。

例1： 在这个第一个例子中，我们展示定理 1 及其推论 1 的结果。为此，我们首先考虑由 (68) 给出的 Sinc核 $\mathbb{K}_c$，其中 $c = 25$. 然后，我们没有采用其通常对应的抽样概率测度 $dP(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)\,dx$，而是考虑了一组服从截断标准正态分布的 $n = 200$ 独立同分布样本，其支撑集为$I = [-1, 1]$，概率密度函数为

$$\rho(x) = \frac{\phi(x)}{\Phi(1) - \Phi(-1)}\mathbf{1}_{[-1,1]}(x), \quad \phi(x) = \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big), \quad \Phi(x) = \frac{1}{2}\big(1 + \mathrm{erf}(x/\sqrt{2})\big). \tag{78}$$

众所周知，积分算子 $T_{\mathbb{K}_c} = \mathcal{Q}_c$ 是单射。此外，容易看出，$d\rho(x)$ 相对于均匀概率测度 $dP(x)$，是绝对连续的。因此，对于这两个概率测度，定理 1 的条件均得到满足。我们取截断阶数 $N = 25$，并计算了完整格拉姆矩阵 B_n^c 及其主 $n \times N$ 子矩阵 A_N^c的特征值在 10 次实现上的平均值，其定义为

$$B_n^c = \frac{1}{n}\left[\mathbb{K}_c(X_i, X_j)\right]_{1\le i,j\le n}, \qquad A_N^c = \frac{1}{n}\left[\mathbb{K}_c(X_i, X_j)\right]_{1\le i\le n, 1\le j\le N}.$$

回顾一下，对于每次实现，这些 Gram 矩阵都是通过对从 $\rho(\cdot)$ 中抽取的 $n = 200$ 个独立同分布随机抽样点应用 Sinc核构造而成的。 在图1(a)中，我们绘制了特征值 $\lambda_j(B_n^c)$、奇异值 $\mu_j((A_N^c))$以及 *Sinc*核积分算子的真实特征值 $()$ 的平均值 $()$ 和 $()$ 的图像，并以真实特征值为横轴。此外，为了呈现前述特征值序列的快速衰减率，我们在图1(b)中绘制了这些特征值平均值的对数图像。此外，为了说明式（22）和（28）给出的 B_n^c 和 A_N^c 的迹尾项 $\mathcal{T}_k(B_n^c)$ 和 $\mathcal{T}_k(A_N^c)$ 的衰减估计，我们在图2(a)中绘制了 $\mathcal{T}_k(B_n^c)$ 和 $\mathcal{T}_k(A_N^c)$ 的图像，并以真实特征值 $\lambda_k(T_{\mathbb{K}_c})$ 为横轴 $1 \le k \le N$. 在图2(b)中，对于不同的 $1 \le k \le N.$，我们绘制了 $\log(\mathcal{T}_k(B_n^c))$、$\log(\mathcal{T}_k(A_N^c))$ 关于真实特征值 $\log(\lambda_k(T_{\mathbb{K}_c}))$ 的图像。图 1 和图 2 中的数值模拟结果与定理 1 和推论 1 的理论结果相符。

此外，我们考虑了高斯核 $K_G(x,\ y) = e^{-\xi(x-y)^2},\ x,\ y \in I = [-1, 1]$，其中$\xi = 25$. 相关的格拉姆矩阵 B_n^G及其主子矩阵 A_N^G，对于 $N < n$ 定义为

$$B_n^G = \frac{1}{n}\left[\mathbb{K}_G(X_i, X_j)\right]_{1\le i,j\le n}, \qquad A_N^G = \frac{1}{n}\left[\mathbb{K}_G(X_i, X_j)\right]_{1\le i\le n, 1\le j\le N}.$$

这里，X_i 是从 $I = [-1, 1].$ 上的均匀分布中抽取的独立同分布随机样本。尽管均匀概率测度并非核 $\mathbb{K}_G$ 对应的常用高斯测度，但正如定理 1及其推论 1所预测的，B_n^G和 A_N^G 的特征值具有与真实特征值序列 λ_k，相似的衰减率，该序列由式（75）给出，其中 $\xi = 25$ 且 $c = 1.$。图3中，我们根据式（75）绘制了积分算子 $\lambda_j(T_{\mathbb{K}_G})$，的特征值，以及 $\lambda_j(B_n^G)$ $\overline{\mu_j}(A_N^G)$，即 B_n^G的特征值和 A_N^G的奇异值在 10 次实现中的平均值，其中 $n = 200$ 且 $N = 25.$

例2： 在这个例子中，我们展示定理 2、命题1和命题 2的结果。为此，我们考虑如下非参数回归问题（1）：

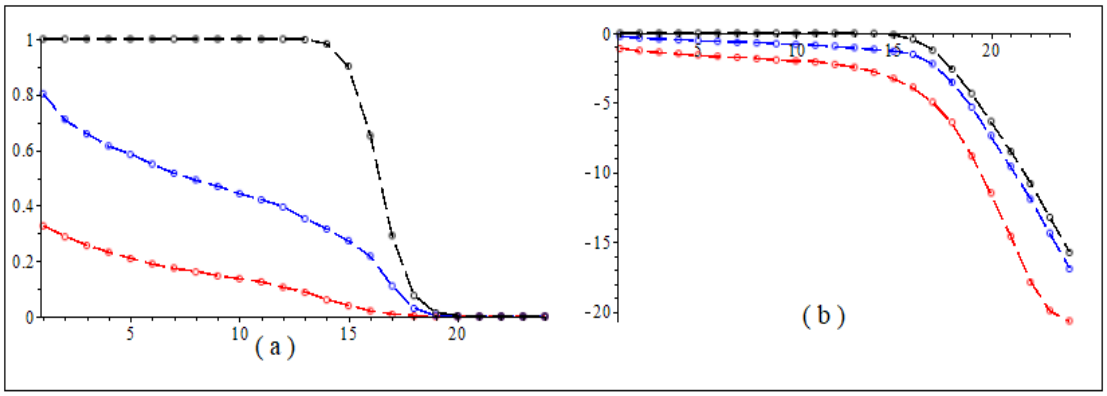


图1：(a) $\lambda_j(T_{\mathbb{K}_c})$（黑色）与 $\lambda_j(B_n^c)$（蓝色）的图，以及奇异值$\mu_j(A_N^c)$，其中$c = 25$, $n = 200$, $N = 25$。(b) $\log(\lambda_j(T_{\mathbb{K}_c}))$（黑色）与 $\log(\lambda_j(B_n^c))$（蓝色）的图，以及 $\log(\mu_j(A_N^c))$（红色）。

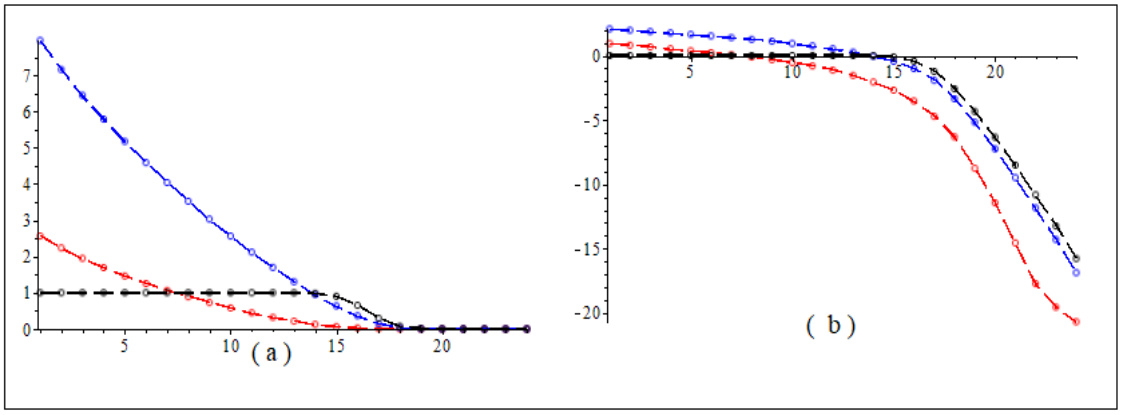


图2：(a) 迹尾 $\mathcal{T}_k(B_n^c)$（蓝色）和 $\mathcal{T}_k(A_N^c)$（红色）相对于 $\lambda_k(T_{\mathbb{K}_c})$（黑色）的图，c=25,$n = 200$, $N = 25$. (b) $\log(\mathcal{T}_k(B_n^c))$（蓝色）和 $\log(\mathcal{T}_k(A_N^c))$（红色）相对于 $\log(\lambda_k(T_{\mathbb{K}_c}))$（黑色）的图。

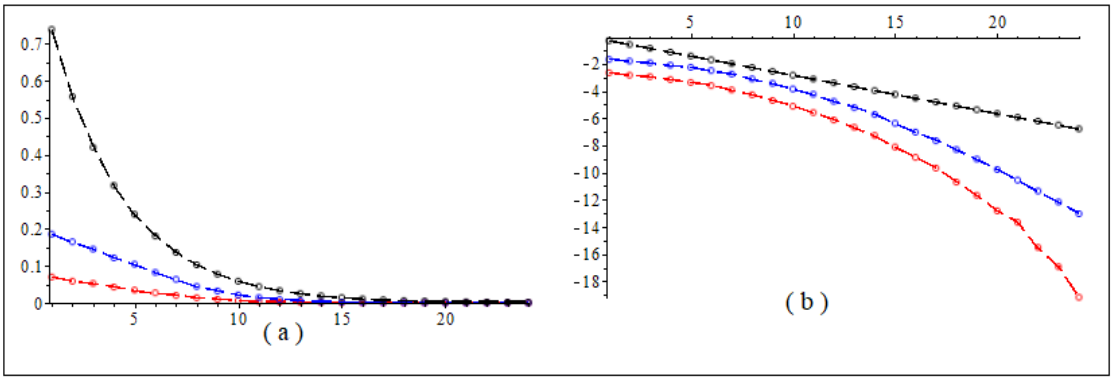


图3：(a) $\lambda_j(T_{\mathbb{K}_G})$（黑色）与 $\lambda_j(B_n^G)$（蓝色）的图，以及奇异值 $\mu_j(A_N^G)$，其中$n = 200$，$N = 25$，$\xi = 25$，$c = 1.$；(b) $\log(\lambda_j(T_{\mathbb{K}_G}))$（黑色）与 $\log(\lambda_j(B_n^G))$（蓝色）的图，以及 $\log(\mu_j((A_N^G))$（红色）。

$d = 1$ 以及真实回归函数 f^*，由下式给出

$$f^*(x) = \frac{\sin(20\,x)}{20\,x},\ x \in I = [-1, 1].$$

我们使用 Mercer 核 $\mathbb{K} = \mathbb{K}_c$，即取特殊值为 $c = 25$ 的 Sinc核。众所周知，当其与 I 上的均匀测度结合时，与 相关联的再生核希尔伯特空间，是由 $L^2(\mathbb{R})$ 中傅里叶变换支撑于 $[-c, c]$. 的函数构成的空间。因此，前述测试函数 $f^* \in \mathcal{H}$. 。我们考虑了 $n = 200$ 个独立同分布的 X_i，它们取自前一例中的截断正态分布。此外，添加的独立同分布噪声 ε_i 取自中心正态分布，其标准差有两个不同取值： $\sigma = 0.1$ 和 $\sigma = 0.5.$ 。然后，我们构造了估计量 (19)，其正则化参数 λ，按照规则 (58) 选取，其中 $b = 2$ 且 $N_b = \left\lceil \frac{ec}{2} \right\rceil = 34$，即 $\lambda = 2.47e - 4.$ 。我们考虑了四个截断阶数：$N = 20$，25，30，50。我们计算了 *10* 次实现的经验风险误差 和理论经验风险误差，的均值，后者由 (70) 给出。即

$$\widehat{\mathcal{R}}_N = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}_{N,\lambda}(X_i) - f^*(X_i)\right)^2, \quad \mathcal{R}_n = O\left(\sigma^2\frac{ec}{2\,n}\right).$$

数值结果见表1。这些结果表明，定理 2 和命题1给出的理论经验风险误差界相当紧。此外，正如我们此前所提到的，对于 Sinc核情形，精细化后的最优截断阶数由式 (72)-(73) 给出。具体而言，对于 $c = 25$，该最优截断阶数在 $\sigma = 0.1$时为 $N_1 = 21$，在 $\sigma = 0.5.$ 时为 $N_2 = 20$ 。这解释了为什么表1中针对 $N = 20$ 得到的经验风险误差与较大 $N.$ 值下得到的误差相近。

例3： 在最后这个例子中，我们考虑高斯核 $K_G(x, y) = e^{-\xi(x-y)^2}$, $x, y \in I = [-1, 1]$ 其中 $\xi = 25$ 且 $c = 1.$ 接着，我们进行与前面类似的数值模拟

σ	N	$\widehat{\mathcal{R}}_N$	$\mathcal{R}_n$	σ	N	$\widehat{\mathcal{R}}_N$	$\mathcal{R}_n$
0.1	20	$9.05e-3$	$1.70e-3$	0.5	20	$2.07e-2$	$4.25e-2$
–	25	$4.64e-3$	$1.70e-3$	–	25	$2.21e-2$	$4.25e-2$
–	30	$1.02e-3$	$1.70e-3$	–	30	$2.20e-2$	$4.25e-2$
–	50	$1.00e-3$	$1.70e-3$	–	50	$1.92e-2$	$4.25e-2$

表1：例2的数值结果。

例子，不过使用新的合成回归函数

$$f^*_{\xi,c}(x) = \exp\Big(-\sqrt{c^2+c\xi}\Big)\cdot\Big(1+\sum_{j=1}^{10}\frac{x^j}{j}\Big), \quad \xi = 25,\, c = 1. \tag{79}$$

注意，当与式（74）给出的高斯概率测度 dP_c相关联时，对应于 $\mathbb{K}_G$ 的再生核希尔伯特空间 $\mathcal{H}$由式（76）给出的厄米特函数 ϕ_k， $k \geq 1$张成。因此，前述测试回归函数 $f^*_{\xi,\ c} \in \mathcal{H}$. 此外，在每次的10次实现中，我们都考虑一组新的 $n = 200$ 独立同分布随机抽样点 X_i，这些点从 I. 上的均匀分布中抽取。在表 ⌢2中，我们列出了经验风险误差 $\mathcal{R}_N$ 的均值，以及由式（60）给出的理论经验风险误差$\mathcal{R}_n$，其中 $N_b = 20$ 且 $b = \underset{\xi}{\underline{^{c+\xi+\sqrt{c^2+2c\xi}}\ \overline{\approx\ \frac{4}{3}\cdot\circ}}}$ 也就是说

$$\widehat{\mathcal{R}}_N = \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{f}_{N,\lambda}(X_i) - f^*_{\xi,c}(X_i)\Big)^2, \quad \mathcal{R}_n = O\Big(\frac{\lambda}{2} + \frac{2\sigma^2 N_b}{n} + \lambda_N\Big).$$

这里， λ 和 λ_N 分别由式（58）和（75）给出。同样，表2中的数值结果与定理2和命题1的理论结果高度一致。

σ	N	$\widehat{\mathcal{R}}_N$	$\mathcal{R}_n$	σ	N	$\widehat{\mathcal{R}}_N$	$\mathcal{R}_n$
0.1	20	$1.13e-2$	$4.75e-3$	0.5	20	$1.20e-2$	$5.01e-2$
–	25	$5.26e-3$	$4.75e-3$	–	25	$1.43e-2$	$5.01e-2$
–	30	$1.91e-3$	$4.75e-3$	–	30	$1.04e-2$	$5.01e-2$
–	50	$1.10e-3$	$4.75e-3$	–	50	$8.72e-3$	$5.01e-2$

表 2：例3的数值结果。

致谢 作者谨向 Aline Bonami 教授致谢，感谢她就定理 1 的结果及其可能的推广所进行的富有成效的讨论并提出宝贵建议。

参考文献

[1] A. E. Alaoui 和 M.W. Mahoney，《具有统计保证的快速随机化核方法》，神经信息处理系统进展，(2015) 775–783。

[2] A . A. Amini, Spectrally-t$_{\text{runca}}$ ted kernel ridge regress ¡on an d its free lunch, *Electron. J. Stat* **15** ., (2021) 3743–3761.[3] H. Avron、K. L. Clarkson 和 D. P. Woodruff, 《利用草图和预条件加速核岭回归》, SIAM矩阵分析与应用杂志, 38 (4) (2017) 1116– 1138.[4] H. Avron、M. Kapralov、C. Musco、M. Christopher、A. Velingker 和 A. Zandieh, 《用于核岭回归的随机傅里叶特征：近似界与统计保证》, 第34届国际机器学习会议论文集, PMLR, 70 (2017) 253–262.[5] F. Bach, 《核求积规则与随机特征展开之间的等价性》, 机器学习研究杂志, 18 (21) (2017) 1–38.[6] Bauer, B. 和 Kohler, M. (2019)。《深度学习能否弥补非参数回归中的维数灾难》, 统计年刊, 47 (4), 2261–2285.[7] A. Bonami 和 A. Karoui, 《时频限制算子的谱衰减》, 应用与计算调和分析, 42 (2017) 1–20.[8] A. Bonami 和 A. Karoui, 《有限傅里叶变换的随机离散化及相关核随机矩阵》, J. Fourier Anal. Appl., 26 (2020) https://doi.org/10.1007/s00041-020-09736-8[9] A. Bonami、Ph. Jaming 和 A. Karoui, 《sinc 核算子谱的非渐近行为及相关应用》, J. Math. Phys., 62 (3) (2021).[10] C. Boutsidis 和 A. Gittens, 《通过子采样随机哈达玛变换改进矩阵算法》, SIAM矩阵分析与应用杂志, 34 (2013) 1301–1340.[11] J.A. Hogan 和 J.D. Lakey, 《时长与带宽限制：长球函数、采样及其应用》, 应用与数值调和分析丛书, Birkh¨auser, Springer, 纽约, (2013).[12] R. A. Horn 和 C. R. Johnson, 《矩阵分析》, 第二版, 剑桥大学出版社, (2013).[13] M. G. Genton, 《机器学习的核函数类别：统计学视角》, 机器学习研究杂志, 2 (2001) 299–312.[14] E. R. Gizewski、L. Mayer、B. A. Moser、D-H. Nguyen、S. Pereverzyev、S. V. Pereverzyev、N. Shepeleva 和 W. Zellinger, 《再生核希尔伯特空间中无监督领域自适应的正则化》, 应用与计算调和分析, 57 (2022) 201–227.[15] T. Hofmann、B. Sch¨olkopf 和 A. J. Smola, 《机器学习中的核方法》, 统计年刊, 36 (3) (2008) 1171–1220.[16] Y. K. Lee、E. Mammen、J. P. Nielsen 和 B. U. Park, 。《借助参数模型的非参数回归》, 电子统计学杂志, 14(2) (2020) 3845–3868.[17] A. Rahimi 和 B. Recht, 《面向大规模核机器的随机特征》, 载于《神经信息处理系统进展》, (2007) 1177–1184.[18] C. E. Rasmussen 和 K. I. Williams, 《机器学习中的高斯过程》, 麻省理工学院出版社, (2006)。

[19] A. Rudi 和 L. Rosasco, 《利用随机特征进行学习的泛化性质》, 神经信息处理系统进展 (2017) 3218–3228.[20] A. Rudi、L. Carratino 和 L. Rosasco, 《FALKON：一种最优的大规模核方法》, *NIPS'17*：第*31*届神经信息处理系统国际会议论文集(2017) 3891–3901.[21] J. Shawe-Taylor、K. I. Williams、N. Cristianini 和 J. Kandola, 《格拉姆矩阵的特征谱与核主成分分析的广义误差》, *IEEE*信息论汇刊**51** (7) (2005) 2510–2522.[22] S. Smale 和 D. X. Zhou, 《通过积分算子及其近似进行学习理论估计》。构造性近似**26** (2) (2007) 153–172.[23] R.C. Thompson, 《主子矩阵 IX：子矩阵奇异值的交错不等式》, 线性代数及其应用**5** (1972) 1–12.[24] J. A. Tropp, 《矩阵集中不等式导论》, 机器学习基础与趋势丛书**8** No. 1–2, Now Publishers Inc, (2015).[25] Yang、Y. Pilanci、M 和 Wainwright, M. J. (2017)。《核的随机草图：快速且最优的非参数回归》。统计年刊**45**，991–1023.[26]T. Zhang 和 N. Simon, 《再生核希尔伯特空间中非参数回归的在线投影估计量》, *Sinica*统计学报(2022) doi:https://doi.org/10.5705/ss.202021.0018.[27] G. Zhao 和 Y. Ma, 《稳健非参数核回归估计量》, 统计学与概率论通讯**116**(2016) 72–79.[28] L. Shu、J-F. Ton、D. Oglic 和 D. Sejdinovic, 《迈向随机傅里叶特征的统一分析》, 机器学习研究杂志**22** (2021) 1–51.[29] H. Zhu、C. K. I. Williams、R. J. Rohwer 和 M. Morciniec, 《高斯回归与最优有限维线性模型》, 载于 Bishop, C. M. 编, 《神经网络与机器学习》, Springer-Verlag，柏林 (1998)。