arXiv:2303.09206v1 [math.ST] 16 Mar 2023

无界采样下正则化三角线性回归的误差分析：统计学习视角

Anna Scampicchio、Elena Arcari、Melanie N. Zeilinger

摘要——非参数基于核的方法在函数估计方面效果显著，但代价是计算复杂度较高，这限制了它们在自适应模型控制中的应用。受稀疏谱高斯过程所采用的近似技术启发，本文聚焦于正则化三角线性回归模型。本文在统计学习框架下分析了这类估计设置的性能。具体而言，我们针对有限维空间推导出一个新的样本误差上界，并将支撑可能无界的噪声纳入考量。随后，我们研究近似误差，并通过结合这两个上界，讨论正则化参数对偏差-方差权衡的影响。

I. 引言

正则化函数估计的非参数方法是机器学习中的重要工具，已成功应用于系统辨识 [1] 和基于学习的控制 [2], [3]等领域。然而，由于其计算复杂度随数据量呈立方增长，这类方法难以应用于实时场景。为实现基于核的方法的快速自适应，已有策略可分为两大类：输入位置选择和核的低秩近似 [4]。在第二类方法中，稀疏谱高斯过程 [5], [6], 取得了显著成功：对平移不变核进行运算，可通过傅里叶特征的线性组合得到参数化近似。

本文受到后一种方法的启发，在由 E 预定义三角函数的张成空间所确定的有限维假设空间 $\mathscr{H}$ 中进行正则化回归。这样的设定放宽了对平移不变核的假设；与未正则化的基于投影的方法 [7]相比，它对于潜在的基函数设定错误更为稳健。关于基于傅里叶特征的参数化方法综述，请参阅 [8, 第 1 章7]。我们的目标是评估所提出估计器的性能，这是为利用此类模型的数据驱动、基于模型的控制方案推导可靠保证的关键一步（例如参见 [9], [10], [11]）。

我们在统计学习设定 [12]–[13]下开展这项分析。待估计函数（即回归函数 f_ρ）被定义为在一个（部分）未知的概率分布上使期望风险最小化的函数，该概率分布共同定义定义在输入-输出空间上，并从中抽取独立同分布样本。因此，这种表述也能处理完全非线性的测量模型。此外，f_ρ 通常假设属于平方可积函数空间 $\mathscr{L}^2$，而假设空间通常取为无限维再生核希尔伯特空间（RKHS），它与 $\mathscr{L}^2$ 通过插值空间论证相关（[14, 定理 2], [15]）。与经典的非参数设定不同，这里不假设回归函数属于假设空间。因此，可以定义两个对象：实际的基于数据的估计量 f_z 及其无数据极限 $f_{\mathscr{H}}$。误差分析旨在量化逼近误差（即偏差）$\|f_\rho - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ 以及样本误差（即方差）$\|f_{\mathscr{H}} - f_z\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$。关于后者，统计学习文献中已有大量研究。其中大多数研究针对输出上的概率测度具有有界支撑的情形，因此利用诸如 Hoeffding 或 Bennett 不等式之类的集中不等式获得上界[16], [13, 第 3 章。1]。这方面的研究例如有 [17]–[13], [18], [19], [20]。考虑无界采样的研究包括 [21], [22], [23]。这些研究所推导的上界利用所谓的矩假设，该假设放宽了有界性假设，并且也适用于（次）高斯噪声。这类结果依赖于对覆盖数的计算，以量化假设空间的容量[24], ，并展示了最优收敛速率；然而，在非渐近情形下，它们的实际应用价值往往有限，因为乘法系数的取值较大，而且通常难以计算。本文针对由三角函数构成的有限维假设空间进行误差分析。我们的第一项贡献是给出一个样本误差界；即使考虑具有无界支撑的噪声，该界也比文献中已有的界更不保守。第二项贡献是研究正则化三角回归设定中的偏差-方差权衡。为此，我们获得两个逼近误差上界，将它们与样本复杂度结果结合起来，并分析确保存在唯一正则化参数值 γ 以实现最优权衡的条件。我们通过一项蒙特卡洛研究考察了两种正则化参数估计方法的差异。结果显示，其中一种准则给出的 γ 能够体现预言机行为（即使总体误差最小化），从而得出无需预先进行超参数选择的快速估计方案。

作者隶属于苏黎世联邦理工学院动态系统与控制研究所。{ascampicc,earcari,mzeilinger}@ethz.ch

II. 问题设定

设输入的度量空间 $\mathscr{X}$ 为$\mathbb{R}$的一个紧子集：不失一般性，取 $\mathscr{X}=[-X/2, X/2]$，其中 $X\in\mathbb{R}_+$（此处仅为便于可视化而介绍标量情形；多维情形可直接推广）。假设输出空间为 $\mathscr{Y}=\mathbb{R}$。在 $\mathscr{Z}=\mathscr{X}\times$上定义了一个概率测度 $\rho_{\mathscr{Y}}$，根据富比尼定理，它可分解为 $\rho_{\mathscr{X}}(x)$ 和 $\rho(y|x)$。在所考虑的设定中，定义在 $\mathscr{X}$ 上的概率测度为标准均匀分布：记 μ 为勒贝格测度，则对于感兴趣的σ–代数中的任意集合 A，有 $\rho_{\mathscr{X}}(A)=\mu(A\cap\mathscr{X})/\mu(\mathscr{X})=\mu(A\cap\mathscr{X})/X$。因此，$\rho_{\mathscr{X}}$ 是一个博雷尔非退化的 σ–有限测度。关于 $\rho(y|\cdot)$，我们假设它未知，且定义在 $\mathbb{R}$上。

给定从 ρ 中抽取并收集于数据集 $\mathcal{D}=\{(x_t, y_t)\}_{t=1}^N$ 的 N 个独立样本，目标是估计回归函数

$$f_\rho(x)=\int_{\mathscr{Y}} y d\rho(y|x). \tag{1}$$

我们采用以下假设。

假设 1：回归函数 f_ρ 属于 $\mathscr{X}$ 上平方可积函数的空间，记为 $\mathscr{L}^2_{\rho_{\mathscr{X}}}$，且满足 $\|f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}=\sqrt{\int_{\mathscr{X}} f^2(x)d\rho_{\mathscr{X}}(x)}=\sqrt{\int_{\mathscr{X}} f^2(x)d\mu(x)}/X\le B_f$。此外，还满足 $\sigma_\rho^2=\int_{\mathscr{X}}\sigma_\rho^2(x)d\rho_{\mathscr{X}}(x)=\int_{\mathscr{Z}}(y-f_\rho(x))^2d\rho\le B_\sigma^2$。□

换言之，我们假设已知待估计未知函数的能量上界，以及加性噪声的方差上界。

空间 $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ 是一个可分希尔伯特空间，其由三角函数 [25] 构成的完备标准正交基为

$$\left\{\sqrt{2}\sin\left(\frac{2\pi q}{X}x\right),\sqrt{2}\cos\left(\frac{2\pi q}{X}x\right)\right\}_{q\in\mathbb{N}} \tag{2}$$

$$=\{\bar{\varphi}_q^s(x),\bar{\varphi}_q^c(x)\}_{q\in\mathbb{N}}\quad\text{with } x\in\mathscr{X}. \tag{3}$$

因此，任意函数 $f\in\mathscr{L}^2_{\rho_{\mathscr{X}}}$ 都可以表示为$f(\cdot)=\sum_{q\in\mathbb{N}}\alpha_q^s\bar{\varphi}_q^s(\cdot)+\alpha_q^c\bar{\varphi}_q^c(\cdot)$，也可以简洁地写作 $f(\cdot)=\sum_{q\in\mathbb{N}}\alpha_q\bar{\varphi}_q(\cdot)$，其中 $\sum_{q\in\mathbb{N}}\alpha_q^2<\infty$。在这种表示下，我们将目标函数记为$f_\rho(\cdot)=\sum_{q\in\mathbb{N}}\bar{\alpha}_q^s\bar{\varphi}_q^s(\cdot)+\bar{\alpha}_q^c\bar{\varphi}_q^c(\cdot)=\sum_{q\in\mathbb{N}}\bar{\alpha}_q\bar{\varphi}_q(\cdot)$。

无法在 $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ 中进行函数估计，因为逐点求值没有良好定义。因此，我们在具有再生核希尔伯特空间结构的假设空间中执行此任务。具体而言，我们考虑由 (3) 中一个基数为 E 的函数子集所生成的再生核希尔伯特空间，其中 E 根据我们的计算容量进行选择。记 Q 为所选频率的集合，即 $Q=\{q_j\}_{j=1}^{E/2}\subset\mathbb{N}$，并考虑使用 Q 定义的、从 (3) 中提取的以下函数，其中 $j=1,...,E/2$，定义为

$$\varphi_i(x)=\begin{cases}\bar{\varphi}_{q_j}^s(x), & i=j\\ \bar{\varphi}_{q_j}^c(x), & i=j+\frac{E}{2}.\end{cases} \tag{4}$$

那么，我们感兴趣的再生核希尔伯特空间由以下核诱导：

$$\mathscr{K}(x_a,x_b)=\phi^\top(x_a)\Sigma_\alpha\phi(x_b), \tag{5}$$

其中 $\Sigma_\alpha=\mathrm{diag}(\lambda_1,...,\lambda_E)$ 是一个正定矩阵，且向量 $\phi(\cdot)\in\mathbb{R}^E$ 满足$\phi^\top(x)=[\varphi_1(x)\ \dots\ \varphi_E(x)]$. 显然，(5) 是一个 Mercer核（[26, (6), 第 346] 页）；它满足 Mercer条件 $\int_{\mathscr{X}}\int_{\mathscr{X}}\mathscr{K}(x,x')^2d\rho_{\mathscr{X}}(x)d\rho_{\mathscr{X}}(x')=\sum_{i=1}^E\lambda_i^2$，并且当且仅当对所有 $i=1,\dots,E/2$，$\lambda_i\neq\lambda_{i+E/2}$ 时，它是非平稳的。此外，利用 [27, 第 4 章3]中的论证，可得

$$C_{\mathscr{K}}=\sup_{x_a,x_b\in\mathscr{X}}\sqrt{\mathscr{K}(x_a,x_b)}$$

$$\le\sqrt{\sum_{i=1}^{E/2}\max\{\lambda_i,,\lambda_{i+E/2}\}}<+\infty. \tag{6}$$

由于其为 Mercer核，根据 Moore-Aronszajn定理 [26]，$\mathscr{K}$ 与函数希尔伯特空间（$\mathscr{H}$，$\langle\cdot,\ \cdot\rangle_{\mathscr{H}}$）一一对应，该空间为

$$\mathscr{H}=\{f\in\mathscr{L}^2_{\rho_{\mathscr{X}}}:f(\cdot)=\phi^\top(\cdot)\alpha,\ \alpha\in\mathbb{R}^E\} \tag{7}$$

其内积定义为：对于 $f^{(\natural)}(\cdot)=\phi^\top(\cdot)\alpha^{(\natural)}$ 以及 $\natural=a, b$:

$$\langle f^{(a)},f^{(b)}\rangle_{\mathscr{H}}=\langle\Sigma_\alpha^{-1/2}\alpha^{(a)},\Sigma_\alpha^{-1/2}\alpha^{(b)}\rangle_2. \tag{8}$$

在假设空间中，我们可以根据数据集 $\mathcal{D}$ 计算估计值，如下所示。考虑采样算子 $\mathcal{S}_{\mathscr{X}}:\mathscr{H}\to\mathbb{R}^N$，使得 $\mathcal{S}_{\mathscr{X}}(f)=[f\ x_1\ \dots\ f\ x_N\ ]^\top$ $\mathcal{S}_{\mathscr{X}}^\top\ \mathbb{R}^N\to(\ )$ $\mathscr{H}$ $(\ )$，以及其伴随算子：得到 $\mathcal{S}_{\mathscr{X}}^\top c=\sum_{t=1}^N c_t\mathscr{K}(x_t,\ )$。因此，考虑$Y=[y_1,...,y_N]^\top$和正则化参数 $\gamma>0$，我们有

$$f_z=\arg\min_{f\in\mathscr{H}}\frac{1}{N}\sum_{t=1}^N(y_t-f(x_t))^2+\gamma\|f\|^2_{\mathscr{H}} \tag{9}$$

$$=\left(\frac{1}{N}\mathcal{S}_{\mathscr{X}}^\top\mathcal{S}_{\mathscr{X}}+\gamma I\right)^{-1}\frac{1}{N}\mathcal{S}_{\mathscr{X}}^\top Y. \tag{10}$$

误差分析旨在量化 f_z 与 f_ρ之间的差异。为此，我们还考虑式（9）的无数据极限：

$$f_{\mathscr{H}}=\arg\min_{f\in\mathscr{H}}\int_{\mathscr{X}}(f(x)-f_\rho(x))^2d\rho_{\mathscr{X}}(x)+\gamma\|f\|^2_{\mathscr{H}} \tag{11}$$

$$=(L_{\mathscr{K}}+\gamma I)^{-1}L_{\mathscr{K}}f_\rho, \tag{12}$$

其中 $L_{\mathscr{K}}(f)(\bar{x})=\int_{\mathscr{X}}\mathscr{K}(\bar{x},x)f(x)d\rho_{\mathscr{X}}(x)$ 是一个积分算子，由于 $\mathscr{K}$ 的性质，它具有以下性质：(a) 自伴且严格正定；(b) 连续且紧致；(c) 满足谱定理 [13, 定理 4.3]，其特征对为 $\{(\varphi_i(\cdot)\ \lambda_i)\}_{i=1}^E$。得益于这些性质，对于任意 $\mathscr{L}^2_{\rho_{\mathscr{X}}}$函数 $f(x)=\sum_{q\in\mathbb{N}}\alpha_q\bar{\varphi}_q(x)$，利用基底的线性性和正交归一性，我们有

$$L_{\mathscr{K}}(f)(\bar{x})=\sum_{i=1}^E\lambda_i\alpha_i^\pi\varphi_i(\bar{x}), \tag{13}$$

其中，按照式（4）的思路，我们将向量 α^π for$i = 1, ..., E$ 的第 $i-$ 个分量定义如下：

$$\text{For } j = 1, ..., \frac{E}{2}, \quad \alpha_i^\pi = \begin{cases} \alpha_{q_j}^s, & i = j \\ \alpha_{q_j}^c, & i = j + E/2. \end{cases} \tag{14}$$

此外，得益于性质 (a)，我们还可以将积分算子[1] 的 r 次幂定义为 [12]:

$$L_{\mathscr{K}}^r(f)(\bar{x}) = \sum_{i=1}^{E} \lambda_i^r \alpha_i^\pi \varphi_i(\bar{x}). \tag{15}$$

下文将研究由数据集 $\mathcal{D}$有限所引入的样本误差 $\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$，以及由假设空间的选择决定的逼近误差 $\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$。这两者共同限定了总体误差的上界 $\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \le \|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} + \|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$；该上界需作为正则化参数 γ的函数进行最小化。

III. 样本误差

本节给出关于式 (10) 和 (12) 中引入的 f_z 与 $f_{\mathscr{H}}$ 之间误差的新结果。证明见附录 B。

定理 1：设假设 1 成立。考虑式 (6) 中引入的 $C_{\mathscr{K}}$，并定义 $\lambda = \min_{i=1,...,E} \lambda_i$。则以至少 $1 - \delta$ 的置信度，有

$$\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \le \frac{C_{\mathscr{K}}^3}{\gamma} \sqrt{\frac{B_f^2 + B_\sigma^2}{\breve{\lambda} N \delta}}. \tag{16}$$

□

注 1：我们没有研究 $\mathbb{E}_{\mathscr{Z}}[\rho_N(\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}]$的上界，因为它们通常会给出保守的数值。关于无界采样的结果，例如见 [28, 命题 20]。另请注意，根据 [29] 中给出的定义，我们的概率保证属于“可靠”上界，而非“精确”上界：这意味着，对于用户选择的置信水平 δ，结果成立的置信度为“至少 $1 - \delta$”，而非具有“精确概率$1 - \delta$”。

IV. 逼近误差

现在，我们研究由假设空间 $\mathscr{H}$ 的选择所导致的误差，即解 $f_{\mathscr{H}}$（在 (12) 中引入）与回归函数 f_ρ（在 (1) 中定义）之间的 $\mathscr{L}^2_{\rho_{\mathscr{X}}}$-距离。我们首先给出 $f_{\mathscr{H}}$ 的一个表达式：将回归函数表示为基函数的形式

[1]请注意，情形 $r = -1/2$ 在连接假设空间中函数的 $\mathscr{L}^2_{\rho_{\mathscr{X}}}$范数与 $\mathscr{H}$范数方面起着关键作用。事实上，考虑 $f(\cdot) = \sum_{i=1}^E \alpha_i \varphi_i(\cdot)$，根据 $\mathscr{H}$的定义可得$\|f\|^2_{\mathscr{H}} = \|\Sigma_\alpha^{-1/2}\alpha\|_2^2 = \sum_{i=1}^E \alpha_i^2/\lambda_i$。另一方面，我们有 $L_{\mathscr{K}}^{-1/2}(f)(\cdot) = \sum_{i=1}^E \alpha_i/\sqrt{\lambda_i}\varphi_i(\cdot)$，根据帕塞瓦尔定理，其 $\mathscr{L}^2_{\rho_{\mathscr{X}}}$-范数等于 $\sum_{i=1}^E \alpha_i^2/\lambda_i$ $\|f\|^2_{\mathscr{H}} =$。因此，$\|L_{\mathscr{K}}^{-1/2} f\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ 我们得到 。

对于 $\mathscr{L}^2_{\rho_{\mathscr{X}}}$，有 $f_\rho(\cdot) = \sum_{q\in\mathbb{N}} \bar{\alpha}_q \bar{\varphi}_q(\cdot)$；回顾式 (4) 中再生核希尔伯特空间基函数 $\varphi_i(\cdot)$ 的定义以及式 (14) 中系数 α_i^π 的定义，可得

$$f_{\mathscr{H}}(\cdot) = \sum_{i=1}^{E} \frac{\lambda_i}{\lambda_i + \gamma} \bar{\alpha}_i^\pi \varphi_i(\cdot). \tag{17}$$

借助这一结果，我们根据向量$\bar{\alpha}^\pi$ （在 (14) 中定义）的不同范数，推导出逼近误差的两个上界。关于它们性能的讨论留待第 VI-C 节。我们在以下命题中给出该结果，证明见附录 C。

命题 1：在第 II 节提出的三角线性回归框架中，逼近误差$\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$满足以下上界：

$$\text{(a)} \qquad \frac{\gamma}{\breve{\lambda} + \gamma} \|\bar{\alpha}^\pi\|_2 + \sqrt{\sum_{q\in\mathbb{N}\setminus Q} \bar{\alpha}_q^2} \tag{18}$$

$$\text{(b)} \qquad \|\bar{\alpha}^\pi\|_\infty \gamma \sum_{i=1}^{E} \frac{1}{\lambda_i} + \sqrt{\sum_{q\in\mathbb{N}\setminus Q} \bar{\alpha}_q^2}. \tag{19}$$

□

V. 偏差-方差权衡

本节结合定理 1 和命题 1 分别推导出的样本误差与逼近误差上界，并研究估计的总体误差 $\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$随正则化参数 γ变化的情况。在完全指定再生核希尔伯特空间 $\mathscr{H}$之后，即确定 Q 和 $\{\lambda_i\}_{i=1}^E$之后，我们进行分析。

主要结果见下述命题，该命题的证明见附录 D。

命题 2：

(a)考虑式 (18) 中的逼近误差界。若数据数量 N 和置信参数 δ 满足

$$\sqrt{N\delta} > \frac{C_{\mathscr{K}}^3}{\breve{\lambda}^{3/2}} \sqrt{\frac{B_f^2 + B_\sigma^2}{\sum_{i=1}^E (\bar{\alpha}_i^\pi)^2}}, \tag{20}$$

存在唯一的 $\gamma = \hat{}\gamma_{(a)}$ 使估计误差$\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$最小。

(b) 现在取式 (19) 中的逼近误差界。此时，总是存在唯一的 $\gamma = \hat{}\gamma_{(b)}$ 使估计误差$\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$最小。□

$\hat{}\gamma_{(a)}$ 和$\hat{}\gamma_{(b)}$ 的闭式表达式见证明。

VI. 讨论

我们首先将第三节给出的样本误差界与 [14] 和 [28]中给出的其他上界进行比较，以研究其性能。接下来，我们讨论命题 2 的结果，特别展示 $\gamma^{(b)}$ 捕捉使整体误差最小化的预言机 γ 行为的能力。

A. 与 *[14, 定理 5]*中的样本误差界比较

在数值设置中，我们假设均匀分布的随机噪声（信噪比为 150）会影响回归函数$f_\rho(x) = \sum_{q\in\mathbb{N}} \bar{\varphi}_q(x)\bar{\alpha}_q$ 的测量值，其中 $x \in [-1250, 1250]$。假设该函数由 20 对正弦/余弦对 $\{\bar{\varphi}_q\}$刻画，其中 q 是从集合 $\{1, ..., 30\}$中随机且不重复地抽取的。假设空间 $\mathscr{H}$由 $E/2 = 10$ 对正弦/余弦对的一个子集刻画，这些正弦/余弦对是从定义回归函数的正弦/余弦对中随机选取的。

我们进行了 500 次蒙特卡洛研究，每次均重新抽取一组基函数，用以定义回归函数和假设空间。回归函数的系数 $\bar{\alpha}_q$取自高斯分布$\mathcal{N}(0, \lambda)$，其中 λ从 $(0, 5)$ 上的均匀分布中抽取；此外，如式 (8) 所示，它也以 $\lambda_i = \lambda$的形式参与假设空间的定义，对所有 $i = 1, ..., E$均如此。每次运行中，数据点数 N 从集合 $\{100, 101, ..., 1000\}$ 中随机抽取。我们考虑置信水平 $\delta = 0.1$。然后，针对满足 [14, 定理 5], 中上界的 γ 最小值，评估相应的样本误差界，并计算其与该 γ下实际样本误差之间的相对差异。结果见图 1。两个上界均随 $1/\sqrt{N}$衰减，但（16）在置信水平方面表现得更为有利，至少当 δ 小于 $(0, 1]$,中 $1/\sqrt{\delta} = \log(4/\delta)$ 的解，即 ≈ 0.0539时是如此。[14, 定理5] 中的上界之所以较为保守，主要是因为它与输出值界 M呈线性关系。使上界（16）更为保守的关于 M 的显式条件如下：

$$M \le \frac{C_{\mathscr{K}}^2}{12}\sqrt{\frac{B_f^2 + B_\sigma^2}{\check{\lambda}\gamma}}\frac{1}{\sqrt{\delta}\log(4/\delta)}. \tag{21}$$

这类数值往往非常小：例如，在蒙特卡洛测试中，上界 (21) 得出的均值为 3.50 ± 2.02，而由（相当有利的）信噪比产生的真实值 M 的均值为 39.02 ± 14.66。

B. 与 *[28, 命题20]*中的样本误差界进行比较

我们考虑与上一节相同的数值设置，并利用马尔可夫不等式，将 [28, 命题 20]的上界转化为与定理 1 同类型的表述。为进一步适应第 II 节所述情形，我们令 $p = 2$和 $\mathcal{N}(\gamma) = \sum_{i=1}^{E} \lambda_i/(\lambda_i + \gamma)$。相较于 (16), [28, 命题 20] 的上界随数据量 N 增加而呈现出较慢的变化；此外，该上界还依赖于通常未知的逼近误差。我们通过将$\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ 设为其真实值进行了蒙特卡洛研究，结果非常保守，如图 1 所示。

C. 从偏差-方差权衡的角度看 γ 的选择

现在，我们进行一项蒙特卡洛研究，以讨论命题 2 给出的结果。考虑 $\mathscr{X} = [-5 \times 10^5, 5 \times 10^5]$

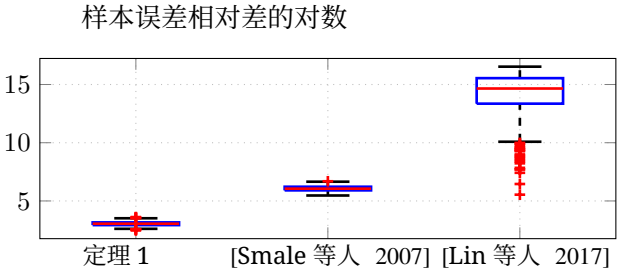


图 1：第 VI-A 节和第 VI-B 节蒙特卡罗试验中样本误差界的表现。采用的评分是上界与真实样本误差之差，再除以真实样本误差。对于定理 1，此类误差的平均值为 21.44 ± 4.093；对于 [14] 中的上界，该值为 440.03 ± 99.33；对于 [28] 中的上界，该值为 $3.40 \times 10^6 \pm 3.32 \times 10^6$。我们以对数刻度显示这些数值，以便于观察。

作为输入域。回归函数由 30 对正弦/余弦 $\{\bar{\varphi}\}_{q=1}^{30}$表征，其中每个 q 均从集合 $\{1, ..., 100\}$中随机选取且不重复，系数向量 $\bar{\alpha}$ 的每个分量均取自均值为零、方差为 λ =1 的高斯随机变量。后一个超参数也用于定义再生核希尔伯特空间 $\mathscr{H}$。频率集合 Q 是从表征回归函数的频率中随机选取的一个基数为 10 的子集。将信噪比固定为 50 后，我们抽取 50 个随机回归函数，并为假设空间选取基函数。每次运行的输入/输出对数量为 $N = 2500$，置信参数取 $\delta = 0.5$。对于每次运行，我们按照命题 2 计算 $\gamma^{(a)}$ 和 $\gamma^{(b)}$，按照定理 1 和命题 1 计算样本误差界和逼近误差界，并将其与由 $\gamma^{(a)}$ 和 $\gamma^{(b)}$得到的真实误差进行比较。我们观察到，使用 $\gamma^{(a)}$ 计算的误差界更接近真实值。相关数值列于表 I。

$\gamma_{(a)}$	真实值	上界
$\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$	0.036 ± 0.007	0.419 ± 0.033
$\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ (a)	9.313 ± 0.077	10.05 ± 0.992
$\gamma_{(b)}$	真实值	上界
$\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$	0.387 ± 0.076	14.04 ± 1.942
$\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ (b)	7.351 ± 0.575	20.41 ± 2.193

表 I：样本误差和逼近误差界相对于真实误差的总体数值（均值 ± 标准差）。

上述测试是在固定正则化参数的情况下进行的，重点考察单项误差。若改为考察总体误差$\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$，并将 $\gamma^{(a)}$ 和 $\gamma^{(b)}$ 的值与使其最小化的预言机值 γ^*（通过网格搜索获得）进行比较，我们会发现 $\gamma^{(b)}$ 的表现最佳。$\gamma^{(a)}$ 表现不佳，是因为要满足条件 (20) 需要大量数据，这会导致正则化参数被高估。在这项特定测试中，γ^* 取网格中的最小值，即 $\gamma^* = 0.1$；$\gamma^{(a)}$ 和 $\gamma^{(b)}$ 的均值分别为 7.7703 ± 0.2115 和0.2308 ± 0.0230。

VII. 结论

本文在统计学习框架下分析了正则化三角回归中出现的估计误差。据作者所知，文献中尚无此类研究，现有研究大多聚焦于非参数方法或非正则化三角回归。我们推导出了一个新的样本误差上界，该上界不要求输出分布的支撑集有限；数值测试表明，至少在非渐近情形下，它比经典上界更不保守。接着，我们计算了两个逼近误差上界，并将它们与样本误差界结合起来，从而得到正则化参数 γ的选择准则，以优化估计偏差与方差之间的权衡。具体而言，我们证明，其中一个准则给出的正则化参数值接近预言机值，因此原则上可用于加快超参数选择。我们强调，此类分析可以推广到 $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ 的任何其他正交基。此外，我们预计，这一框架的普适性将有助于对偏差学习进行抽象处理；这是本研究计划的后续拓展方向。未来的研究还将关注随数据量 N 和基函数 E变化的渐近行为。

参考文献

[1] G. Pillonetto、F. Dinuzzo、T. Chen、G. De Nicolao 和 L. Ljung，“系统辨识、机器学习与函数估计中的核方法：综述”，Automatica，第 50 卷，第 3 期，第 657–682 页，2014 年。[2] L. Hewing、K. P. Wabersich、M. Menner 和 M. N. Zeilinger，“基于学习的模型预测控制：迈向控制中的安全学习”，《控制、机器人与自主系统年度评论》，第 3 卷，第 1 期，第 269–296 页，2020 年。[3] T. Koller、F. Berkenkamp、M. Turchetta 和 A. Krause，“面向安全探索的基于学习的模型预测控制”，载于 2018 IEEE决策与控制会议（CDC），2018 年，第 6059–6066 页。[4] H. Liu、Y.-S. Ong、X. Shen 和 J. Cai，“当高斯过程遇上大数据：可扩展高斯过程综述”，IEEE神经网络与学习系统汇刊，第 31 卷，第 11 期，第 4405–4423 页，2020 年。[5] M. Lázaro-Gredilla 、J. Quiñonero-Candela、C. E. Rasmussen 和 A. R. Figueiras-Vidal，“稀疏谱高斯过程回归”，《机器学习研究杂志》，第 11 卷，第 1865–1881 页，2010 年。[6] A. Rudi 和 L. Rosasco，“使用随机特征进行学习的泛化性质”，载于神经信息处理系统进展，第 30 卷。Curran Associates, Inc.，2017 年。[7] A. E. Hoerl 和 R. W. Kennard，“岭回归：非正交问题中的有偏估计”，Technometrics，第 12 卷，1970 年，第 1 期，第 55–67 页。[8] A. B. Tsybakov，“非参数估计导论”，载于 Springer统计学丛书，2009 年。[9] Y. Pan、X. Yan、E. A. Theodorou 和 B. Boots，“稀疏谱高斯过程中不确定性下的预测及其在滤波与控制中的应用”，载于第 34 届国际机器学习会议论文集，PMLR 系列，第 70 卷，2017 年 8 月，第 2760–2768 页。[10] E. Arcari、A. Scampicchio、A. Carron 和 M. N. Zeilinger，“基于有限维模型的贝叶斯多任务学习：比较研究”，载于 2021 年第 60 届 IEEE决策与控制会议（CDC），2021 年，第 2218–2225 页。[11] E. Arcari、M. V. Minniti、A. Scampicchio、A. Carron、F. Farshidian、M. Hutter 和 M. N. Zeilinger，“面向机器人移动操作的贝叶斯多任务学习 MPC”，2022 年。 [在线]。网址：https://arxiv.org/abs/2211 10270 。[12] F. Cucker 和 S. Smale，“关于学习的数学基础”，美国数学会通报，第 39 卷，2002 年，第 1 期，第 49–55 页。[13] F. Cucker 和 D. X. Zhou，《学习理论：近似理论视角》，剑桥应用与计算数学专著系列。剑桥大学出版社，2007 年。 [14] S. Smale 和 D.-X. Zhou，“通过积分算子及其逼近进行学习理论估计”，《构造性逼近》，第 26 卷，第 153–172 页，2007 年。[15] A. Lunardi，《插值理论》，比萨高等师范学校出版物系列。比萨高等师范学校出版社，2009 年。[16] S. Boucheron、G. Lugosi 和 O. Bousquet，《集中不等式》。柏林、海德堡：施普林格柏林海德堡，2004 年，第 208–240 页。[17] Q. Wu、Y. Ying 和 D.-X. Zhou，“最小二乘正则化回归的学习速率”，《计算数学基础》，第 171–192 页，2006 年。[18] F. Cucker 和 S. Smale，“学习理论中正则化参数的最佳选择：偏差—方差问题”，《计算数学基础》，第 2 卷，第 413–428 页，2008 年 3 月。[19] S. Mendelson 和 J. Neeman，“ 核学习中的正则化 ”，《统计年鉴》，第 38 卷，第 1 期，第 526–565 页，2010 年。[20] C. Wang 和 D.-X. Zhou，“无界采样下最小二乘正则化回归的最优学习速率”，《复杂性杂志》，第 27 卷，第 1 期，第 55–67 页，2011 年。[21] A. Caponnetto 和 E. de Vito，“正则化最小二乘算法的最优速率”，《计算数学基础》，第 7 卷，第 331–368 页，2007 年。[22] C. Wang 和 D.-X. Zhou，“无界采样下最小二乘正则化回归的最优学习速率”，《复杂性杂志》，第 27 卷，第 1 期，第 55–67 页，2011 年。[23] Z.-C. Guo 和 D.-X. Zhou，“无界采样学习的集中估计”，《计算数学进展》，第 38 卷，第 207–223 页，2013 年。[24] D.-X. Zhou，“学习理论中再生核空间的容量”，《IEEE 信息论汇刊》，第 49 卷，第 1743–1752 页，2003 年 8 月。[25] N. Akhiezer 和 I. Glazman，《希尔伯特空间线性算子理论》，多佛数学丛书系列。多佛出版社，2013 年。[26] N. Aronszajn，“再生核理论”，《美国数学会汇刊》，第 68 卷，第 3 期，第 337–404 页，1950 年。[27] I. Steinwart 和 A. Christmann，《支持向量机》，第 1 版。施普林格出版公司，2008 年。[28] S.-B. Lin、X. Guo 和 D.-X. Zhou，“基于正则化最小二乘的分布式学习”，《机器学习研究杂志》，第 18 卷，第 92 期，第 1–31 页，2017 年。[29] P. L. Davies、A. Kovac 和 M. Meise，“非参数回归、置信区域与正则化”，《统计年鉴》，第 37 卷，第 2597–2625 页，2007 年。[30] P. Niyogi 和 F. Girosi，“基于分散噪声数据进行函数逼近的泛化界”，《计算数学进展》，第 10 卷，第 1 期，1999 年。[31] S. Berberian，《希尔伯特空间导论》。牛津大学出版社，1961 年。[32] G. Wahba 和 Y. Wang，《表示定理》。美国癌症协会，2019 年，第 1–11 页。[33] S. Smale 和 D.-X. Zhou，“香农采样 II：与学习理论的联系”，《应用与计算调和分析》，第 19 卷，第 3 期，第 285–302 页，2005 年。[34] ——，“香农采样与基于点值的函数重构”，《美国数学会通报》，第 41 卷，第 279–306 页，2004 年 7 月。[35] T. Zhang，“利用有效数据维数的核回归学习界”，《神经计算》，第 17 卷，第 9 期，第 2077–2098 页，2005 年 9 月。[36] B. Mu、T. Chen 和 L. Ljung，“关于基于核的正则化方法中超参数估计量的渐近性质”，《Automatica》，第 94 卷，第 381–395 页，2018 年。[37] W. Gilks、S. Richardson 和 D. Spiegelhalter，《实践中的马尔可夫链蒙特卡罗》，Chap man & Hall/CRC 跨学科统计丛书系列。泰勒-弗朗西斯，1995 年。[38] G. Pillonetto 和 A. Chiuso，“基于正则化核的回归与线性系统辨识中的复杂度调节：边际似然估计量的稳健性”，《Automatica》，第 58 卷，117，2015 年，106 页。–。[39] J. Maritz，《经验贝叶斯方法及其应用》。CRC 出版社，2018 年。

附录

我们在附录 A 中介绍统计学习框架下的误差分析。附录 B、C 和 D 分别给出第 III、IV 和 V 节中理论结果的证明。为便于查阅，第 VI 节中用作基准的定理陈述列于附录 E。附录 F 对 [22] 中报告的无界噪声支持下的样本误差界作了进一步比较。附录 G 展示了与第 VI-C 节实验相关的更多图表。附录 H 讨论正则化的优势。

A. 统计学习框架

设输入的度量空间 $\mathscr{X}$ 为紧致空间，输出空间为 $\mathscr{Y}=\mathbb{R}$。在 $\mathscr{Z}=\mathscr{X}\times\mathscr{Y}$ 上定义了概率测度 ρ，根据富比尼定理，它可分解为 $\rho(y|x)$ 和 $\rho_{\mathscr{X}}(x)$：对于可积函数 $\psi:\mathscr{Z}\to\mathbb{R}$，有

$$\int_{\mathscr{Z}}\psi(z)d\rho(z)=\int_{\mathscr{X}}\Big(\int_{\mathscr{Y}}\psi(x,y)d\rho(y|x)\Big)d\rho_{\mathscr{X}}(x).$$

假设在数据集 $\mathcal{D}=\{(x_t,y_t)\}_{t=1}^N$ 中收集从 ρ 抽取的 N 个独立样本。统计学习的目标是估计 ρ 的回归函数，其定义为

$$f_\rho(x)=\int_{\mathscr{Y}}yd\rho(y|x).$$

此函数是期望风险 $\mathcal{I}[f]=\mathbb{E}_\rho[\ y-f\ x\ \ ^2]=\int_{\mathscr{Z}}\ y-f_\rho\ x\ +f_\rho\ x\ -f\ x\ \ ^2d\rho=\int_{\mathscr{Z}}\ f_\rho\ x\ -f\ x\ \ ^2d\rho$ [30]

(( )) (( ) ( ) ( )) (( ) ( )) 的极小化函数，也可视为 ( ) 的第一个统计矩。其方差定义为

$$\sigma_\rho^2(x)=\int_{\mathscr{Y}}(y-f_\rho(x))^2d\rho(y|x),$$

其在 $\mathscr{X}$ 域上的积分为

$$\sigma_\rho^2=\int_{\mathscr{X}}\sigma_\rho^2(x)d\rho_{\mathscr{X}}(x)=\int_{\mathscr{Z}}(y-f_\rho(x))^2d\rho=\mathcal{I}[f_\rho].$$

注意，σ_ρ^2 表示最小化 $\mathcal{I}(f)$ 时不可避免的代价，并衡量 ρ 的条件是否良好：换言之，它类似于数值线性代数中的条件数概念 [12]。

假设回归函数属于某个目标空间 $\mathscr{T}$，但由于 ρ 未知，实际中无法计算。因此，在便于进行计算的假设空间 $\mathscr{H}$ 中寻找估计值 f_z。为此，选择 $\mathscr{H}$ 作为一个带有内积 $\langle\cdot,\cdot\rangle_{\mathscr{H}}$的再生核希尔伯特空间（RKHS），并通过求解以下蒂霍诺夫正则化问题来估计回归函数：

$$f_z=\arg\min_{f\in\mathscr{H}}\frac{1}{N}\sum_{t=1}^N(y_t-f(x_t))^2+\gamma\|f\|_{\mathscr{H}}^2.$$

再生核希尔伯特空间的关键特征在于，定义域 $\mathscr{X}$ 中任意点处的函数取值都由一个线性连续泛函良好定义。根据 Moore-Aronszajn定理 [26], ，再生核希尔伯特空间与一个正半定核算子一一对应（例如，参见 [13, 定义 2.8]）

$$\mathscr{K}:\mathscr{X}\times\mathscr{X}\to\mathbb{R}$$

使再生性质成立，即 $f(x)=\langle\mathscr{K}(x,\cdot)\ f\rangle_{\mathscr{H}}$.这些事实与里斯-弗雷歇定理（例如参见 [31][第 V 章，定理 1] ）共同导出了所谓的表示定理 [32], ，该定理指出，解 f_z 是 $\{\mathscr{K}(x_t,\cdot)\}_{t=1}^N$的线性组合，也就是说，是以 $\mathscr{X}$中给定输入位置为中心的核截面的线性组合。同一结果也可以通过采样算子 $\mathcal{S}_{\mathscr{X}}$ [33], [14]来表示：

引理 1：设 $\mathcal{S}_{\mathscr{X}}:\ \mathscr{H}\to\mathbb{R}^N$是一个算子，满足 $\mathcal{S}_{\mathscr{X}}(f)=[f(x_1)\ \ \dots\ \ f(x_N)]^\top$，并考虑其伴随算子 $\mathcal{S}_{\mathscr{X}}^\top:\ \mathbb{R}^N\to\mathscr{H}$，得到 $\mathcal{S}_{\mathscr{X}}^\top c=\sum_{t=1}^N c_t\mathscr{K}(x_t,\cdot)$。引入 $Y=[y_1\ \dots\ y_N]^\top$，可得

$$f_z=\Big(\frac{1}{N}\mathcal{S}_{\mathscr{X}}^\top\mathcal{S}_{\mathscr{X}}+\gamma I\Big)^{-1}\frac{1}{N}\mathcal{S}_{\mathscr{X}}^\top Y.$$

证明。我们先推导 $\mathcal{S}_{\mathscr{X}}^\top$ 的表达式。根据伴随算子的定义，必有 $\langle\mathcal{S}_{\mathscr{X}}f,c\rangle_2=\langle f,\mathcal{S}_{\mathscr{X}}^\top c\rangle_{\mathscr{H}}$。现在，$\langle\mathcal{S}_{\mathscr{X}}f,c\rangle_2=\sum_{t=1}^N c_tf(x_t)$，根据再生性质，这等于$\sum_{t=1}^N c_t\langle f,\mathscr{K}(x_t,\cdot)\rangle_{\mathscr{H}}=\langle f,\sum_{t=1}^N c_t\mathscr{K}(x_t,\cdot)\rangle_{\mathscr{H}}$。由伴随算子的定义可知，$\mathcal{S}_{\mathscr{X}}^\top c=\sum_{t=1}^N c_t\mathscr{K}(x_t,\cdot)$。

现在我们来求出 f_z 的表达式。将目标函数写成

$$\begin{aligned}&\frac{1}{N}\Big(\langle \mathcal{S}_{\mathscr{X}} f, \mathcal{S}_{\mathscr{X}} f\rangle_2 + \|Y\|_2^2 - 2\langle y, \mathcal{S}_{\mathscr{X}} f\rangle_2\Big) + \gamma\langle f, f\rangle_{\mathscr{H}}\\ =&\Big\langle\Big(\frac{1}{N}\mathcal{S}_{\mathscr{X}}^{\top}\mathcal{S}_{\mathscr{X}} + \gamma I\Big) f, f\Big\rangle_{\mathscr{H}} - \frac{2}{N}\langle \mathcal{S}_{\mathscr{X}}^{\top} y, f\rangle_{\mathscr{H}} + \frac{1}{N}\|Y\|_2^2.\end{aligned}$$

对最后一个表达式取泛函导数即可得到解。 ■

误差分析关注的首要问题是量化 f_z 对 f_ρ的逼近效果。为此，我们将 f_z 的数据自由极限定义为

$$f_{\mathscr{H}} = \arg\min_{f\in\mathscr{H}} \int_{\mathscr{X}} (f(x) - f_\rho(x))^2 d\rho_{\mathscr{X}}(x) + \gamma\|f\|_{\mathscr{H}}^2.$$

其解由积分算子给出

$$L_{\mathscr{K}}(f)(\bar{x}) = \int_{\mathscr{X}} \mathscr{K}(\bar{x}, x) f(x) d\rho_{\mathscr{X}}(x),$$

其表达式为（[12], II.2，定理 3；以及 III.6，命题 7）

$$f_{\mathscr{H}} = (L_{\mathscr{K}} + \gamma I)^{-1} L_{\mathscr{K}} f_\rho.$$

定义了 f_ρ、 f_z 和 $f_{\mathscr{H}}$后，考虑度量 $\|f_z - f_\rho\|_\natural$，其中 $\natural$ 表示所关注的范数类型。因此，总误差可分解为

$$\|f_z - f_\rho\|_\natural \le \|f_z - f_{\mathscr{H}}\|_\natural + \|f_{\mathscr{H}} - f_\rho\|_\natural.$$

第一项称为样本误差，表示由于使用的数据量有限而导致的再生核希尔伯特空间 $\mathscr{H}$中的误差。第二项称为逼近误差，源于假设空间的选择。在平方可积函数空间中考察误差时，这两项也分别称为方差和偏差。假设空间 $\mathscr{H}$的大小对它们的影响相反：假设空间越大， $f_{\mathscr{H}}$与 f_ρ 之间的距离可能越小；另一方面，模型越复杂，拟合它所需的数据就越多。

备注 2：统计学习视角最初与所谓的函数估计抽样理论方法并列提出 [34], [30]。前者可视为一种更灵活的误差分析框架，并将后者作为一个特例纳入其中。事实上，可以设想 f_ρ 是待估计的“真实”函数，并假设数据按 $y_t = f_\rho(x_t) + e_t$ 生成，且噪声的均值为零、方差为 $\sigma_\rho^2(x_t)$。更多评论请参阅 [33, 第7]节。

B. 定理 1 的证明

定义 $\xi_t:\ \mathscr{X} \to \mathscr{H}$使得 $\xi_t(\cdot) = (y_t - f_{\mathscr{H}}(x_t))\mathscr{K}(x_t, \cdot)$，则有 $\mathbb{E}_{\mathscr{Z}}[\xi_t](\cdot) = L_{\mathscr{K}}(f_\rho - f_{\mathscr{H}})(\cdot) = \gamma f_{\mathscr{H}}(\cdot)$。由此，并回顾采样算子的定义，可知 $f_z(x) - f_{\mathscr{H}}(x)$等于 [14]

$$\Big(\frac{1}{N}\mathcal{S}_{\mathscr{X}}^{\top}\mathcal{S}_{\mathscr{X}} + \gamma I\Big)^{-1}\left[\frac{1}{N}\sum_{t=1}^{N}\xi_t(x) - \mathbb{E}_{\mathscr{Z}}[\xi](x)\right].$$

现在我们可以研究上述表达式的$\mathscr{L}^2_{\rho_{\mathscr{X}}}$–范数。由于 $\mathscr{X}$是紧致的，且其上定义的测度 $\rho_{\mathscr{X}}$是概率测度，因此对于任意函数$f \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$都有$\|f\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \le \|f\|_\infty$：由此，$\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$的上界为

$$\Big\|\Big(\frac{1}{N}\mathcal{S}_{\mathscr{X}}^{\top}\mathcal{S}_{\mathscr{X}} + \gamma I\Big)^{-1}\Big\|_\infty \Big\|\frac{1}{N}\sum_{t=1}^{n}\xi_t - \mathbb{E}_{\mathscr{Z}}[\xi]\Big\|_\infty.$$

由于算子范数可由 $\frac{C_{\mathscr{K}}}{\gamma\sqrt{\breve{\lambda}}}$上界限定（证明见本小节末尾），现在我们可以研究 $\rho_N(\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} > \epsilon)$的上界；对于任意 $\epsilon > 0$，该上界为

$$\rho_N\left(\Big\|\frac{1}{N}\sum_{t=1}^{n}\xi_t - \mathbb{E}_{\mathscr{Z}}[\xi]\Big\|_\infty > \frac{\epsilon\gamma\sqrt{\breve{\lambda}}}{C_{\mathscr{K}}}\right). \tag{22}$$

在任意输入位置 $x \in \mathscr{X}$ 和给定的 $\bar{\epsilon} \in (0, 1)$ 下，由切比雪夫不等式可得 s

$$\rho_N\Big(\Big|\frac{1}{N}\sum_{t=1}^{N}\xi_t(x) - \mathbb{E}_{\mathscr{Z}}[\xi](x)\Big| > \bar{\epsilon}\Big) \le \frac{\mathrm{var}(\xi)(x)}{N\bar{\epsilon}^2},$$

注意到 $\{\xi_t\}_{t=1}^{N}$ 是独立同分布的。利用这一结果，我们可以进一步将 (22) 上界为

$$\rho_N(\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} > \epsilon) \le \frac{C_{\mathscr{K}}^2}{\gamma^2 \breve{\lambda}} \frac{\|\mathrm{var}(\xi)\|_\infty}{N\epsilon^2}. \tag{23}$$

方差项可上界为

$$\sup_{\bar{x}\in\mathscr{X}} \mathrm{var}(\xi)(\bar{x}) \le \sup_{\bar{x}\in\mathscr{X}} \int_{\mathscr{Z}} \mathscr{K}(\bar{x},x)^2 (y - f_{\mathscr{H}}(x))^2 d\rho \le C_{\mathscr{K}}^4 \int_{\mathscr{Z}} (y - f_{\mathscr{H}}(x))^2 d\rho \le B_f^2 + B_\sigma^2,$$

其中，最后一个不等式由以下事实得出：对于任意 :,

$$\int_{\mathscr{Z}} f\ x - y\ ^2 d\rho - \int_{\mathscr{Z}} f_\rho\ x - y\ ^2 d\rho = \|f - f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \qquad f \quad \mathscr{X} \to \mathscr{Y}(\,)(\,)(\,), \text{ 且}$$

[14] $f_{\mathscr{H}} \le \mathcal{J}\ 0$ 23

$\|f_{\mathscr{H}} - f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} + \gamma\|f_{\mathscr{H}}\|^2_{\mathscr{H}} = \mathcal{J}(\,)(\,) = \|f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \le B_f^2$。回到 (), 我们有

$$\rho_N\Big(\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} > \epsilon\Big) \le \frac{C_{\mathscr{K}}^6}{\gamma^2 \breve{\lambda}} \frac{(B_f^2 + B_\sigma^2)}{N\epsilon^2} = \delta. \tag{24}$$

从等式（24）中的 δ 求出 ϵ 的表达式，即可完成证明。 ■

算子范数界的证明：根据定义，我们要寻找一个常数 $\mathfrak{C}_\infty$，使得对于任意 $u \in \mathscr{H}$，

$\| \mathcal{S}_{\mathscr{X}}^\top \mathcal{S}_{\mathscr{X}}/N + \gamma I\ ^{-1} u\|_\infty \le \mathfrak{C}_\infty \|u\|_\infty$ $\Big\|\Big(\frac{1}{N}\mathcal{S}_{\mathscr{X}}^\top \mathcal{S}_{\mathscr{X}} + \gamma I\Big)^{-1} u\Big\|_\infty = \quad _{\bar{x}\in\mathscr{X}} \Big|\Big\langle \Big(\frac{1}{N}\mathcal{S}_{\mathscr{X}}^\top \mathcal{S}_{\mathscr{X}} + (\,)$。根据再生性质，sup ()(,)，再由 $\gamma I\Big)^{-1} u \cdot \ \mathscr{K}\ \bar{x} \cdot \Big\rangle_{\mathscr{H}}\Big|$ $C_{\mathscr{K}} \Big\|\Big(\frac{1}{N}\mathcal{S}_{\mathscr{X}}^\top \mathcal{S}_{\mathscr{X}} + \gamma I\Big)^{-1}\Big\|_{\mathscr{H}} \|u\|_{\mathscr{H}}$

柯西-施瓦茨不等式和 (6 $\mathscr{H}$ [14, 3 5], $\Big\|\Big(\frac{1}{N}\mathcal{S}_{\mathscr{X}}^\top \mathcal{S}_{\mathscr{X}} +)$，可进一步将

$\gamma I\Big)^{-1} u\Big\|_\infty \le \frac{C_{\mathscr{K}}}{\gamma} \|L_{\mathscr{K}}^{-1/2}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \|u\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \le \frac{C_{\mathscr{K}}}{\gamma} \|L_{\mathscr{K}}^{-1/2}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \|u\|_\infty$

其上界为。现在，根据方程中给出的算子范数上界，我们有。通过推导

$\|L_{\mathscr{K}}^{-1/2}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ $\|L_{\mathscr{K}}^{-1/2}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \le 1/\sqrt{\breve{\lambda}}$ $f \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ $\|L_{\mathscr{K}}^{-1/2} f\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} = \sqrt{\sum_{i=1}^{E} \frac{\alpha_i^2}{\lambda_i}} \le$ 的算子范数即可完成证明，这是因为对 $\sqrt{\frac{1}{\breve{\lambda}} \sum_{i=1}^{E} \alpha_i^2} \le \frac{1}{\sqrt{\breve{\lambda}}} \|f\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$

于任意的，。

C. 命题 1 的证明

将回归函数表示为式 (17) 中的 $f_\rho = \sum_{q\in\mathbb{N}} \bar{\alpha}_q \bar{\varphi}_q$ 和 $f_{\mathscr{H}}$，我们对 $\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ 应用三角不等式和帕塞瓦尔定理，得到

$$\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} = \Big\| \sum_{i=1}^{E} \frac{\lambda_i}{\lambda_i + \gamma} \bar{\alpha}_i^\pi \varphi_i - \sum_{q\in\mathbb{N}} \bar{\alpha}_q \bar{\varphi}_q \Big\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \le \sqrt{\sum_{i=1}^{E} \left(\frac{\gamma}{\lambda_i + \gamma}\right)^2 (\bar{\alpha}_i^\pi)^2} + \sqrt{\sum_{q\in\mathbb{N}\setminus Q} \bar{\alpha}_q^2}.$$

现在来看右侧的第一项。考虑 $(\lambda_i + \gamma)^{-1} \le (\breve{\lambda} + \gamma)^{-1}$，即可得到第一个上界 (18)。对于第二个上界，取 $\bar{\alpha}_i^\pi \le \|\bar{\alpha}^\pi\|_\infty$，将和的平方根上界为各项平方根之和，并取 $(\lambda_i + \gamma)^{-1} \le (\lambda_i)^{-1}$。

D. 命题 2 的证明

(a) 分别考虑式 (16) 和 (18) 中得到的样本误差和逼近误差。引入以下记号：

$$A = C_{\mathscr{K}}^3 \sqrt{\frac{B_f^2 + B_\sigma^2}{N\delta\breve{\lambda}}}, \quad b = \breve{\lambda}, \quad B = \sqrt{\sum_{i=1}^{E} (\bar{\alpha}_i^\pi)^2}, \quad C = \sqrt{\sum_{q\in\mathbb{N}\setminus Q} \bar{\alpha}_q^2}, \tag{25}$$

则整体误差可上界为：

$$\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \le \frac{A}{\gamma} + \frac{B\gamma}{b + \gamma} + C = F(\gamma). \tag{26}$$

函数 $F(\gamma)$ 对于 $\gamma > 0$ 始终为正。我们的目标是找出使 γ 存在唯一有限值且能够最小化 $F(\gamma)$ 的条件。为此，我们来研究一阶导数：

$$\frac{dF}{d\gamma} = 0 \longrightarrow \gamma^2(Bb - A) - 2Ab\gamma - Ab^2 = 0. \tag{27}$$

应用笛卡尔法则可知，确保正实轴上存在唯一根的条件是 $Bb - A > 0$，即式 (20)。该条件意味着在 $\gamma > 0$ 上存在唯一拐点：这是由以下事实得出的

$$\lim_{\gamma\to+\infty} F(\gamma) = B + C,$$

但也可以通过对 $d^2F(\gamma)\,d^2$ 应用笛卡尔法则来验证这一结论
最后，通过求解 (27) 可得到最优的 γ，其表达式如下：$\overline{\gamma}$.

$$\hat{\gamma}_{(a)} = \frac{b(A + \sqrt{ABb})}{Bb - A}.$$

(b) 我们沿用前述论证思路，但考虑如 (19) 所示的逼近误差界。考虑以下系数：

$$A \text{ as in (25)}, \qquad D = \sum_{i=1}^{E} \frac{\|\bar{\alpha}^{\pi}\|_{\infty}}{\lambda_i}, \tag{28}$$

通过证明函数 $F(\gamma) = \frac{A}{\gamma} + D\gamma$ 在 $\gamma > 0$处有唯一极小值，即可得出该结论。我们通过研究一阶和二阶导数，并利用 A和 D均为正这一事实来证明这一点。由此得到的最优 γ 始终存在，其取值如下：

$$\hat{\gamma}_{(b)} = \sqrt{\frac{A}{D}}.$$

E. 第 VI 节中基准定理的陈述

为便于引用，我们列出第 VI-A 节和第 VI-B 节中分别用到的 [14, 定理 5] 和 [28, 命题 20] 的陈述。

定理 2（*[14],* 定理 5）：设 ρ 几乎必然满足 $|y| \le M$。那么，对于任意 $0 < \delta < 1$

$$\|f_z - f_{\mathscr{H}}\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \le \frac{12 C_{\mathscr{K}} M \log(4/\delta)}{\sqrt{N\gamma}} \tag{29}$$

前提是

$$\gamma \ge \frac{8 C^2_{\mathscr{K}} \log(4/\delta)}{\sqrt{N}}. \tag{30}$$

□

命题 3（*[28],* 命题 20）：假设 $\mathbb{E}[y^2] < \infty$，且对于某个 $1 \le p \le \infty$，有 $\sigma_\rho^2 \in \mathscr{L}^p_{\mathscr{X}}$。此外，定义有效维数 $\mathcal{N}$（γ）$=$ Tr（$(L_{\mathscr{K}} + \gamma I)^{-1} L_{\mathscr{K}}$）[35]。那么，

$$\begin{aligned}\mathbb{E}_{\mathscr{Z}}[\|f_z - f_{\mathscr{H}}\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}] \le & (2 + 56C^4_{\mathscr{K}} + 57C^2_{\mathscr{K}})\Big(1 + \frac{1}{(N\gamma)^2} + \frac{\mathcal{N}(\gamma)}{N\gamma}\Big) \\ & \times \left\{ C^{\frac{1}{p}}_{\mathscr{K}} \sqrt{\|\sigma_\rho\|_p} \Big(\frac{\mathcal{N}(\gamma)}{N}\Big)^{\frac{1}{2}(1-\frac{1}{p})} \Big(\frac{1}{N\gamma}\Big)^{\frac{1}{2p}} + + C_{\mathscr{K}} \frac{\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}}{\sqrt{N\gamma}} \right\}.\end{aligned}$$

□

F. 与 *[22]* 中样本误差界的比较

现在，我们在另一种设定下开展另一项蒙特卡洛研究，其中噪声具有无界支撑。我们将上界 (16) 与以下内容进行比较：

定理 3（*[22],* 定理 1）：假设存在常数 $\tilde{M} > 0$ 和 $C > 0$，使得所谓的矩假设成立：

$$\int_{\mathscr{Y}} |y|^\ell d\rho(y|x) \le C\ell!\tilde{M}^\ell \qquad \forall \ell \in \mathbb{N},\, x \in \mathscr{X}. \tag{31}$$

此外，假设存在某个 $0 < \beta \le 1$ 和一个常数 $C_\beta > 0$，使得

$$\|f_{\mathscr{H}} - f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} + \gamma \|f_{\mathscr{H}}\|^2_{\mathscr{H}} \le C_\beta \gamma^\beta. \tag{32}$$

那么，如果核 $\mathscr{K}$在 $\mathscr{X} \times \mathscr{X}$上无限次可微，则对于任意 $0 < \varepsilon < 1$和 $0 < \delta < 1$，以置信度 $1 - \delta$，取 $\gamma = N^{\varepsilon - 1}$可得，

$$\|f_z - f_{\mathscr{H}}\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \le \tilde{C}_\varepsilon N^{\varepsilon - 1} \log(4/\delta)^{\frac{4}{\varepsilon} + 2}. \tag{33}$$

□

~ ~在介绍数值实验的细节之前，我们先推导 C、 M、 C_β 和 C_ε 的表达式。定理 3 是以下结果的推论：

˜定理 4（*[22],* 定理 2）：假设矩假设 (31) 满足常数 C 和 M；此外，设条件 (32) 在 $0<\beta\le 1$ 和常数 C_β 下成立。定义 $\mathscr{N}(\mathcal{B}_1,\ \eta)$ 为覆盖球 $\mathcal{B}_1=\{f\in\mathscr{H}$ 的半径为 η 的圆盘的最小数量：$\|f\|_{\mathscr{H}}\le 1\}$，并假设 $\mathscr{H}$ 的多项式复杂度指数为 $s>0$，即：

$$\log\mathscr{N}(\mathcal{B}_1,\eta)\le C_0\Big(\frac{1}{\eta}\Big)^s. \tag{34}$$

若 $0<\varepsilon<\frac{\beta}{1+s}$，则通过取 $\gamma=N^{\frac{\varepsilon}{\beta}-\frac{1}{s+1}}$，对于任意 $0<\delta<1$，以 $1-\delta$ 的置信度，我们有

$$\|f_z-f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}\le\tilde{C}_\varepsilon N^{\varepsilon-\frac{\beta}{s+1}}\Big(\log\Big(\frac{4}{\delta}\Big)\Big)^{\frac{\beta(1+\beta)}{(s+1)\varepsilon}+2}. \tag{35}$$

˜常数 C_ε 可按如下方式计算：

$$\tilde{C}_\varepsilon=\frac{C_5}{\varepsilon^2}C_2^{\frac{4\beta}{\varepsilon(s+1)}}\Big(1+\log\Big(1+\frac{2}{\varepsilon(s+1)}\Big)\Big)^{\frac{\beta(1+\beta)}{\varepsilon(s+1)}+2},$$

其中我们有

$$\begin{cases}C_5 &=38C_\beta+2(C_1+32^2(C+1)^2)C_4^2(2/(s+1))^2+480(C_{\mathscr{K}}+1)^2C_\beta\\ C_4 &=\tilde{M}(2C_{\mathscr{K}}(C+(1+2\sqrt{2C})+1))+C_3\\ C_3 &=\sqrt{38C_\beta}+(C_{\mathscr{K}}+1)\sqrt{480C_\beta}+\tilde{M}\\ C_2 &=\sqrt{2[C_1+32^2(C+1)^2]}\\ C_1 &=6C_{\mathscr{K}}+6C+8(1+\sqrt{2C})/\tilde{M}+520(C_{\mathscr{K}}+C+2(C+1))^2(C_0+1).\end{cases} \tag{36}$$

□

从上述结果得到定理 3 的陈述时，[22] 中的选择是，假设 f_ρ 属于假设空间 $\mathscr{H}$，并令 $\beta=1$ 和 $s=\frac{\varepsilon}{1-\varepsilon}$ 满足 $0<\varepsilon<1/2$（然后将 2ε 重新缩放为 ε）。现在我们的任务是˜求出常数 C、M、C_β 和 C_0的显式取值。$C_{\mathscr{K}}$的上界见式 (6)。

˜矩假设 (31) 的常数（C和 M）：根据文献 [22], 中的例 1，我们得到矩假设 (31) 对于 $C=4$ 和 $\tilde{M}=\max\{\sqrt{B_0},B_\infty\}$成立，其中 $B_0=B_\sigma^2$（见假设 1），且 $B_\infty\ge\|f_\rho\|_\infty$。注意，$B_\infty<\infty$，因为 f_ρ 属于 $\mathscr{H}$。

进入上界 (32) 的常数 (C_β)：我们在 $f_{\mathscr{H}}=(I+\gamma L_{\mathscr{K}}^{-1})^{-1}f_\rho$ 处计算代价 (11)。由此得到

$$\begin{aligned}\|f-f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}+\gamma\|L_{\mathscr{K}}^{-1/2}f\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}&=\|[(I+\gamma L_{\mathscr{K}}^{-1})^{-1}-I]f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}+\gamma\|L_{\mathscr{K}}^{-1/2}(I+\gamma L_{\mathscr{K}}^{-1})^{-1}f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}\\ &=\sum_{i=1}^E\Big[\Big(\frac{\lambda_i}{\lambda_i+\gamma}-1\Big)^2+\gamma\Big(\frac{\sqrt{\lambda_i}}{\lambda_i+\gamma}\Big)^2\Big](\bar{\alpha}_i^\pi)^2\\ &=\gamma\sum_{i=1}^E\frac{1}{\lambda_i+\gamma}(\bar{\alpha}_i^\pi)^2=\gamma\sum_{i=1}^E\Big(\frac{\lambda_i^\beta}{\lambda_i+\gamma}\Big)\lambda_i^{-\beta}(\bar{\alpha}_i^\pi)^2\end{aligned}$$

对于任意 β，属于 $(0,1]$。我们可以将 这样的表达式上界如下（完整推导见 [12] 中的第 II.2 章定理 3）：

$$\gamma\sum_{i=1}^E\Big(\frac{\lambda_i^\beta}{\lambda_i+\gamma}\Big)\lambda_i^{-\beta}(\alpha_i^\pi)^2\le\gamma\Big(\sup_\tau\frac{\tau^\beta}{\tau+\gamma}\Big)\|L_{\mathscr{K}}^{-\beta/2}f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}\le\gamma^\beta\|L_{\mathscr{K}}^{-\beta/2}f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}.$$

因此，通过对 $\|L_{\mathscr{K}}^{-\beta/2}f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ 取上界即可得到常数 C_β。在我们感兴趣的情形 $\beta=1$ 下，一个可能的解是 $C_{\beta=1}\le B_f^2/\breve{\lambda}$。

多项式复杂度对应的常数 (34) (C_0)：由于 (5) 中的 $\mathscr{K}$ 无限可微，已知对于任意 $s>0$，条件 (34) 均成立；然而，由于我们考虑的是满足 $0<\varepsilon<1/2$ 的 $s=\frac{\varepsilon}{1-\varepsilon}$，因此将关注范围限制在 $0<s<1$。

由于 $\mathscr{H}$ 是维数为 E 的有限维空间，根据定理 5.3 [13]，可得 $\mathscr{N}(\mathcal{B}_1,\eta)\le(1+2/\eta)^E$。对两边取对数，得到

$$\log\mathscr{N}(\mathcal{B}_1,\eta)\le E\log\Big(1+\frac{2}{\eta}\Big)\le 2EG(s)\Big(\frac{1}{\eta}\Big)^s.$$

现在我们需要求出 $G(s)$。其表达式总结如下。

引理2：考虑 $x>0$。为了使 $\log(1+x)<G(s)x^s$ 对 $0<s<1$成立，必须满足

$$G(s)>\Big(\frac{1-s}{s}\Big)^{1-s}.$$

证明。考虑函数 $g(x) = G(s)x^s - \log(1+x)$。我们希望它对 $x > 0$始终为正。由于 $g(0) = 0$，这相当于要求 $\frac{dg(x)}{dx}$ 始终为正。因此，考察一阶导数可得

$$\frac{dg(x)}{dx} = \frac{G(s)s(x+1) - x^{1-s}}{x^{1-s}(1+x)} > 0 \longrightarrow \frac{x}{(1+x)^{\frac{1}{1-s}}} < \Big(G(s)s\Big)^{\frac{1}{1-s}}.$$

该结论可通过对左侧项取最大值而得。 ■

现在可以进行数值测试了。我们对回归函数和数据点数量都进行随机化。前者的抽取方式与第 VI-A 节相同，而数据集基数 N 从集合 $\{300, 315, ..., 6990\}$中抽取。与第 VI-A 节一样，信噪比设为 150，但现在噪声服从高斯分布。用于控制 $\bar{\alpha}_q$ 采样和假设空间 $\mathscr{H}$ 的超参数 $\{\lambda_i\}_{i=1}^{20}$ 均设为 10。计算样本误差界的数值时，γ 按照定理 3 的陈述进行设置。对于每次蒙特卡洛运行，我们在网格 $\{0.05, 0.1, ..., 0.95\}$. 上使用 ε 计算这些上界。为了比较上界的数值，我们采用的评分是上界与真实样本误差之差除以真实样本误差。结果如图 2 所示。从 (33) 可以清楚看出，与 (16) 中给出的结果相比，关于数据 N 数量的收敛速率更好；然而，采用 γ 得到的数值极为保守。

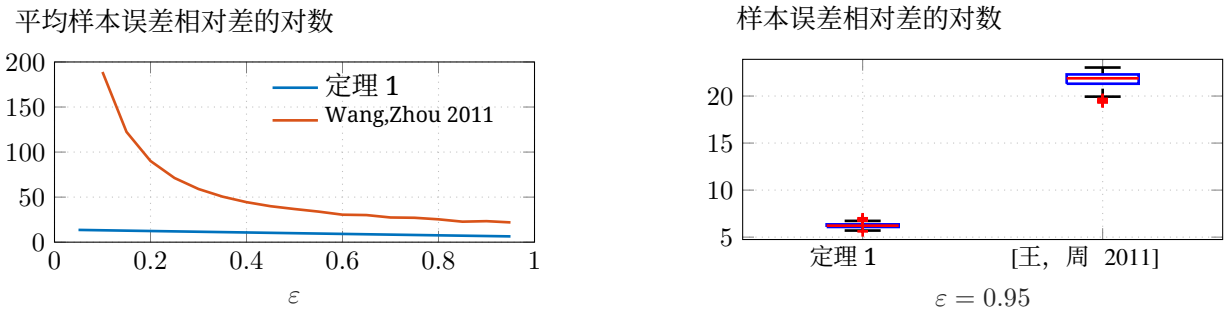


图 2：上界与真实样本误差之差相对于真实样本误差的归一化值的统计量，采用对数刻度。上图：蒙特卡洛迭代中均值随 ε 变化的情况。下图：$\varepsilon = 0.95$ 情况下各次蒙特卡洛运行的箱线图。定理 1 得出的值（采用对数刻度）为 6.238 ± 4.694，而使用定理 3 得出的值为 21.961 ± 21.421。

G. 第 VI-C 节测试的更多结果

我们回顾一下，蒙特卡洛测试所选取的数值如下：

- $\mathscr{X} = [-5 \times 10^4, 5 \times 10^4]$;
- $\delta = 0.5$;
- 回归函数 $f_\rho = \sum_q \bar{\alpha}_q \bar{\varphi}_q$ 由 30 对正弦/余弦对 $\{\bar{\varphi}\}_{q=1}^{30}$表征，其中每个 q 均从集合 $\{1, ..., 100\}$ 中随机选取，且不重复；
- $\bar{\alpha} \in \mathbb{R}^{30}$ 的所有分量均为独立样本，服从均值为零、方差为 $\lambda = 1$的高斯分布；
- 在假设空间中，对于所有 $i = 1, \quad \lambda_i = \lambda$....., E;
- Q 从刻画回归函数的频率集合中选取一个基数为 10 的随机子集；
- 信噪比 = 50;
- 在每次迭代中，我们随机抽取一个回归函数，并选取 $\mathscr{H}$的基函数；
- $N = 2500$.

我们进一步在图 3 中展示表 I 报告的结果。在每个箱线图中，我们考察上界与真实误差之差，并将其除以真实误差进行归一化。

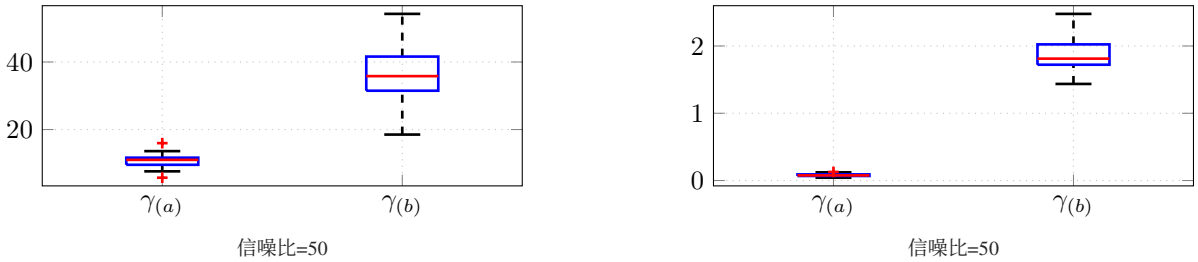


图 3：上界与真实值之间归一化相对差异的箱线图。左图：样本误差的结果；右图：近似误差的结果。这些结果补充了表 I 和第 VI-C 节中给出的结果。

H. 正则化的优势：加性噪声模型的情形

现在，我们讨论式（9）中提出的估计方案相较于不进行正则化时所能达到的方案的性能。我们在加性噪声模型的假设下进行分析，并将回归函数视为需要根据数据估计的“真实”函数。在此设定下，对于每个 $t = 1, ..., N$，测量模型为 $y_t = \phi^\top(x_t)\bar{\alpha}^\pi + r(x_t) + e_t$，其中第一项是回归函数在由式（4）定义并用于 $\phi^\top(\cdot)$ 的基函数 $\{\varphi_i(\cdot)\}_i\}_{i=1}^E$ 所张成的子空间上的投影；第二项汇集了未纳入该子空间的频率所产生的贡献，并假设其有界；最后一项是加性噪声，假设其为均值为零、方差 σ^2 已知的独立同分布噪声。需要求解的总体估计问题如下：

$$\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^E} \|Y - \Phi\alpha\|^2 + N\sigma^2 \alpha^\top P^{-1} \alpha, \tag{37}$$

其中，$\Phi \in \mathbb{R}^{N \times E}$ 将所有的 $\{\phi^\top(x_t)\}_{t=1}^N$ 堆叠在一起，Y 定义为式（10）中所用的量，而 P^{-1} 是一个正则化矩阵。解（37）可写为

$$\hat{\alpha} = (\Phi^\top\Phi + N\sigma^2 P^{-1})^{-1}\Phi^\top Y. \tag{38}$$

我们的目标是讨论在将 $P^{-1} = \frac{2\Sigma_\alpha^{-1}}{\sigma^2}$ 作为 γ 的函数来考察时，此类估计量的性能；这与式（9）直接相关。为此，我们将 $\mathcal{M}(\gamma) = \mathbb{E}_e[\|\hat{\alpha} - \bar{\alpha}^\pi\|^2]$ 视为性能指标，其中 $\mathbb{E}_e[\cdot]$ 表示对噪声分布求期望。采用上述特定的正则化项进行计算，并定义 $r = [r(x_1), ..., r(x_N)]^\top$ 和 $\tilde{\Sigma}_\alpha = \Sigma_\alpha / N$ 后，可得

$$\mathcal{M}(\gamma) = \mathrm{Tr}\left((\gamma\tilde{\Sigma}_\alpha^{-1} + \Phi^\top\Phi)^{-1} \left[\sigma^2\Phi^\top\Phi + \gamma^2\tilde{\Sigma}_\alpha^{-1}\bar{\alpha}^\pi(\bar{\alpha}^\pi)^\top\tilde{\Sigma}_\alpha^{-1} + \Phi^\top r r^\top \Phi + 2\gamma\tilde{\Sigma}_\alpha^{-1}\bar{\alpha}^\pi r^\top\Phi \right] (\gamma\tilde{\Sigma}_\alpha^{-1} + \Phi^\top\Phi)^{-1} \right). \tag{39}$$

注意，$\mathcal{M}(0) = \mathrm{Tr}((\Phi^\top\Phi)^{-1}[\Phi^\top r r^\top\Phi + \sigma^2\Phi^\top\Phi](\Phi^\top\Phi)^{-1})$。用 R 表示式（39）方括号中的表达式，可得

$$\frac{d\mathcal{M}(\gamma)}{d\gamma} = 2\mathrm{Tr}\left((\gamma\tilde{\Sigma}_\alpha^{-1} + \Phi^\top\Phi)^{-1} \left[\gamma\tilde{\Sigma}_\alpha^{-1}\bar{\alpha}^\pi(\bar{\alpha}^\pi)^\top\tilde{\Sigma}_\alpha^{-1} + \tilde{\Sigma}_\alpha^{-1}\bar{\alpha}^\pi r^\top\Phi - \tilde{\Sigma}_\alpha^{-1}(\gamma\tilde{\Sigma}_\alpha^{-1} + \Phi^\top\Phi)^{-1}R \right] (\gamma\tilde{\Sigma}_\alpha^{-1} + \Phi^\top\Phi)^{-1} \right).$$

考察其在 $\gamma \to 0^+$ 时的极限，可得

$$2\mathrm{Tr}\left((\Phi^\top\Phi)^{-1}\tilde{\Sigma}_\alpha^{-1} \left[\bar{\alpha}^\pi r^\top\Phi - (\Phi^\top\Phi)^{-1}(\Phi^\top r r^\top\Phi + \sigma^2\Phi^\top\Phi) \right] (\Phi^\top\Phi)^{-1} \right). \tag{40}$$

唯一容易研究的情形是不存在残差项（即回归函数属于假设空间，因此对于所有 x，$r(x) = 0$）。在这种情况下，上述表达式显然为负，这证明了至少在原点的某个小邻域内，$\mathcal{M}(\gamma) < \mathcal{M}(0)$：另见 [36] 中的命题 2。

为数值检验正则化的影响，我们进行了包含 500 次试验的蒙特卡洛研究。我们考虑输入域 $\mathcal{X} = [-25, 25]$，以及一个由 50 个基函数刻画的回归函数，这些基函数从前 80 个基函数中随机选取（按照式 (3) 中 q 递增的顺序排列），并由一个从均值为零、分量独立同分布且方差为 $\lambda = 10$ 的高斯分布中抽取的向量进行线性组合。后一个超参数也用于定义 $\mathscr{H}$；定义该量的频率集合 Q 是一个随机维数为 $E/2$（属于 $\{5, 6, ..., 50\}$）的样本，从定义 f_ρ 的频率中选取。因此，在这一测试情形下，由残差 $r(\cdot)$ 决定的逼近误差不为 0。我们考虑信噪比为 100 的情形，噪声服从高斯分布、均值为零且独立同分布。每次运行时，数据数量 N 都在区间 $\{5, ..., E/2\}$ 内随机选取。对于正则化参数，我们既采用命题 2(b) 中的 $\gamma^{(b)}$，也采用通过边际似然优化估计得到的值 $\hat{}\gamma$；该优化使用 Gibbs 采样方案，并利用了式 (37) 中问题的贝叶斯解释（[37,第 1] 章；关于边际似然超参数估计稳健性的详细讨论，另见 [38]。更多细节见下文）。随后，我们研究整体误差 $\mathcal{M}(\gamma) = \|f_\rho - f_z\|_{\mathscr{L}^2_{\rho_\mathscr{X}}}$ 的取值，其中 $f_z(\cdot) = \phi^\top\hat{}\alpha$，$\hat{}\alpha$ 如式 (38) 所示；并将 $\gamma^{(b)}$、$\gamma = 0$ 和 $\hat{}\gamma$ 所达到的误差与通过网格搜索计算 $\gamma = \gamma^*$ 所得到的预言机值进行比较。结果汇总于图 4。我们注意到，两种正则化估计量的结果相当，这意味着 $\gamma^{(b)}$ 是 $\hat{}\gamma$ 的良好估计量（且计算速度更快）；此外，两者的表现始终优于未使用正则化的情形。具体而言，例如，使用 $\gamma = 0$ 和 $\hat{}\gamma$ 时的相对偏差中位数为 24.99%，最小值和最大值分别为 0.64% 和 66.97%。

最后，我们给出蒙特卡洛测试中用作基准的超参数 γ 选择规则的详细说明。考虑式 (37) 中的估计问题，即

$$\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^E} \|Y - \Phi\alpha\|^2 + N\sigma^2\alpha^\top P^{-1}\alpha = (\Phi^\top\Phi + N\sigma^2 P^{-1})^{-1}\Phi^\top Y, \tag{41}$$

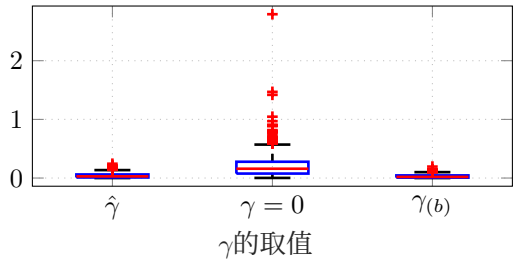


图 4：正则化和非正则化情形下的相对偏差 100%($\bar{\mathcal{M}}(\hat{\gamma}) - \bar{\mathcal{M}}(\gamma^*))/\bar{\mathcal{M}}(\gamma^*)$。

其中， $\Phi \in \mathbb{R}^{N\times E}$ 是由所有 $\{\phi^\top(x_t)\}_{t=1}^N$ 叠合而成的矩阵（即包含基函数 $\{\varphi_i(\cdot)\}_{i=1}^E$ 的行向量，如式 (4) 所定义）， P^{-1} 是正则化矩阵，而 Y是输出测量向量 $[y_1, ..., y_N]^\top$。
(41)中的目标函数具有一种随机解释。考虑测量模型 $Y = \Phi\alpha + E$，其中 E 是一个均值为零、协方差已知的N维高斯向量 $\sigma^2 I_N$，而 $\alpha \in \mathbb{R}^E$未知。从贝叶斯视角来看， α 被建模为一个随机向量；具体而言，假设其服从均值为零、协方差为 P/N的高斯分布。在这种设定下，(41)中的目标函数（除去不依赖于 α的常数）是后验概率 $\alpha|Y$的负对数：也就是说，$\hat{}\alpha$ 被计算为最大后验（MAP）估计；在我们所考虑的高斯情形下，它对应于最小方差线性估计。事实上，(41)的解确实就是后验均值的表达式。

在这一框架下，假设 P^{-1} 或 σ^2 中涉及的（部分）超参数未知，并将它们汇集为一个向量 η。一种可行的策略是利用所谓的经验贝叶斯方法 [39]从数据中估计这些参数：具体而言，通过最大化证据 $Y|\eta$ （或者更方便地，最小化其负对数）来计算 η 的估计值，这等价于对联合分布 $Y, \alpha|\eta$，其中已对 α 的依赖进行积分消去。在高斯情形下，这可写为

$$\hat{\eta} = \arg\min_{\eta}\ Y^\top \Sigma_Y(\eta)^{-1} Y + \log\det\Sigma_Y(\eta), \tag{42}$$

其中 $\Sigma_Y = \Phi P \Phi^\top/N + \sigma^2 I_N$。这是一个非凸问题，确定性优化程序可能会因对初始条件敏感而返回不可靠的结果。解决这一问题的一种方法是采用马尔可夫链蒙特卡罗（MCMC）。具体而言，其思路是运行一个（单分量）Metropolis-Hastings 算法，构造一个不变分布与所关注的边际似然成正比的马尔可夫链：这是一种从该分布中抽取样本并以基于样本的形式求解（42）的方法。关于马尔可夫链蒙特卡罗的介绍，请参阅 [37]。

现在，我们将 (41) 与 (9) 中给出的原始函数估计问题联系起来，以便为 P^{-1} 给出一个表达式，并详细说明基于 MCMC 的边际似然优化流程。函数估计为

$$f_z = \arg\min_{f\in\mathscr{H}} \frac{1}{N}\sum_{t=1}^{N}(y_t - f(x_t))^2 + \gamma\|f\|_{\mathscr{H}}^2.$$

根据定义 (7) 和 (5) 所指定的再生核希尔伯特空间 $\mathscr{H}$，可以写成 $f_z(\cdot) = \phi^\top(\cdot)\hat{}\alpha$。这样，估计 $f_z(\cdot)$ 就转化为在特定选择 $P^{-1} = \frac{\gamma\Sigma_\alpha^{-1}}{\sigma^2}$ 下求解 (41)，其中 Σ_α 由假设空间的选择给出（例如，参见 (5)），而 γ 是待调节的正标量。

现在，我们详细说明上述情形下的 MCMC 流程。首先，将感兴趣的概率密度函数记为 $p(\cdot)$，注意到

$$p(\alpha, \gamma|Y) = p(\alpha|\gamma, Y)p(\gamma|Y) \propto p(\alpha|\gamma, Y)p(Y|\gamma),$$

这表明，从 $\alpha, \gamma|Y$ 中抽取样本是探索边际似然的一种合适方式。因此，我们构建一个吉布斯采样器（单分量 Metropolis-Hastings 算法的一种特例），其不变分布的密度为 $p(\alpha, \gamma|Y) = \pi(\alpha, \gamma)$。在这一特定情形下，全条件分布的样本很容易计算：事实上，记 $\tilde{}$ 为 $\Sigma_\alpha = \Sigma_\alpha/N$，并令 $\Gamma(\mathfrak{a}, \mathfrak{b})$ 表示均值为 $\mathfrak{a}/\mathfrak{b}$ 的伽马分布，利用共轭分布的性质可得

$$\pi(\alpha|\gamma) \longleftarrow \alpha|\gamma, Y \sim \mathcal{N}\Big((\tilde{\Sigma}_\alpha^{-1} + \Phi^\top\Phi/\gamma)^{-1}\Phi^\top Y/\gamma,\ (\tilde{\Sigma}_\alpha^{-1} + \Phi^\top\Phi/\gamma)^{-1}\Big) \tag{43}$$

$$\pi\ (\gamma|\alpha)\ \longleftarrow \gamma|\alpha,\ \ Y \sim \Gamma\Big(\frac{N}{2},\ \frac{\|Y - \Phi\alpha\|^2}{2}\Big). \tag{44}$$

最后，可以使用吉布斯采样器获得的最后 $N_g < N_G$ 个样本来计算使边际似然最大化的 $\hat{}\gamma$ 。