投稿至《统计年鉴》

具有极小极大最优性的非参数回归个性化黑盒模型

BY S人工智能 LI[1,a]，和 LINJUN ZHANG[2,b]

[1]统计系　　学与数据科学，清华大学，[a]saili@tsinghua　　.edu.cn

[2]罗格斯大学统计学系，[b]linjun.zhang@rutgers.edu

近年来，包括深度神经网络和大语言模型在内的大规模模型取得了长足进展，显著提升了其在广泛学习任务中的性能。此类预训练模型的广泛可用性为数据高效的统计学习创造了新的机遇，前提是能够将其有效整合到下游任务中。在这一背景下，我们研究了少样本个性化问题，即使用有限数量的样本，将预训练黑盒模型适配到目标域。我们针对非参数回归中的少样本个性化建立了理论框架，并提出了能够将黑盒预训练模型纳入回归过程的算法。我们确立了个性化问题的极小极大最优速率，并证明所提方法能够达到该速率。我们的结果阐明了在样本稀缺情况下利用预训练模型的统计优势，并在预训练模型不具备信息性的情况下提供稳健性保证。我们通过模拟以及使用多个预训练模型对加利福尼亚住房数据集的应用，展示了这些方法在有限样本下的性能。

arXiv:2601.01432v1 [stat.ME] 4 Jan 2026

1. 引言。近年来，大规模黑盒模型显著提升了各类任务的预测性能。强大的预训练黑盒模型为高效利用数据进行学习创造了新的机遇，前提是能够将这些模型有效且严谨地整合到下游应用中。这进而推动了将预训练模型纳入学习流程的原则性统计框架的发展。

在异质环境中利用预训练模型，一个核心挑战是个性化：利用有限的有标签数据，将在某一人群或任务上训练的大规模模型适配到新的人群或任务。少样本个性化旨在仅使用少量目标样本，将黑盒预训练模型适配到新的用户或任务 [14]。例如，NeuralCVD [29] 是一种基于神经网络的模型，利用英国生物样本库的数据训练，用于预测心血管风险轨迹。若不经适应便将其应用于不同人群，其性能可能会大幅下降。然而，由于数据稀缺或计算约束，从头开始重新训练这类模型往往不可行。在这种情况下，利用有限的目标观测值适配现有预训练模型，是一种自然且高效的替代方案。

形式化地，令 $y \in \mathbb{R}$ 表示响应变量，$\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ 表示协变量。令 $f^{(\mathrm{ptr})}: \mathcal{X} \to \mathbb{R}$ 表示从外部来源训练得到的预训练模型。我们的目标是构建一个针对目标分布 $P^*(\boldsymbol{x}, y)$ 的个性化预测器，将 $f^{(\mathrm{ptr})}$ 与从目标总体中抽取的少量样本相结合，同时无需访问 $f^{(\mathrm{ptr})}$ 的内部机制。当预训练模型为目标任务提供了有用信号时，理想的做法是借助 $f^{(\mathrm{ptr})}$ 的优势，并调整源分布与目标分布之间的差异。我们的框架允许预训练模型是一个黑

MSC2020 学科分类：主要分类 62G05、68T05；次要分类 62C20。关键词和短语：个性化、非参数回归、极小极大最优性。

盒模型，并将其预测作为输入。我们无需访问模型参数、架构，也不需要预先了解其与目标问题的相关性。这种设定反映了实际场景：用户希望借助专有模型（如 GPT-4 和 Gemini）改进小样本数据分析，但无法访问其内部机制。即使是开源模型（如 Llama），直接微调在计算上仍可能难以承受。因此，个性化为整合大规模黑盒模型提供了一种计算效率更高的替代方案。

我们研究非参数回归场景中的个性化。给定协变量向量$\boldsymbol{x} \in \mathcal{X}$，目标总体中的响应满足

$$y = f^*(\boldsymbol{x}) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid \boldsymbol{x}] = 0, \tag{1}$$

其中，ε 表示一个条件方差为$\mathrm{Var}(\varepsilon \mid \boldsymbol{x}) = \sigma^2(\boldsymbol{x})$的噪声变量，允许存在异方差性。未知的回归函数 $f^* : \mathcal{X} \to \mathbb{R}$ 是我们关注的目标。一般而言，$f^*(\cdot)$ 不一定与预训练模型 $f^{(\mathrm{ptr})}(\cdot)$ 一致，并且可能与之存在显著差异。因此，利用目标数据进行适应对于将 $f^{(\mathrm{ptr})}$ 校准至目标总体至关重要。由于收集带标签的目标样本通常成本高昂，我们假设采样预算有限，规模为 n，并提出一种专门设计用于促进个性化的样本收集方案。

总而言之，个性化流程的输入包括一个预训练模型 $f^{(\mathrm{ptr})}$、一个目标协变量定义域 $\mathcal{X}$以及一个采样预算 n。基于 n 由（1）生成的带标签观测值，我们开发算法来构造 $f^*(\cdot)$的个性化估计量。本框架在其他场景下的扩展将在后文讨论。

1.1. 与相关工作的联系。少样本个性化在大语言模型 [14] 和联邦学习 [33], 领域日益重要，其共同目标是使预训练模型适应不同用户或任务。然而，目前对其统计性质的理解有限。统计意义上的最优方法在很大程度上仍属未知，极小极大最优收敛率也尚未得到研究。我们首先回顾一些相关主题，并着重介绍少样本个性化所面临的独特挑战。

为整合外部信息，迁移学习和领域泛化是常用方案，它们利用源数据提升目标数据上的学习性能。近期许多研究探讨了适用于非参数回归[6, 24], 和高维参数模型 [18, 30, 20] 等场景的迁移学习方法。领域泛化是另一种分布外预测范式，其重点是利用多个源领域训练预测规则，使其能够在未见过的目标领域上表现良好。当前最先进的方法在因果模型和机器学习模型中研究不变预测规则 [23] 。在因果结构模型下，[23, 5, 26] 提出了适用于线性模型的不变方法。在不作因果假设的情况下，[3] 提出领域不变投影方法，学习在不同定义域间保持不变的协变量变换。 [25] 提出不变风险最小化框架，旨在发现多个训练环境间的不变表示，同时排除伪特征。[10] 在无限样本条件下研究了用于领域泛化的不变最小二乘方法。除不变性框架之外， [19] 还研究了回归系数可组织为低秩张量时的领域泛化。总之，在迁移学习和领域泛化中，用户可以访问个体级源数据，从而构建理想的预训练模型，以实现最佳的知识迁移。相比之下，个性化将预训练模型视为黑盒，无法访问源数据，并明确设置了样本收集阶段。

近年来，预测驱动推断（PPI）框架[1, 2] 被提出，用于在半监督设定下结合任意预训练模型进行统计推断。在PPI框架中，$f^{(\mathrm{ptr})}(\boldsymbol{x})$ 可被视为用于推断总体层面参数[12]的替代变量。PPI与个性化之间有两个关键区别。首先，

PPI的目标是总体均值，而不是回归函数或个体预测。其次，PPI考虑的是半监督设定：将预训练模型与少量带标签样本及更多无标签样本相结合。假设这些样本是从目标分布中随机收集的。相比之下，在个性化问题中，我们只利用少量带标签数据。此外，带标签数据的协变量分布由用户指定，可能不同于目标协变量分布。一项更为相关的工作是 [34]，该工作研究了如何在PPI设定下选择需要标注的样本，但其重点仍然是对总体均值进行推断。我们提出了一种针对个性化问题的不同抽样规则，并为所提出的方法建立了理论保证。

近期关于大语言模型（LLMs）个性化的研究开始探索如何根据个体用户的偏好、写作风格和历史记录定制模型输出 [32, 21]。在方法层面，[27] 通过实证表明，与通用大语言模型相比，检索增强生成和微调（或提示）能够显著提升面向单个用户的文本分类和生成任务的个性化效果。近期的算法进展还提出了学习用户专属提示 [15] 或用户专属奖励 [28] 的方法，仅利用少量用户反馈，就能引导大语言模型的输出符合个体偏好。尽管相关研究日益丰富，这些工作主要侧重于实证或算法，缺乏对有限数据条件下适应性的正式统计分析，也未提供极小极大最优性保证。

1.2. 主要结果。本文提出了一个用于非参数回归个性化的统计框架和高效算法。我们的贡献有两方面。

首先，我们提出一种个性化流程，通过局部平滑将黑盒预训练模型融入非参数回归估计量。这一步对于在保持统计效率的同时实现对模型误设的稳健性至关重要。此外，我们还引入了一种样本检索方案，旨在异方差噪声条件下提高抽样效率。这些组件共同为将大规模预训练模型融入经典非参数推断提供了一种有原则的方法。

在理论方面，我们刻画了个性化非参数回归的极小极大最优速率，并证明所提估计量在温和条件下能够达到这一速率。我们的结果从极小极大的意义上刻画了预测准确度对目标回归函数 Hölder 范数的依赖关系。分析揭示了利用预训练模型如何降低估计问题的赫尔德复杂度，并在某些情形下提高有效光滑阶数，从而相较于仅使用目标样本进行估计获得更快的收敛速率。重要的是，所提方法具有无害保证：在温和条件下，无论预训练模型的相关性如何，都可以证明其性能不会差于仅基于目标样本的非参数估计量。据我们所知，这是最早一批为个性化提供严格保证的极小极大最优统计处理研究之一。

1.3. 结构与符号。 本文其余部分安排如下。第 2 节介绍用于非参数回归的所提少样本个性化方法。第3 节给出理论分析和极小极大最优收敛率。第 4 节将该框架扩展至可获得目标域额外无标签协变量的情形。第 5 节报告模拟研究，并将所提方法与现有方法进行比较；第 6 节将该方法应用于真实数据分析，预测加利福尼亚州房价。证明和技术细节见补充材料。

对于集合 $\mathcal{S} \in \mathbb{R}^d$，令 $|\mathcal{S}|$ 表示其勒贝格测度。令 $a_n = O(b_n)$ 和 $a_n \lesssim b_n$ 表示当 n 足够大时关于某个常数 c 的 $|a_n/b_n| \le c$。令 $a_n = o(b_n)$ 和 $a_n \ll b_n$ 表示当 $n \to \infty$ 时的 $a_n/b_n \to 0$。我们使用 C、C_0、C_1、……、c、c_0、c_1、…… 表示通用常数，这些常数在不同陈述中可能不同。

2. 非参数函数的个性化估计。本节介绍我们所提方法，用于将预训练模型 $f^{(\mathrm{ptr})}(\cdot)$ 个性化，以在定义域 $\mathcal{X} \subseteq \mathbb{R}^d$ 上估计目标回归函数 $f^*(\boldsymbol{x})$。我们首先定义所研究的函数类，然后介绍完整的流程。

定义 1（Hölder 类）。对于 $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$，定义在 $H(\boldsymbol{\theta})$ 上的 Hölder 类 $\mathbb{L} \subseteq \mathbb{R}^d$，其中某些有限的 $\theta_1 \geq 0$ 和 $0 < \theta_2 \leq 1$，定义为满足以下条件的函数集合 $f\colon \mathbb{L} \to \mathbb{R}$，对任意 $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{L}$，

$$|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)| \leq \theta_1 \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^{\theta_2}.$$

若 $f \in H(\boldsymbol{\theta})$，则称 $\boldsymbol{\theta}$ 为 $f(\cdot)$ 的 Hölder 类参数。

在定义 1中，我们关注光滑阶为 $\theta_2 \leq 1$ 的 Hölder 类，这类在非参数回归文献 [9, 31] 和迁移学习文献 [6]中都很常见。不同于经典理论假设 Hölder 范数 θ_1 固定不变，我们只要求 θ_1 有限，并允许其随着样本量 $n \to \infty$趋于零。这种灵活性使我们能够更精细地刻画光滑度参数如何影响收敛速率，并在个性化分析中发挥重要作用。

全文中，我们假设目标回归函数满足 $f^*(\boldsymbol{x}) \in H(\boldsymbol{\theta}^*)$ 在 $\mathcal{X}$上，对某个有限的 $\theta_1^* > 0$ 和 $0 < \theta_2^* \leq 1$成立。为简单起见，我们取 $\mathcal{X} = [0,1]^d$，而一种扩展将在第4.1节中研究。对于风险准则，我们将一般函数 $f\colon \mathcal{X} \to \mathbb{R}$ 的均方积分误差定义为

$$\mathrm{MISE}(f) = \mathbb{E}\left[\int_{\mathcal{X}} \{f(\boldsymbol{x}) - f^*(\boldsymbol{x})\}^2 d\boldsymbol{x}\right], \tag{2}$$

这是经典非参数文献中的一种标准度量 [31]。

2.1. 主要步骤概述。我们首先概述所提出的个性化流程。该方法包含三个主要步骤。

- 步骤 1：样本检索。我们首先根据第2.4节所述的数据收集规则，检索 n 协变量点 $\boldsymbol{x}_i \in \mathcal{X}$，$i = 1, \ldots\ldots, n$。对于每个选出的 $\boldsymbol{x}_i$，我们收集由目标模型（1）生成的相应响应 y_i。

- 步骤 2：平滑偏差校正。我们通过估计预训练模型 $f^{(\mathrm{ptr})}(\boldsymbol{x})$的偏差函数 $\delta(\boldsymbol{x}) = f^*(\boldsymbol{x}) - f^{(\mathrm{ptr})}(\boldsymbol{x})$。由于 $\delta(\boldsymbol{x})$不一定平滑，直接对其进行估计可能不稳定。因此，我们首先对预训练模型应用一种称为$\boldsymbol{\theta}$局部平滑的局部平滑操作，该操作由调优参数 $\boldsymbol{\theta}$索引。然后，我们ˆ构造一个基于核的估计量来估计所得偏差，记为 $\delta_{\boldsymbol{\theta}}(\boldsymbol{x})$。相应的个性化估计量为

$$\hat{f}_{\boldsymbol{\theta}}^{(\mathrm{fsp})}(\boldsymbol{x}) = f^{(\mathrm{ptr})}(\boldsymbol{x}) + \hat{\delta}_{\boldsymbol{\theta}}(\boldsymbol{x}).$$

- 步骤 3：适应。我们使用验证样本选择调优参数 $\boldsymbol{\theta}$，ˆ并输出最终的小样本个性化估计器

$$\hat{f}^{(\mathrm{fsp})}(\boldsymbol{x}) = \hat{f}_{\hat{\boldsymbol{\theta}}}^{(\mathrm{fsp})}(\boldsymbol{x}).$$

该算法的核心思想是以经济高效且具有统计效率的方式纠正预训练模型的偏差。具体而言，所提出的样本检索规则旨在噪声具有异方差性时获取更重要的样本。由于预训练模型的黑盒特性，局部平滑步骤确保我们的方法能够抵御

对抗性预训练模型。通过在步骤 3 中选择最优调优参数，所提出的方法能够自动适应预训练模型的最佳使用方式。

我们将在以下各节中详细介绍每个步骤。在第 2.2 节和 2.3节中，我们将在通用抽样方案下介绍偏差校正和适应步骤。在此风险分析的基础上，我们随后将在第 2.4节中介绍所提出的样本检索规则，该规则近似于最优抽样策略。本节末尾的算法 2 总结了完整算法。

2.2.平滑偏差校正。 假设我们已经获得了 n 带标签样本$\{(\boldsymbol{x}_i^\top, y_i)\}_{i=1}^n$，其中协变量 $\boldsymbol{x}_i$ 根据预先指定的抽样方案选取，响应变量 y_i 由目标模型（1）生成。在本小节中，我们介绍所提出的少样本个性化框架的第 2 步。

如前文所述，预训练模型 $f^{(\mathrm{ptr})}$ 可能不规则，或与目标回归函数的匹配程度较低。为确保后续偏差校正的稳定性，我们首先通过局部平滑操作对 $f^{\mathrm{ptr})}$ 进行正则化。

对于任意 $\boldsymbol{x}^*$， $\boldsymbol{x}' \in \mathcal{X}$以及给定函数 $g(\cdot)$，定义

$$\omega_{\boldsymbol{\theta},\boldsymbol{x}^*} \circ g(\boldsymbol{x}) = g(\boldsymbol{x}^*) + \min(|g(\boldsymbol{x}) - g(\boldsymbol{x}^*)|, \theta_1 \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^{\theta_2}) \mathrm{sgn}(g(\boldsymbol{x}) - g(\boldsymbol{x}^*)), \tag{3}$$

其中，$\mathrm{sgn}(\cdot)$ 表示符号函数。我们将这一操作称为**“$\boldsymbol{\theta}$-局部平滑”**。粗略地说，变换 $\omega_{\boldsymbol{\theta},\boldsymbol{x}^*}(\cdot)$ 通过截断 $g(\cdot)$ 中过大的局部变化，强制在 $\boldsymbol{x}^*$ 附近满足局部赫尔德正则性。重要的是，该变换是确定性的，并且不依赖于观测数据。

为形式化由此产生的光滑性性质，我们引入如下定义。

DEFINITION 2（局部光滑性）。对于定义在 $\mathcal{X}$上的任意 $f(\cdot)$ 以及给定的 $\boldsymbol{x} \in \mathcal{X}$，如果 $f \in L_{\boldsymbol{x}}(\boldsymbol{\theta})$ 满足

$$\sup_{\boldsymbol{x}' \in \mathcal{X}} \frac{|f(\boldsymbol{x}') - f(\boldsymbol{x})|}{\|\boldsymbol{x}' - \boldsymbol{x}\|_2^{\theta_2}} \le \theta_1$$

则称其对某个有限的 $\theta_1 > 0$ 和 $\theta_2 \ge 0$成立。

定义 2 中的局部光滑性概念弱于全局赫尔德光滑性。特别地，如果 $f \in H(\boldsymbol{\theta})$ 满足 $0 < \theta_2 \le 1$，则对于所有 $\boldsymbol{x} \in \mathcal{X}$，都有 $f \in L_{\boldsymbol{x}}(\boldsymbol{\theta})$。因此，如果 $g(\cdot)$ 已经满足参数为 $\boldsymbol{\theta}$ 的赫尔德条件，那么 $\omega_{\boldsymbol{\theta},\boldsymbol{x}^*} \circ g = g$。反之，局部光滑性并不对参考点 $\boldsymbol{x}^*$之外的区域施加任何全局正则性。根据构造，对于定义在 $\mathcal{X}$上的任意给定函数 $g(\cdot)$，变换后的函数 $\omega_{\boldsymbol{\theta},\boldsymbol{x}} \circ g$属于 $L_{\boldsymbol{x}}(\boldsymbol{\theta})$，并且对于任意给定的 $\boldsymbol{x}^*$，满足 $\omega_{\boldsymbol{\theta},\boldsymbol{x}^*} \circ g(\boldsymbol{x}^*) = g(\boldsymbol{x}^*)$。

图1 给出了一个示例，其中有 $\boldsymbol{\theta} = (1, 0.5)^\top$。我们看到，函数 $f_1($ x $=$ $0.7|x|^{1/4} \notin L_0$ $\boldsymbol{\theta}$ $x \in [-0.5, 0.5]$ $\tilde{f}_1$ $\boldsymbol{x}$ $= \omega_{\boldsymbol{\theta}\,0} \circ f_1$ $\boldsymbol{x}$ $)\,()$ for 被转换为 $()$，$()$，其在零附近更加平坦。函数 $f_2(x) = 0.7|x|^{1/2} \in H(\boldsymbol{\theta})$ and 不受 (3) 的影响。我们将对 $f^{(\mathrm{ptr})}(\cdot)$ 应用这一变换，并说明该变换能够使我们的最终估计量达到所需的理论性能。

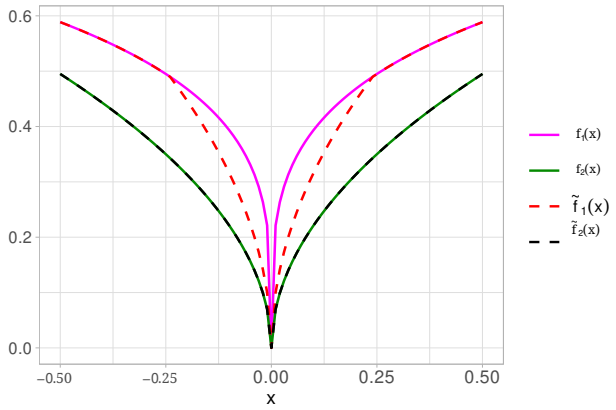


图 1. $\boldsymbol{\theta}$ 局部平滑示意图。对于 $x \in [-0.5, 0.5]$，我们设置 $f_1(x) = 0.7|x|^{1/4}$ 和 $f_2(x) = 0.7|x|^{1/2}$。虚线对应于通过 (3) 实现的 $\tilde{f}_1(\boldsymbol{x}) = \omega_{\boldsymbol{\theta},\,0} \circ f_1(x)$ 和 $\tilde{f}_2(\boldsymbol{x}) = \omega_{\boldsymbol{\theta},\,0} \circ f_2(x)$，其中 $\boldsymbol{\theta} = (1, 0.5)^\top$。

接下来，我们介绍偏差校正步骤。给定 n 检索样本，我们使用其中较大比例的样本来估计模型，并使用较小比例的样本作为验证样本，以选择调优参数。样本划分步骤在第 2.4 节中进行了描述。令 $\mathcal{N}^{(tr)} \subseteq [n]$表示训练样本的索引集，并暂时假设其已给定。对于任意给定的 $\boldsymbol{x} \in \mathcal{X}$，令

$$\hat{\delta}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\sum_{i \in \mathcal{N}^{(tr)}} (y_i - \omega_{\boldsymbol{\theta},\boldsymbol{x}} \circ f^{(\mathrm{ptr})}(\boldsymbol{x}_i)) \mathbb{I}(\|\boldsymbol{x}_i - \boldsymbol{x}\|_\infty \le h)}{1 \vee \sum_{i \in \mathcal{N}^{(tr)}} \mathbb{I}(\|\boldsymbol{x}_i - \boldsymbol{x}\|_\infty \le h)}, \tag{4}$$

$\hat{}$估计量 $\delta_{\boldsymbol{\theta}}(\boldsymbol{x})$ 对应于偏差函数 $\delta_{\boldsymbol{\theta}}$ 的核估计量，$_{\boldsymbol{x}}(\boldsymbol{x}) = f^*(\boldsymbol{x}) - \omega_{\boldsymbol{\theta}}$，$_{\boldsymbol{x}} \circ f(\mathrm{ptr})(\boldsymbol{x})$ 利用了 $\delta_{\boldsymbol{\theta}}$ 在 $\boldsymbol{x}$ 附近的局部光滑性。由于方程（3）表明 $\omega_{\theta\boldsymbol{x}} \circ f(\mathrm{ptr})(\boldsymbol{x}) = f^{(\mathrm{ptr})}(\boldsymbol{x})$，估计量 $\hat{}\,\delta_{\boldsymbol{\theta}}(\boldsymbol{x})$ 也是 $f^*(\boldsymbol{x}) - f^{(\mathrm{ptr})}(\boldsymbol{x})$ 的适当估计。因此，我们将校准估计量定义为

$$\hat{f}_{\boldsymbol{\theta}}^{(\mathrm{fsp})}(\boldsymbol{x}) = f^{(\mathrm{ptr})}(\boldsymbol{x}) + \hat{\delta}_{\boldsymbol{\theta}}(\boldsymbol{x}),\ \boldsymbol{x} \in \mathcal{X}. \tag{5}$$

鉴于（4）和（5），对于任意给定的 $\boldsymbol{\theta}$ 和 $\boldsymbol{x} \in \mathcal{X}$，计算 $\hat{}\,f_{\boldsymbol{\theta}}^{(\mathrm{fsp})}(\boldsymbol{x})$ 只需查询在检索到的 $\boldsymbol{x}_i, i \in \mathcal{N}^{(tr)}$ 处的 $f^{(\mathrm{ptr})}$ 以及每个测试点，因此具有较高的计算效率。

注意，当我们对任意 $\boldsymbol{x}' \in \mathcal{X}$ 应用 $\boldsymbol{\theta}$-局部平滑时，使用 $\theta_1 = 0, \omega_{\boldsymbol{\theta},\boldsymbol{x}} \circ f_{(\mathrm{ptr})}(\boldsymbol{x}') = f^{(\mathrm{ptr})}(\boldsymbol{x})$，因此 $\hat{}\,f_{\boldsymbol{\theta}}^{(\mathrm{fsp})}(\cdot)$ 退化为仅基于目标样本的单任务核估计。在这种情况下（$\theta_1 = 0$），预训练模型不会被用于个性化估计。对于 $\theta_1 > 0$，通过局部平滑以受控方式引入预训练模型中的信息。因此，调优参数 $\boldsymbol{\theta}$ 控制预训练模型对最终估计量的影响程度，即使 $f_{(\mathrm{ptr})}$ 与 f^* 的匹配度较低，也能确保稳健性。

接下来，我们将利用验证样本选择合适的 $\boldsymbol{\theta}$，具体内容将在下一小节中详述。

2.3. 适应。我们利用交叉验证技术选择调优参数 $\boldsymbol{\theta}$。令 $\Theta = \{0, c_1/\text{对数 } n,\ 2c_1/\text{对数 } n, \ldots\ldots, c_1\} \times \{0, 1/\text{对数 } n,\ 2/\text{对数 } n, \ldots\ldots, 1\}$ 为用于搜索最优 $\boldsymbol{\theta}$ 的网格，其中 c_1 是预先确定的常数。我们首先基于训练样本 $\mathcal{N}^{(tr)}$ 拟合 $\{\hat{}\,f_{\boldsymbol{\theta}}^{(\mathrm{fsp})}\}_{\boldsymbol{\theta} \in \Theta}$。令 $\mathcal{N}^{(va)} \subseteq [n] \setminus \mathcal{N}^{(tr)}$ 表示验证样本集合。令

暂且假设 $\mathcal{N}^{(tr)}$ 和 $\mathcal{N}^{(va)}$ 已知，我们将在下一小节讨论如何将 n 个检索样本划分为训练样本和验证样本。具体而言，令

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{i\in\mathcal{N}^{(va)}} (y_i - \hat{f}_{\boldsymbol{\theta}}^{(\mathrm{fsp})}(\boldsymbol{x}_i))^2. \tag{6}$$

也就是说，我们选择 $\boldsymbol{\theta}$，使其成为验证样本中平均预测误差最小的最佳调优参数。在网格 Θ 中纳入取值为 $\theta_1 = 0$ 的候选项，可确保模型能够根据预训练模型所提供信息的丰富程度进行自适应。特别是，当预训练模型无法为目标任务提供有用信号或仅提供少量有用信号时，该过程可以选择一个有效忽略 $f^{(\mathrm{ptr})}$ 的调优参数，从而退化为仅目标估计量。

2.4. 样本检索方案。本节介绍所提出的检索方案，并详细说明样本划分。为说明其设计原理，我们首先分析在 $\mathcal{X}$ 上采用一般预先确定的抽样方案 $p_X(\boldsymbol{x})$ 时，固定调优参数 $\boldsymbol{\theta}$ 下估计量 $\hat{f}_{\boldsymbol{\theta}}^{(\mathrm{fsp})}(\boldsymbol{x})$ 的理论性质。该分析揭示了抽样分布如何影响均方积分误差（MISE），并为构建高效的检索策略提供指导。

我们首先给出非参数回归的标准正则性条件。

条件1（全局光滑性）。假设 $f^*(\boldsymbol{x}) \in H(\boldsymbol{\theta}^*)$ 在 $\mathcal{X}$ 上满足某个有限的$\theta_1^* \geq 0$ 和 $0 < \theta_2^* \leq 1$。

条件2（次高斯噪声）。假设 ϵ_i 相互独立，服从均值为零、方差为 $\sigma^2(\boldsymbol{x}_i)$ 的次高斯分布，其中 $i = 1, \ldots, n$。存在某个正数常量 $\sigma_{\max}$，使得 上确界$_{\boldsymbol{x}\in\mathcal{X}}\, \sigma^2(\boldsymbol{x}) \leq \sigma_{\max}^2$。函数 $\sigma(\boldsymbol{x}) \in H(\boldsymbol{\theta}^{(\sigma)})$ 在 $\mathcal{X}$ 上满足某个有限的$\theta_1^{(\sigma)} \geq 0$ 和 $\theta_2^{(\sigma)} > 0$。

条件1 规定了 $f^*(\cdot)$ 的赫尔德光滑性。正如前文所述，这是迁移学习和经典非参数文献中的标准假设。条件 2允许存在异方差次高斯噪声，并对方差函数施加了温和的光滑性约束，从而有助于方差估计。

在下述引理中，我们首先展示抽样分布 p_X（$\boldsymbol{x}$） 对 MISE 的影响，以此说明所提出检索方案的动机。由于预训练模型$f^{(\mathrm{ptr})}$（·） 是从独立于检索样本的外部数据源中学习得到的，在不失一般性的情况下，我们在分析中将其视为确定性的。

引理 1 （一般检索方案下的平均积分平方误差）。假设条件 *1* 和 *2* 成立。设 $\boldsymbol{x}_i$ 根据某个预先确定的 p_X（$\boldsymbol{x}$）生成，相应的 y_i 根据（*1*）生成。则对于任意有限的 $\theta_1 \geq 0$ 以及任意 $0 < \theta_2 \leq 1$，都有

$$\mathrm{MISE}(\hat{f}_{\boldsymbol{\theta}}^{(\mathrm{fsp})}) \lesssim \gamma_1^2(\boldsymbol{\theta})(\sqrt{d}h)^{2\gamma_2(\boldsymbol{\theta})} + \int_{\mathcal{X}} \mathbb{E}[\frac{\sigma^2(\boldsymbol{x}) + \sigma_{\max}\theta_1^{(\sigma)} h^{\theta_2^{(\sigma)}}}{n_h(\boldsymbol{x}) \vee 1}] d\boldsymbol{x},$$

其中 n_h（$\boldsymbol{x}$）$= \sum_{i=1}^n$ 𝕀（$\|\boldsymbol{x}_i - \boldsymbol{x}\|_\infty \leq h$） 和 $\boldsymbol{\gamma}$（$\boldsymbol{\theta}$） 是光滑度参数，使得对于所有 $\boldsymbol{x} \in \mathcal{X}$，$\delta_{\boldsymbol{\theta},\,\boldsymbol{x}} \in L_{\boldsymbol{x}}$（$\boldsymbol{\gamma}$（$\boldsymbol{\theta}$）

引理1 上界中的第一项对应于估计量的偏差平方，取决于局部平滑后偏差函数的局部光滑性 $\delta_{\boldsymbol{\theta}}(\boldsymbol{x})$。下一节将讨论 $\gamma(\boldsymbol{\theta})$ 的大小，本节其余部分则聚焦于样本检索带来的方差

降低效应。上界中的第二项反映方差贡献，与局部样本数量 $n_h(\boldsymbol{x})$成反比。在带宽和检索分布满足温和条件时，项 $\sigma_{\max}\theta_1^{(\sigma)}h^{\theta_2^{(\sigma)}}$可以忽略，且 $n_h(\boldsymbol{x}) \propto nh^d p_X(\boldsymbol{x})$。因此，$p_X(\boldsymbol{x})$的理想选择应使 $\int_{\boldsymbol{x}\in\mathcal{X}} \sigma^2(\boldsymbol{x})/p_X(\boldsymbol{x})d\boldsymbol{x}$ 最小，这表明最优采样密度应在噪声水平较高的区域分配更多样本。由此得到以下最优抽样分布：

$$p_X^*(\boldsymbol{x}) = \frac{\sigma(\boldsymbol{x})}{\int_{\mathcal{X}} \sigma(\boldsymbol{x})d\boldsymbol{x}},\ \boldsymbol{x}\in\mathcal{X}, \tag{7}$$

这种形式符合直觉：噪声方差较大的区域需要更密集的采样，才能实现准确估计。

接下来，我们推导最优检索方案下相应的均方积分误差界。设

$$\bar{\sigma} = \frac{\int_{\mathcal{X}} \sigma(\boldsymbol{x})d\boldsymbol{x}}{|\mathcal{X}|}$$

表示平均噪声水平。对于给定的带宽 h，定义 $r_n = $最大值$\{\sqrt{\sigma_{\max}\theta_1^{(\sigma)}h^{\theta_2^{(\sigma)}}}, \sigma_{\max}$对数$nh^d{}^{n}\}$以及 $\mathcal{X}_0 = \{\boldsymbol{x}\in\mathcal{X}:\ \sigma(\boldsymbol{x}) \le c_r r_n\}$，其中常数 c_r 足够大。

推论 1（最优检索方案下的均方积分误差）。假设条件 *1* 和*2*成立。设 $\boldsymbol{x}_i, i=1,\dots,n$ 按照 $p_X^*(\boldsymbol{x})$ 生成，相应的 y_i 则按照（*1*）生成。若对于某个常数 $c_0\ge 0$，$h = n^{-c_0}$，且对于某个常数 $c_1<1$，$|\mathcal{X}_0| \le c_1$最小值$\{\bar{\sigma}^2/(r_n^2 nh^d)\ 1\}$，那么对于任意有限的 $\theta_1\ge 0$ 和 $\theta_2\ge 0$，我们有

$$\text{MISE}(\hat{f}_{\boldsymbol{\theta}}^{(\text{fsp})}) \lesssim \gamma_1^2(\boldsymbol{\theta})(\sqrt{d}h)^{2\gamma_2(\boldsymbol{\theta})} + \frac{\bar{\sigma}^2}{nh^d}, \tag{8}$$

其中 $\boldsymbol{\gamma}(\boldsymbol{\theta})$ 为光滑度参数，满足对于所有 $\boldsymbol{x}\in\mathcal{X}$，$\delta_{\boldsymbol{\theta},\boldsymbol{x}}(\cdot) \in L_{\boldsymbol{x}}(\boldsymbol{\gamma}(\boldsymbol{\theta}))$。

推论 1 确立了在最优检索方案 $p_X^*(\cdot)$下个性化估计的收敛速率。式（8）右侧第一项对应偏差平方，第二项则表示方差贡献。推论 1 中的条件要求 $|\mathcal{X}_0|$ 足够小，以确保每个局部核邻域都包含足够多的样本。在最优抽样方案下，估计量的有效噪声水平由平均方差 $\bar{\sigma}^2$刻画。方差降低效应将在注 1中进一步说明。

注释 1（方差比较）。考虑均匀采样方案，即从 $\mathcal{X}=[0,1]$ 中随机抽取 n 个点。假设对于 $x\in[0,b_n-a_n]$，$\sigma(\boldsymbol{x}) = b_n - x$，并且对于 $\boldsymbol{x}\in[b_n-a_n,1]$，$\sigma(\boldsymbol{x}) = a_n$，其中 $b_n > a_n$。在这种均匀随机抽样方案下，方差项为

$$\frac{\int_{\mathcal{X}} \sigma^2(\boldsymbol{x})d\boldsymbol{x}}{nh^d} \ge \frac{\bar{\sigma}^2}{nh^d} + \frac{(b_n-a_n)^3}{12}.$$

最后一项表示当 $a_n \ne b_n$ 时，随机抽样相对于最优采样方案所造成的效率损失。

如果 $b_n - a_n = o$（1）且 $a_n = o$（$b_n - a_n$），那么

$$\frac{\int_{\boldsymbol{x}\in\mathcal{X}} \sigma^2(\boldsymbol{x})d\boldsymbol{x}}{nh^d} \gg \frac{\bar{\sigma}^2}{nh^d}, \tag{9}$$

这意味着，最优采样方案的方差阶严格小于均匀采样。

备注 1 的证明留待补充材料。备注 1 表明，当噪声具有异方差性时，与均匀采样相比，最优采样方案能够显著降低估计量的方差。特别是，公式（9）表明，当噪声方差在足够大的区域内较小时，方差阶本身也能得到改善。分类任务就是噪声具有异方差性的典型例子。具体而言，如果 $\mathbb{P}(y_i=1|\boldsymbol{x}_i)=f^*(\boldsymbol{x}_i)$，那么 $\sigma^2(\boldsymbol{x})=f^*(\boldsymbol{x}($ $1-f^*(\boldsymbol{x}))$。因此，在 $f^*(\boldsymbol{x})$ 接近0或1的区域，噪声方差较小；而当 $f^*(\boldsymbol{x})$ 接近1/2时，噪声方差最大，这使得自适应采样尤其有利。

虽然最优密度 $p_X^*(\boldsymbol{x})$ 取决于未知的方差函数 $\sigma^2(\boldsymbol{x})$，但可以根据数据对其进行近似估计。为此，我们将一小部分采样预算用于估计 $\sigma^2(\boldsymbol{x})$，并利用所得估计来指导样本检索。该过程总结于算法 1。

算法 1 样本检索方案。

输入：采样预算 n 和目标区域 $\mathcal{X}$。输出：检索到的样本 $(\boldsymbol{x}_i^\top, y_i)$、$i=1$、……、$n, \mathcal{N}(tr)$，以及 $\mathcal{N}(va)$。步骤 1. 对于某个较小的常数 c，从 $\mathcal{X}$ 中随机检索 $n_0=cn$ 个带标签样本 $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n_0}$。对于 $h_\sigma=n^{-1/(d+2)}$ 和 $\boldsymbol{x}\in\mathcal{X}$，计算

$$\hat{\sigma}^2(\boldsymbol{x})=\frac{\sum_{i=1}^{n_0/2} y_i^2 K_{h_\sigma}(\boldsymbol{x}_i,\boldsymbol{x})}{1\vee\sum_{i=1}^{n_0/2} K_{h_\sigma}(\boldsymbol{x}_i,\boldsymbol{x})}-\frac{\{\sum_{i=1}^{n_0/2} y_i K_{h_\sigma}(\boldsymbol{x}_i,\boldsymbol{x})\}^2}{1\vee\{\sum_{i=1}^{n_0/2} K_{h_\sigma}(\boldsymbol{x}_i,\boldsymbol{x})\}^2}, \tag{10}$$

其中 $K_h(\boldsymbol{x},\boldsymbol{x}')$ =取最大值$\{0, h-\|\boldsymbol{x}-\boldsymbol{x}'\|_\infty\}$。
步骤 2. 从以下分布中抽取 $\boldsymbol{x}_i, i=n_0+1, \ldots\ldots, n$：

$$\hat{p}_X(\boldsymbol{x})=\frac{\hat{\sigma}(\boldsymbol{x})}{\int_{\mathcal{X}}\hat{\sigma}(\boldsymbol{x})d\boldsymbol{x}}$$

设定 $\mathcal{N}\ (tr)=\{n_0+1,\ldots,n\}$ 以及 $\mathcal{N}\ (va)=\{n_0/2+1,\ldots,n_0\}$。输出 $\boldsymbol{x}_i$ 及其对应的响应变量 y_i、$i=1$、……、n。

现在我们进一步展示算法 1 的示例。在步骤 1 中，我们首先从 $\mathcal{X}$ 中均匀随机地检索 n_0 个样本。其中一半样本用于计算 $\hat{}\sigma^2$（$\boldsymbol{x}$），另一半将用于自适应步骤中的验证。估计出方差函数后，即可估计所提出的检索样本 $\hat{}p_X$（·）。我们基于核估计来估计 σ^2（$\boldsymbol{x}$）。核估计量（10）计算简单，不需要预先了解光滑度参数，并且在温和条件下是一致的。在步骤 2 中，可以使用标准的拒绝采样方法 [7, 11] 从 $\hat{}p_X$（$\boldsymbol{x}$）中进行采样。为完整起见，补充材料中的算法A.1给出了拒绝采样的详细算法。

请注意，我们对带宽 h_σ 的选择是为了简便。实际上，只要 h_σ 的宽度为 $o(1)$，且每个局部估计所使用的样本数趋于无穷，估计量$\hat{}\sigma^2(\boldsymbol{x})$ 就是一致的，下一小节中的理论保证仍然成立。

在下一个引理中，我们论证所提出检索方案的性能。令 $\mathcal{X}_1=\{x\in\mathcal{X}:\ \sigma(\boldsymbol{x})\le c_r\tilde{r}_n\}$，其中 $\tilde{r}_n=r_n+\log n\cdot n^{-\min(2\theta_2^*,\theta(\sigma))/(2,1/2)^{4+2d}}$，这里的常数 c_r 足够大。

引理 2（所提出检索方案的速率最优性）。假设条件 1、2。设 $h=n^{-c_0}$ 对某个常数 $c_0\ge 0$ 和 $|\mathcal{X}_1|\le c_1$ 最小值 $\{\bar{\sigma}^2/(\tilde{r}_n^2 nh^d)\ 1\}$ 成立，其中某个

正数常量 $c_1 < 1$。则存在某个正数常量 C，使得

$$\int_{\mathcal{X}} \mathbb{E}[\frac{\sigma^2(\boldsymbol{x}) + \sigma_{\max}\theta_1^{(\sigma)} h^{\theta_2^{(\sigma)}}}{\max\{1, n_h(\boldsymbol{x})\}}] d\boldsymbol{x} \leq C\frac{\bar{\sigma}^2}{nh^d}.$$

引理 2 表明，在所提出的检索方案下，方差项 $\hat{}p_X(\cdot)$ 达到与最优检索方案 $p_X^*(\cdot)$ 相同的速率。该结果证明了所提出检索方案的有效性。对 $|\mathcal{X}_1|$ 的额外限制用于控制估计 $\sigma(\boldsymbol{x})$时产生的误差。最后，我们在算法 2中总结了结合采样阶段和学习阶段的完整个性化算法。

算法2 $f^*(\boldsymbol{x})$ 的个性化非参数估计，$\boldsymbol{x} \in \mathcal{X}$。

输入：预训练模型 f(ptr)、采样预算 n 和目标区域 $\mathcal{X}$。输出：$\hat{}f$(fsp) $\hat{}_{\boldsymbol{\theta}}(\boldsymbol{x})$，$\boldsymbol{x} \in \mathcal{X}$。步骤 1。通过算法 1 获得 $\hat{}p_X(\boldsymbol{x})$ 并生成样本 $(\boldsymbol{x}_i^\top, y_i)$、$i = 1$、…、$n$。步骤 2。对于每个 $\boldsymbol{\theta} \in \Theta$，通过公式 (5) 使用 $\mathcal{N}(tr) = \{n_0 + 1, \ldots, n\}$ 计算 $\hat{}f$(fsp) $_{\boldsymbol{\theta}}(\boldsymbol{x})$。步骤 3。通过公式 (6) 使用 $\mathcal{N}(va) = \{n_0/2 + 1, \ldots, n_0\}$ 估计模型权重 $\hat{}\boldsymbol{\theta}$，并输出 $\hat{}f$(fsp) $\hat{}_{\boldsymbol{\theta}}(\boldsymbol{x})$。

带宽 h 也可以通过交叉验证来选择。例如，可以联合搜索$(\boldsymbol{\theta},\ h) \in \Theta \times \mathcal{H}$，其中 $\mathcal{H} = \{1, 1/2, \ldots, 1/n\}$。相应的理论分析与（6）的论证类似，为简洁起见，此处省略。

3. 算法2的理论性质。 本节将为算法2中提出的少样本个性化过程建立理论保证。

首先，我们在所提出的检索方案下对个性化估计量的风险给出上界。

T定理 1（所提检索方案下的均方积分误差）。假设条件*1* 和*2* 成立。设 $h = n^{-c_0}$ 对某个常量 $c_0 \geq 0$ 成立，且 $|\mathcal{X}_1| \leq c_1 \min\{\bar{\sigma}^2/(\tilde{r}_n^2 nh^d)\ 1\}$ 对某个正数常量 $c_1 < 1$成立。则

$$\text{MISE}(\hat{f}_{\boldsymbol{\theta}}^{(\text{fsp})}) \lesssim \gamma_1^2(\boldsymbol{\theta}) h^{2\gamma_2(\boldsymbol{\theta})} + \frac{\bar{\sigma}^2}{nh^d},$$

其中 $\boldsymbol{\gamma}(\boldsymbol{\theta})$ 是光滑度参数，满足 $\delta_{\boldsymbol{\theta},\ \boldsymbol{x}}(\cdot) \in L_{\boldsymbol{x}}(\boldsymbol{\gamma}(\boldsymbol{\theta}))$ 对所有 $\boldsymbol{x} \in \mathcal{X}$均成立。

If we take $h \asymp \min\{(\bar{\sigma}/\gamma_1(\boldsymbol{\theta}))^{\frac{2}{2\gamma_2(\boldsymbol{\theta})+d}} n^{-\frac{1}{2\gamma_2(\boldsymbol{\theta})+d}}, 1\}$, *then*

$$\text{MISE}(\hat{f}_{\boldsymbol{\theta}}^{(\text{fsp})}) \lesssim (\gamma_1(\boldsymbol{\theta}))^{\frac{2d}{2\gamma_2(\boldsymbol{\theta})+d}} \bar{\sigma}^{\frac{4\gamma_2(\boldsymbol{\theta})}{2\gamma_2(\boldsymbol{\theta})+d}} n^{-\frac{2\gamma_2(\boldsymbol{\theta})}{2\gamma_2(\boldsymbol{\theta})+d}} + \frac{\bar{\sigma}^2}{n}. \tag{11}$$

定理 1 给出了在所提出的数据驱动检索方案下，固定 $\boldsymbol{\theta}$ 时 $\hat{}f_{\boldsymbol{\theta}}^{(\text{fsp})}(\boldsymbol{x})$ 的均方积分误差上界。该上界与推论 1 相当，表明该检索过程达到了与预言机采样密度相同的方差阶。取带宽 h 以平衡偏差项和方差项，可得上界（11），我们将在定理 3 中证明该上界是最优的。

界（11）包含一个非参数项和一个参数项。非参数项源于通过（4）在估计 $\delta_{\boldsymbol{\theta}}(\boldsymbol{x})$ 时对偏差和方差进行权衡。与 $n, n^{-\frac{2\gamma_2}{2\gamma_2(\boldsymbol{\theta})(\boldsymbol{\theta})+d}}$ 相关的量，是估计非参数函数时通常的极小极大最优速率

具有赫尔德光滑性 $\gamma_2(\boldsymbol{\theta})$。与将赫尔德常数视为固定值的经典分析不同，我们的界刻画了对 $\gamma_1(\boldsymbol{\theta})$ 的依赖性；在个性化设定中，该量可能很小。（11）的第二项是参数速率。当 $\gamma_1(\boldsymbol{\theta})$ 接近于零时，该项将占主导地位，此时 $\delta_{\boldsymbol{\theta}}(\boldsymbol{x})$ 接近于常数。相应地，基于 n 个噪声水平为 $\sigma(\boldsymbol{x}_i)$ 的样本来估计未知常数时，最优速率为 $\bar{\sigma}^2/n$，这对应于第二项。由于 $\gamma_1(\boldsymbol{\theta})$ 可能非常小，且 $\mathcal{X}$ 的宽度为常数，因此我们令 h 不大于某个常数，从而得到上界中的参数项。

接下来，我们讨论步骤 (3) 中的局部平滑调优参数 $\boldsymbol{\theta}$ 如何影响局部光滑度参数 $\boldsymbol{\gamma}(\boldsymbol{\theta})$。由于 $f^*(\cdot) \in H(\boldsymbol{\theta})$，因此 $\gamma_2(\boldsymbol{\theta}) \geq \min\{\theta_2^*, \theta_2\}$，且 $\gamma_1(\boldsymbol{\theta}) \leq \theta_1 + \theta_1^*$。特别地，选择 $\theta_2 = \theta_2^*$ 可确保 $\gamma_2(\boldsymbol{\theta}) \geq \theta_2^*$。在许多情形下，$\gamma_1(\boldsymbol{\theta})$ 可能显著小于 θ_1^*，而 $\gamma_2(\boldsymbol{\theta})$ 可能大于 θ_2，这反映了纳入预训练模型后局部赫尔德复杂度的降低。以下示例将说明这些影响。

示例 1。考虑 $f^*(\boldsymbol{x}) = f_1^*(\boldsymbol{x}) + f_2^*(\boldsymbol{x})$，其中 $f_1^* \in H(\boldsymbol{\theta})$ 且 $f_2^* \in H(\boldsymbol{\theta}')$，对 $\theta_2 < \theta_2'$ 成立。假设 $f(ptr) = f_1^*$。那么，对于 $\boldsymbol{\theta}^* = (\theta_1 + \theta_1', \theta_2)$，有 $f^*(\boldsymbol{x}) \in H(\boldsymbol{\theta}^*)$。此外，通过设置 $\boldsymbol{\theta} = \boldsymbol{\theta}^*$，对于所有 $\boldsymbol{x}$ 和 $\delta_{\boldsymbol{\theta}^*}$，$\omega_{\boldsymbol{\theta},\boldsymbol{x}} \circ f(\text{ptr})(\cdot) = f(\text{ptr})(\cdot)$。由于 $\theta_2 < \theta_2'$，这表明 $\delta_{\boldsymbol{\theta}^*,\boldsymbol{x}}(\cdot)$ 在局部光滑性刻画意义下比 $f^*(\cdot)$ 更光滑。

示例 2。$f^* \in H\ \boldsymbol{\theta}^*\quad f\ ^{ptr}\ \boldsymbol{x} = \rho f^*\ \boldsymbol{x}\quad 0 < \rho <$。设$()$和$_{()}()\ ()$对 $1\quad \boldsymbol{\theta} = \ \rho\theta_1^*\ \theta_2^*\quad \omega_{\boldsymbol{\theta}\ \boldsymbol{x}} \circ f\quad \cdot = f\quad \cdot\quad \boldsymbol{x}\quad \delta_{\boldsymbol{\theta}^*\ \boldsymbol{x}}\ \cdot\ =\ 1 -$ 某个常数成立。则对于$(,)$，有$_{,(\text{ptr})}()\ _{(\text{ptr})}()$对所有$,()\ (\rho) f^*(\cdot) \in H(\boldsymbol{\theta}')$以及 $\boldsymbol{\theta}' = ((1-\rho)\theta_1^*, \theta_2^*)$成立。这是一个 $\gamma_1(\boldsymbol{\theta}) < \theta_1^*$ 的示例。

鉴于上述两个示例，在（3）中适当选择 $\boldsymbol{\theta}$，可以提高 $\delta_{\boldsymbol{\theta}\boldsymbol{x}}(\cdot)$ 的局部光滑性，从而改善（11）中的速率。现在，我们将（11）中的界与传统单任务非参数估计的均方积分误差进行比较。如果 $\gamma_1(\boldsymbol{\theta})$ 和 $\bar{\sigma}$ 都是有界且远离 0 的常数，那么 MISE($\hat{}$ $f_{\boldsymbol{\theta}}^{(\text{fsp})}$) 的阶为 $n^{-\frac{2\gamma_2}{2\gamma_2}}(_{\theta})\ (_{\theta})^{+d}$。相比之下，如果$_{-\ 2\theta_2^*}$我们仅基于 n 个目标样本来估计 $f^*(\cdot)$，那么 MISE 的阶为 $n^{\ \frac{}{2\theta_2^*+d}}$。为保证 $\hat{}f_{\boldsymbol{\theta}}^{(\text{fsp})}$ 与传统单任务非参数估计量相比不逊色，需要在（3）中适当选择调优参数 $\boldsymbol{\theta}$。具体而言，令 $\theta_2 = \theta_2^*$ 即可保证 $\gamma_2(\boldsymbol{\theta}) \geq \theta_2$。如果 $\gamma_2(\boldsymbol{\theta})$ 严格大于 θ_2（如示例 1 所示），那么 $\hat{}f_{\boldsymbol{\theta}}^{(\text{fsp})}$ 的收敛速率快于单任务估计量，这体现了利用预训练模型的优势。$f_{\boldsymbol{\theta}}^{(\text{fsp})}$ 的另一个潜在收益是，$\gamma_1(\boldsymbol{\theta})$ 可能远小于 θ_1，如示例 2 所示。在这种情况下，$\hat{}f_{\boldsymbol{\theta}}^{(\text{fsp})}$ 的 MISE 的阶也可能小于传统单任务非参数估计量。总而言之，通过适当选择 $\boldsymbol{\theta}$，我们的方法可以通过提高赫尔德光滑性来获得更快的收敛速率。

现在我们论证自适应步骤的有效性。

$\hat{}$T定理 2（自适应后的均方积分误差）。假设定理*1*的条件成立。对于由（ $\boldsymbol{\theta}$ （*6*）得到的，有

$$\text{MISE}(\hat{f}_{\hat{\boldsymbol{\theta}}}^{(\text{fsp})}) \leq C_1 \min_{\boldsymbol{\theta} \in \Theta} \text{MISE}(\hat{f}_{\boldsymbol{\theta}}^{(\text{fsp})}) + \frac{C_2 \sigma_{\max}^2 (\log n)^2}{n}$$

对某些正数常量 C_1 和 C_2。

定理 2 表明，验证步骤会从候选集中选择一个调优参数，使风险达到最优 Θ，且仅相差一个较小的余项。该余项反映了适应成本，其阶为 $(\log n)^2/n$，其中 $(\log n)^2$ 确实是搜索空间的基数 Θ，$|\Theta|$。在主要关注的情形下，该项相对于式（11）中的非参数项可以忽略。所提出的调优参数选择步骤很简单，因为$|\Theta|$ 在当前设定下很小。如果 $|\Theta|$ 较大，则可以采用更先进的模型选择方法 [8, 17]。

接下来，我们证明所提出方法的极小极大最优性。我们定义以下函数空间

$$\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}, \sigma_0) = \Bigg\{ P_{y|\boldsymbol{x}} : f^*(\boldsymbol{x}) \in H(\boldsymbol{\theta}^*) \; \forall \, \boldsymbol{x} \in \mathcal{X}, \; \left(\int_{\mathcal{X}} \sigma(\boldsymbol{x}) d\boldsymbol{x} \right) / |\mathcal{X}| \leq \sigma_0, \tag{12}$$
$$f^*(\cdot) - \omega_{\boldsymbol{\theta},\boldsymbol{x}} \circ f^{(\mathrm{ptr})}(\cdot) \in L_{\boldsymbol{x}}(\boldsymbol{\gamma}) \; \forall \boldsymbol{x} \in \mathcal{X} \Bigg\},$$

其中 $\boldsymbol{\theta}$ 表示局部平滑步骤中的预先设定调优参数，而 $P_{y|\boldsymbol{x}}$表示给定 y 时 $\boldsymbol{x}$的条件分布。该函数空间考虑光滑度参数为 $\boldsymbol{\theta}^*$的赫尔德光滑目标函数 $f^*(\cdot)$。此外，参数 $\boldsymbol{\gamma}$ 表示 $\delta_{\boldsymbol{\theta}}$ 经 $\boldsymbol{\theta}$-局部平滑后偏差函数 $_{\boldsymbol{x}}(\cdot)$的局部光滑度参数。在下一个定理中，我们证明该函数空间（12）中均方积分误差的极小极大速率。

T定理 3（极小极大下界）。令 $p_X(\boldsymbol{x})$ 表示用于生成检索协变量 $\boldsymbol{x}_i$、$i = 1$、.....、n 的抽样分布，并令 $\mathcal{D}_n = \{\boldsymbol{x}_i^\top, y_i\}_{i=1}^n$ 表示检索样本。对于 $0 \leq \theta_1 \leq C_1 \theta_1^*$ 和 $\theta_2 \geq \theta_2^*$，存在某个正数常量 C_2，使得

$$\inf_{p_X, \hat{f}(\mathcal{D}_n, f^{(\mathrm{ptr})})} \sup_{f^* \in \mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}, \sigma_0)} \mathrm{MISE}(\hat{f}) \geq C_2 \gamma_1^{\frac{2d}{2\gamma_2 + d}} \sigma_0^{\frac{4\gamma_2}{2\gamma_2 + d}} n^{-\frac{2\gamma_2}{2\gamma_2 + d}} + C_2 \frac{\sigma_0^2}{n}, \tag{13}$$

其中 $\gamma_1 \leq C_3 \theta_1^*$ 和 $\gamma_2 \geq \theta_2^*$，其中 C_3为某个正数常量。

定理 3 在 $\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}, \sigma_0)$ 上给出了极小极大下界，并对检索分布 p_X 以及基于 $(\mathcal{D}_n, f^{(\mathrm{ptr})})$的所有估计量取下确界。这种表述刻画了用户指定的检索方案对极小极大风险的影响。在关于 $\boldsymbol{\theta}$的既定条件下，局部正则性参数满足 $\gamma_1 = O(\theta_1^*)$ 和 $\gamma_2 \geq \theta_2^*$，这意味着个性化极小极大速率不劣于经典单任务极小极大速率。

定理1 和 3 共同表明，对于固定的调优参数 $\boldsymbol{\theta}$，估计量$\hat{f}_{\boldsymbol{\theta}}^{(\mathrm{fsp})}(\boldsymbol{x})$达到了极小极大最优速率。此外，根据定理 2，估计量 $\hat{f}_{(\mathrm{fsp})\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$在 Θ 上达到了预言机风险，误差至多为对数因子。如上所述，在常见情形下，对数膨胀可以忽略，因为（13）右侧的第一项是非参数速率。

4. 扩展。本节将讨论所提算法在另外两种情形下的扩展。

4.1. 向无穷小 $\mathcal{X}$ 的扩展。在某些应用中，目标协变量区域 $\mathcal{X}$ 可能很小，而非具有恒定的勒贝格测度。例如，我们可能希望将通用预测模型个性化应用于一个较小的亚群体，此时目标协变量支撑集可能满足 $|\mathcal{X}| = o(1)$。我们通过考虑 $\mathcal{X} = [0, \nu_n]^d$（其中 $\nu_n \leq 1$，且允许 ν_n 随着 n 趋于无穷而趋于零）来研究这一情形。

当 ν_n 较小时，方差优化加权抽样方案的重要性会降低，因为区域的直径限制了 $\sigma(\boldsymbol{x})$ 在 $\mathcal{X}$ 上的异质性。因此，我们考虑在 $\mathcal{X}$ 上进行均匀采样，并相应地调整算法 2 。

算法3 $f^*(\boldsymbol{x})$ 的个性化非参数估计， $\boldsymbol{x} \in \mathcal{X}$针对无穷小 $\mathcal{X}$。

输入：预训练模型 $f(\mathrm{ptr})$、采样预算 n 和目标区域 $\mathcal{X} = [0, \nu_n]^d$。输出：$\hat{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$， $\boldsymbol{x} \in \mathcal{X}$。步骤 1. 从 $\mathcal{X}$ 中均匀采样 $\boldsymbol{x}_i, i = 1, \ldots, n$，并根据 (1) 获取对应的响应变量 y_i。步骤 2. 对每个 $\boldsymbol{\theta} \in \Theta$，通过 (5) 计算 $\hat{f}(\mathrm{fsp})_{\boldsymbol{\theta}}(\boldsymbol{x})$，其中使用 $\mathcal{N}(tr) = \{n_0 + 1, \ldots, n\}$ 和带宽 $h \le \nu_n$。步骤 3. 通过 (6) 估计模型权重 $\hat{\boldsymbol{\alpha}}$，其中使用 $\mathcal{N}(va) = \{1, \ldots, n_0\}$，并输出 $\hat{f}(\mathrm{fsp})_{\boldsymbol{\theta}}(\boldsymbol{x})$。

算法3 与算法2 的主要区别在于检索步骤（在 $\mathcal{X}$ 上均匀采样）以及额外的约束条件 $h \le \nu_n$，该约束确保核估计量使用的局部邻域始终包含在目标区域内。以下推论给出了相应的风险界。

推论 2. 假设条件1、2成立。设 $\nu_n = n^{-c_0}$，其中 $c_0 \ge 0$ 为某个常数，且 $\bar{\sigma}^2 \ge \nu_n^{\theta_2^{(\sigma)}}$。若取 $h \asymp$ 最小值$\{\nu^{2\gamma_2 \overset{d}{\boldsymbol{\theta}}()} \overline{n^{\quad +d}} (\bar{\sigma}/\gamma_1(\boldsymbol{\theta}))^{2\gamma_2 \overset{2}{\boldsymbol{\theta}}()+d} n^{-2\gamma_2 \overset{1}{\boldsymbol{\theta}}()+d}$， $\nu_n\}$，则对于任意有限的 $\theta_1 \ge 0$ 和 $\theta_2 \ge 0$，均有

$$\mathrm{MISE}(\hat{f}_{\boldsymbol{\theta}}^{(\mathrm{fsp})})/\nu_n^d \lesssim (\gamma_1(\boldsymbol{\theta}))^{\frac{2d}{2\gamma_2(\boldsymbol{\theta})+d}} \nu_n^{\frac{2d\gamma_2(\boldsymbol{\theta})}{2\gamma_2(\boldsymbol{\theta})+d}} \bar{\sigma}^{\frac{4\gamma_2(\boldsymbol{\theta})}{2\gamma_2(\boldsymbol{\theta})+d}} n^{-\frac{2\gamma_2(\boldsymbol{\theta})}{2\gamma_2(\boldsymbol{\theta})+d}} + \frac{\bar{\sigma}^2}{n}. \tag{14}$$

在（14）中，我们建立了任意给定 $\boldsymbol{\theta}$时，个性化估计量$\hat{f}_{\boldsymbol{\theta}}^{(\mathrm{fsp})}(\cdot)$在收缩定义域上的均方积分误差上界。由于 $\mathcal{X}$ 的测度为 ν_n^d，因此 $\mathrm{MISE}(\hat{f}_{\boldsymbol{\theta}}^{(\mathrm{fsp})})/\nu_n^d$对应于在 $\mathcal{X}$ 上的平均积分均方误差。该界表明，在其他参数保持不变的情况下，随着 ν_n 缩小，平均误差会减小。直观而言，较小的定义域允许采用更小的有效带宽，从而降低非参数估计问题的难度。在这一情形下，均匀采样能够达到与方差加权采样相同的阶，因此显式重新加权带来的额外收益有限。

4.2. 存在来自 $\mathcal{X}$ 的未标记数据。在某些实际场景中，可能无法从目标分布中自由生成带标签样本。少样本个性化的一种常见替代设定是：可以获得来自目标区域的大量未标记协变量，但只能获取有限数量的标签。具体而言，假设我们观测到一个预训练模型 $f^{(\mathrm{ptr})}$、一个目标区域 $\mathcal{X}$，以及未标记协变量 $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_N \in \mathcal{X}$，并且只允许查询这些点中 $n \le N$ 个点的标签。我们考虑 n 可以小于且远小于 N 的情形。

在这种设定下，我们可以设计一种抽样过程，选取 n 无标签样本集中的协变量点，使其经验分布近似于目标抽样密度$p_X(\cdot)$。由此得到的检索算法见算法 4。

算法4 给定来自 $\mathcal{X}$ 的 N 无标签样本时的个性化非参数估计。

输入：抽样预算 n、目标区域 $\mathcal{X}$ 和无标签样本 $\boldsymbol{x}_{i}^{(u)}\}_{i=1}^{N} \in \mathcal{X}, i=1,\ldots,N$。输出：检索样本 $(\boldsymbol{x}_i^{\top}, y_i)$, $i=1,\ldots,n, \mathcal{N}(tr)$，以及 $\mathcal{N}(va)$。步骤 1. 从 $\{\boldsymbol{x}(i\ \}_{i=1})$ 中随机抽取 n_0 个点，并将选定索引集记为 $\mathcal{N}_0$。根据 $\{1,\ldots,n_0/2\}$，按照算法 1 的步骤 1 计算 $\hat{\sigma}^2(\boldsymbol{x})$。步骤 2. 按照算法 1 的步骤 2，从 $\mathcal{X}$ 中随机抽取 $\check{\boldsymbol{x}}_i, i=1,\ldots,N-n_0$ 个点。基于样本 $((\boldsymbol{x}(u)\, i)^{\top}, 0)\ i \in [N] \setminus \mathcal{N}_0$ 和 $(\check{\boldsymbol{x}}_i^{\top}, 1)\ i=1,\ldots,$ $N-n_0$ 运行逻辑回归，并得到系数估计 $\hat{}\,\boldsymbol{\beta}$。令 $\hat{r}(\boldsymbol{x}) = \exp\{\boldsymbol{x}^{\top}\hat{}\,\boldsymbol{\beta}\}$。步骤 3. 根据权重 $\hat{r}(\boldsymbol{x}(u)\, i)$ $/\sum_{i\in[N]\setminus\mathcal{N}_0}\hat{r}(\boldsymbol{x}(u)\, i)$，从 $\boldsymbol{x}(u)\, i, i \in [N] \setminus \mathcal{N}_0$ 中抽取 $n-n_0$ 个点。令 $(\boldsymbol{x}_i^{\top}, y_i)$, $i=1,\ldots,n$ 表示在步骤 1 和步骤 3 中抽取的协变量及其对应的响应变量。定义 $\mathcal{N}(tr) = \{n_0+1,\ldots,n\}$ 和 $\mathcal{N}(va)$ $= \{n_0/2+1,\ldots,n_0\}$。

算法 4 的步骤如下。首先，均匀选取少量无标签点，用于估计方差函数并构造验证集。随后，使用重要性抽样策略 [16] 选择其余带标签样本。令 $q_X(\boldsymbol{x})$ 表示未标记协变量 $\boldsymbol{x}_i^{(u)}$ 的分布。估计得到的比率 $\hat{}\,r(\boldsymbol{x})$ 可近似 $p_X(\boldsymbol{x})/q_X(\boldsymbol{x})$，从而仅利用无标签数据，就能从一个接近 p_X 的分布中进行抽样。

现在我们证明，在温和条件下，所提出的检索方案能够达到与最优抽样分布相同的方差阶。

条件 3（关于密度比的条件）。假设 $\boldsymbol{x}_i^{(u)}, i=1,\ldots,N$ 独立同分布于某个分布 $q_X(\boldsymbol{x})$，且满足 $\sup_{\boldsymbol{x}\in\mathcal{X}} q_X(\boldsymbol{x}) \le C$，并且 $p_X^*(\boldsymbol{x})/q_X(\boldsymbol{x}) = \exp\{\boldsymbol{x}^{\top}\boldsymbol{\beta}\}$，其中 $\|\boldsymbol{\beta}\|_2 \le C$，且 C 是一个正数常量。此外，$\boldsymbol{x}_i^{(u)}$ 的协方差矩阵记为 $\Sigma^{(u)}$，满足 $\Lambda_{\min}(\Sigma^{(u)}) \ge c_0 > 0$，其中 c_0 是某个正数常量。

引理 3（算法4中检索方案的速率最优性）。假设条件*1*、*2*和*3*成立。假设对于某个常量 $c_0 \le 0$，有 $N \ge n$, $h = n^{-c_0}$，且 $|\mathcal{X}_1| \le c\min\{\bar{\sigma}^2/(\tilde{r}_n^2 nh^d)\, 1/\sqrt{\log n}\}$，其中 $c<1$ 是某个正数常量。则存在某个正数常量 C，使得

$$\int_{\mathcal{X}} \mathbb{E}[\frac{\sigma^2(\boldsymbol{x}) + \theta_1^{(\sigma)} h^{\theta_2^{(\sigma)}}}{\max\{1, n_h(\boldsymbol{x})\}}] d\boldsymbol{x} \le C\frac{\bar{\sigma}^2}{nh^d}.$$

引理3 表明，通过对无标签协变量进行加权抽样，算法4 中的检索方案可达到与最优检索分布 $p_X^*(\boldsymbol{x})$ 相同的方差阶。

5. 模拟研究。本节中，我们开展了多项数值研究，以评估所提方法的实证表现。

5.1. 回归。在第一个实验中，我们将真实模型设为

$$f^*(\boldsymbol{x}) = \theta_1^*|x_1| + \theta_1^*|x_2 + 0.3|^{\theta_2^*}$$

对于 $\boldsymbol{\theta}^* = (1, 0.5)^{\top}$，目标区域为 $\mathcal{X} = [-0.5, 0.5]^2$。对于从目标模型中检索到的每个 $\boldsymbol{x}_i$，我们将其响应生成为 $y_i = f^*(\boldsymbol{x}_i) + \epsilon_i$，其中 $\epsilon_i \sim N(0,1)$ 相互独立，$i=1,\ldots,n$。

对于预训练模型，我们从 $[-1,1]^2$ 中均匀随机抽样 $\boldsymbol{x}_i^{(\mathrm{ptr})}$, $i=1,\ldots,N^{(\mathrm{ptr})}$，并

$$y_i^{(\mathrm{ptr})} = 0.8 f^*(\boldsymbol{x}_i^{(\mathrm{ptr})}) + \epsilon_i^{(\mathrm{ptr})}$$

对于 $\epsilon_i^{(\mathrm{ptr})} \sim N(0,1)$，我们基于 $\boldsymbol{x}_i^{(\mathrm{ptr})}$ 和 $y_i^{(\mathrm{ptr})}$、$i=1,\ldots,N^{(\mathrm{ptr})}$，通过核回归估计 $f^{(\mathrm{ptr})}$。作为比较，我们还评估了预训练模型和单任务方法的性能；后者仅基于从 $\mathcal{X}$ 中随机检索的 n 个样本计算回归函数。对于所提方法，我们设定 $n_0 = n/4$。为减轻样本划分造成的效率损失，我们不再进一步划分前 n_0 个样本，而是使用所有样本 $(\boldsymbol{x}_i^\top, y_i)$、$i \in [n_0]$ 来计算 $\hat{\sigma}^2(\boldsymbol{x})$，并将它们用作验证样本。为进行评估，我们从 $n_0+1,\ldots,n_t$ 中随机抽取测试协变量 $\boldsymbol{x}_i^{(t)}$，其中 $n_t = 500$，并计算真实条件均值 $f^*(\boldsymbol{x}_i^{(t)})$。对于任意估计量 $f(\cdot)$，我们报告其均值估计误差

$$\mathrm{MSE}(f) = \frac{1}{n_t}\sum_{i=1}^{n_t}\{f^*(\boldsymbol{x}_i^{(t)}) - f(\boldsymbol{x}_i^{(t)})\}^2.$$

我们考虑不同的 $N^{(\mathrm{ptr})}$ 水平以及采样预算 n。对于每种设置，我们独立重复上述实验 300 次，并在图 2中报告结果。代码可在 https://github.com/saili0103/FSP 获取。

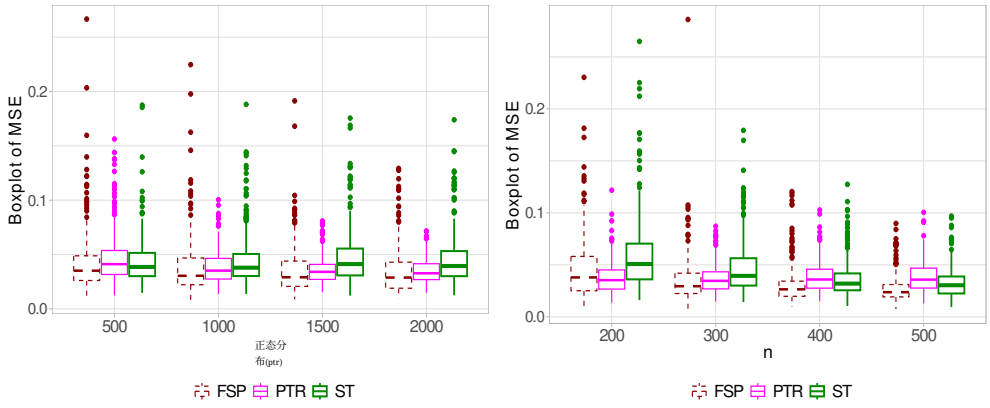


FIG 2. 不同预训练样本量 $N(\mathrm{PTR})$下，单任务方法（ST）、所提方法（FSP）和预训练模型（PTR）的均方误差箱线图。左图中的采样预算固定为 $n=300$，右图中的预训练样本量固定为 $N(\mathrm{PTR})=1000$。

从图2左侧面板可以看出，随着 $N^{(\mathrm{ptr})}$ 增大，预训练估计和所提出的估计的均方误差都会降低。这是因为二者都利用了预留样本，而这些样本对目标任务具有信息价值。单任务方法的均方误差保持不变，因为它没有整合预训练估计。从右侧面板可以看出，随着目标样本量n 增大，单任务估计和所提出的估计的均方误差都会降低，因为二者都利用了来自目标定义域的检索样本。这些结果展示了整合预训练模型的益处，以及我们提出的方法在校正预训练估计偏差方面的有效性。

5.2. 分类。我们进一步考虑一个分类任务 $\mathbb{P}(y_i = 1|\boldsymbol{x}_i) = f^*(\boldsymbol{x}_1)$，其中

$$f^*(\boldsymbol{x}) = \max(\min(\theta_1^*|x_1|^{\theta_2^*} + \theta_1^*|x_2 - 0.3|^{\theta_2^*} - 0.1, 0.9), 0)$$

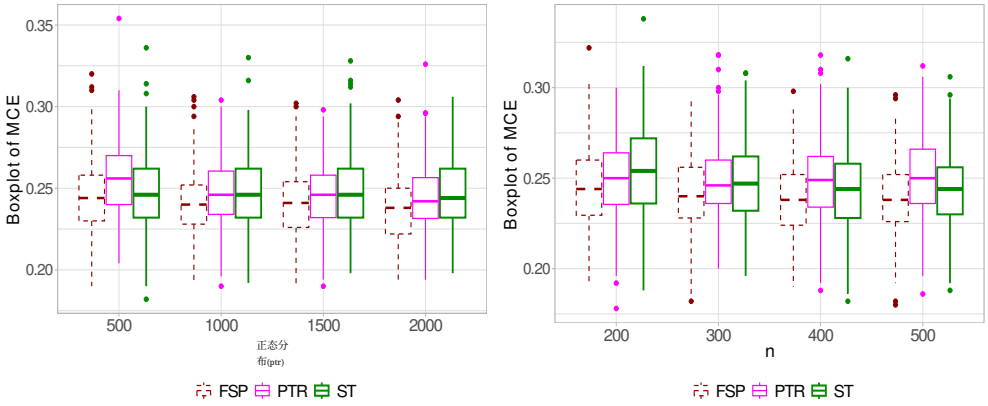
以及 $\boldsymbol{\theta}^* = (1, 0。6)^\top$。我们考虑目标区域 $\mathcal{X} = [-0.2, 0.8]^2$。对于预训练模型，我们首先从 $[-1, 1]^2$ 中均匀生成 $\boldsymbol{x}_i^{(\text{ptr})}$，并

$$\mathbb{P}(y_i^{(\text{ptr})} = 1|\boldsymbol{x}_i^{(\text{ptr})}) = f^*(\boldsymbol{x}_i^{(\text{ptr})}) + 0.1.$$

类似于第 5.1 节，我们从 $\mathcal{X}$中随机生成测试协变量 $\boldsymbol{x}_i^{(t)}$、$i = 1$、.....、n_t，并生成其响应 $y_i^{(t)}$，使其为伯努利随机变量，满足 $\mathbb{P}(y_i^{(t)} = 1|\boldsymbol{x}_i^{(t)} = \boldsymbol{x}) = f^*(\boldsymbol{x})$。对于任意估计量 $f(\cdot)$，其平均误分类误差定义为

$$\text{MCE}(f) = \frac{1}{n_t}\sum_{i=1}^{n_t}\left|y_i^{(t)} - \mathbb{I}(f(\boldsymbol{x}_i^{(t)}) \geq 0.5)\right|.$$

从图 3 中可以看出，在所有设置下，所提方法的平均误分类误差均小于另外两种方法。图中的总体模式与图 2 中的类似，并且我们所提方法的表现与理论结果一致。



F图 3. 单任务方法（ST）、所提方法（FSP）和预训练估计（PTR）在不同预训练样本量 $N(\text{ptr})$下的平均误分类误差箱线图。左图中的采样预算为 $n = 300$，右图中的预训练样本量为 $N(\text{ptr}) = 1000$。

5.3. 集成无信息预训练模型。我们评估了预训练模型不具备信息时所提方法的稳健性。真实模型 $f^*(\cdot)$ 与第5.1 节相同，但对于预训练模型，我们独立生成 $f(\text{ptr})(\boldsymbol{x}_i) \sim N(0, 1)$。在这种情况下，直接集成预训练模型可能会导致估计误差增大。

在表1中，我们报告了在此设置下不同方法的均方误差。可以看出，所提方法的估计误差与单任务方法的误差相当。结果表明，所提方法对抗性预训练模型具有鲁棒性，这与我们的理论分析一致。

N(PTR)	ST	FSP	PTR
500	0.043 (0.02)	0.051 (0.02)	0.712 (0.14)
1000	0.047 (0.03)	0.048 (0.02)	0.700 (0.10)
1500	0.045 (0.03)	0.046 (0.02)	0.707 (0.08)
2000	0.046 (0.03)	0.043 (0.02)	0.703 (0.07)

n	ST	FSP	PTR
200	0.056 (0.03)	0.063 (0.03)	0.703 (0.08)
300	0.046 (0.02)	0.047 (0.02)	0.693 (0.09)
400	0.037 (0.02)	0.039 (0.02)	0.713 (0.10)
500	0.032 (0.02)	0.035 (0.01)	0.698 (0.09)

表1 基于 300 次蒙特卡罗模拟，单任务方法（ST）、所提方法（FSP）和预训练模型（PTR）的均方误差均值（标准差）。左侧表格中的采样预算为 $n = 300$，右侧表格中的预训练样本量为 $N(\mathrm{ptr}) = 1000$。

6. 真实数据研究。在本节中，我们将所提方法应用于预测加利福尼亚州的房价。

该数据集包含来自 1990 年美国人口普查的人口普查街区组层级的 20640 条记录。目标是利用九个特征预测一个街区的房屋价格中位数。该数据最初发表于 [22]，并可在以下网址获取：https://www.kaggle.com/datasets/camnugent/california-housing-pric

e

我们研究了基于五个特征预测近湾区的房屋价格中位数，该区域包含 2290 个样本。这五个特征分别为：街区经度（lon）、街区纬度（lat）、街区房屋年龄中位数（age）、街区组总人口（pop）和街区家庭收入中位数（inc）。我们留出 1000 个样本作为测试数据。对于单任务方法，我们从剩余的 1290 个样本中随机抽取 500 条记录来构建非参数回归模型。对于个性化，我们首先假设这 1290 个样本均未标注，并运行算法 4以检索 $n = 500$ 个样本。我们将这 500 个带标签样本用作检索数据。

我们考虑三个预训练模型。对于第一个模型，我们向 DeepSeek-V3.2 描述这一预测任务，并请人工智能模型给出一个估计加利福尼亚州房价中位数的公式。详细提示词和人工智能的响应可见：https://chat.deepseek.com/share/q55zlavzp48lrc9ntm。DeepSeek-V3.2 给出的预测规则为

$$20 + 15 \times \mathrm{inc} - 0.3 \times \mathrm{age} + 0.005 \times \mathrm{pop} - 1.5 \times \mathrm{lon} - 0.8 \times \mathrm{lat} + 0.02 \times \mathrm{lat}^2 - 0.1 \times \mathrm{inc}^{\mathrm{n}} \times \mathrm{lo}$$

对于第二个预训练模型，我们使用加利福尼亚州湾区以外的样本训练随机森林（RF） [4]，样本量为18350。对于第三个预训练模型，我们考虑LightGBM[13], 这一采用基于树的学习算法的流行梯度提升框架，并同样基于湾区以外的样本对其进行训练。

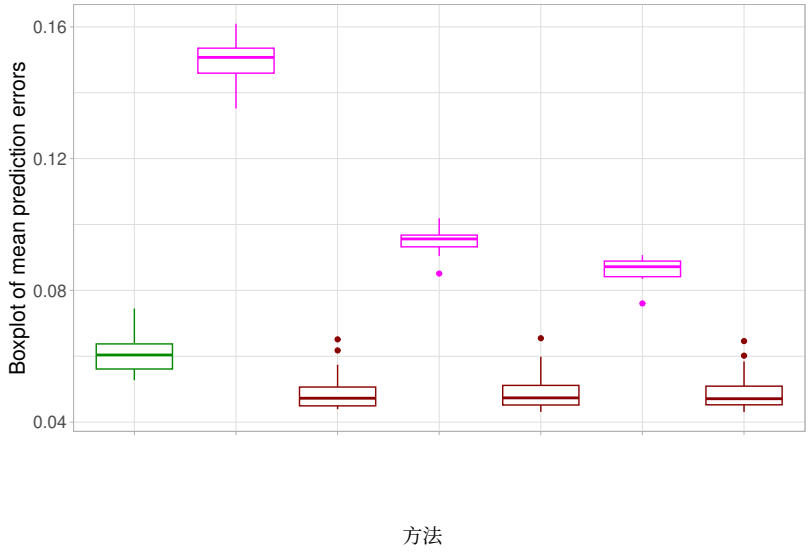


FIG 4. 各方法平均预测误差的箱线图。每个箱线图均基于对测试样本和训练样本进行的 20 次随机划分。比较的方法包括单任务非参数回归（ST）、预训练模型（DeepSeek、随机森林和 LightGBM）以及少样本个性化方法（fsp.DeepSeek、fsp.RF 和 fsp.lightGBM）。

这些实验的结果如图4所示。我们看到，所有预训练模型的预测误差都大于单任务方法。这表明，DeepSeek推导出的模型并不准确，并且湾区与加利福尼亚州其他地区之间的房价存在分布偏移。尤其是，DeepSeek给出的模型在很大程度上独立于数据，其表现不如其他数据驱动模型。然而，采用所提出的个性化方法后，所有预训练模型的预测性能均得到显著提升，超过了单任务方法的准确性。这项分析表明，所提出的个性化方案能够有效借用其他数据集中的信息，并融合来自其他定义域的知识，例如地理和经济学知识。因此，个性化为降低目标域所需的数据收集成本提供了一种途径。

7. 讨论。本文提出了一个少样本个性化的统计框架，并针对非参数回归提出了一种极小极大速率最优个性化方法。我们证明，整合大规模预训练模型能够实现比从头训练更高的估计精度，同时保持稳健性。个性化问题也可以针对其他统计目的进行研究。例如，研究参数模型中的估计与预测个性化颇有意义，如高维线性模型和广义线性模型。

参考文献

[1] ANGELOPOULOS, A. N., BATES, S., FANNJIANG, C., JORDAN, M. I. 和 ZRNIC, T.（2023）。预测驱动推断。《科学》382，669–674。[2] ANGELOPOULOS, A. N., DUCHI, J. C. 和 ZRNIC, T.（2023）。Ppi++：高效的预测驱动推断。arXiv 预印本 arXiv:2311.01453。[3] BAKTASHMOTLAGH, M., HARANDI, M. T., LOVELL, B. C. 和 SALZMANN, M.（2013）。通过领域不变投影实现无监督领域自适应。载于《电气电子工程学会国际计算机视觉会议论文集》，769–776。[4] BREIMAN, L.（2001）。随机森林。《机器学习》45，5–32。

[5] BÜHLMANN, P.（2020）。不变性、因果性与稳健性。《统计科学》35，404–426。[6] CAI, T. T. 和 WEI, H.（2021）。非参数分类的迁移学习。《统计年鉴》49，100–128。[7] CASELLA, G., ROBERT, C. P. 和 WELLS, M. T.（2004）。广义接受-拒绝抽样方案。讲义-专著系列，342–347。[8] DAI, D., RIGOLLET, P. 和 ZHANG, T.（2012）。使用贪心Q聚合进行偏差最优学习。《统计年鉴》，1878–1905。[9] FAN, J.（1993）。局部线性回归平滑器及其极小极大效率。《统计年鉴》，196–216。[10] FAN, J., FANG, C., GU, Y. 和 ZHANG, T.（2024）。环境不变线性最小二乘。《统计年鉴》52，2268–2292。[11] GILKS, W. R. 和 WILD, P.（1992）。用于吉布斯抽样的自适应拒绝抽样。《英国皇家统计学会杂志：C辑（应用统计）》41，337–348。[12] JI, W., LEI, L. 和 ZRNIC, T.（2025）。将预测作为替代指标：在人工智能时代重新审视替代结局。arXiv 预印本 arXiv:2501.09731。[13] KE, G., MENG, Q., FINLEY, T., WANG, T., CHEN, W., MA, W., YE, Q. 和 LIU, T.-Y.（2017）。LightGBM：一种高效的梯度提升决策树。《神经信息处理系统进展》30。[14] KIM, J. 和 YANG, Y.（2024）。错位响应下的大语言模型少样本个性化。arXiv 预印本 arXiv:2406.18678。[15] KIM, J. 和 YANG, Y.（2024）。错位响应下的大语言模型少样本个性化。载于 NAACL 长文，2025。[16] KLOEK, T. 和 VAN DIJK, H. K.（1978）。方程组参数的贝叶斯估计：蒙特卡洛积分的应用。《计量经济学：计量经济学会杂志》，1–19。[17] LECUÉ, G. 和 RIGOLLET, P.（2014）。基于Q聚合的最优学习。《统计年鉴》42，211–224。[18] LI, S., CAI, T. T. 和 LI, H.（2022）。高维线性回归的迁移学习：预测、估计与极小极大最优性。《英国皇家统计学会杂志B辑：统计方法学》84，149–173。[19] LI, S. 和 ZHANG, L.（2025）。具有低秩结构的多维领域泛化。《美国统计学会杂志》120，2522—2534。[20] LI, S., ZHANG, L., CAI, T. T. 和 LI, H.（2024）。知识迁移下高维广义线性模型的估计与推断。《美国统计学会杂志》119，1274–1285。[21] LIU, J., QIU, Z., LI, Z., DAI, Q., YU, W., ZHU, J., HU, M., YANG, M., CHUA, T.-S. 和 KING, I.（2025）。个性化大语言模型综述：进展与未来方向。arXiv 预印本 arXiv:2502.11528。[22] PACE, R. K. 和 BARRY, R.（1997 33 291 297）。稀疏空间自回归。《统计与概率快报》–。[23] PETERS, J., BÜHLMANN, P. 和 MEINSHAUSEN, N.（2016）。利用不变预测进行因果推断：识别与置信区间。《英国皇家统计学会杂志B辑（统计方法学）》，947–1012。[24] REEVE, H. W., CANNINGS, T. I. 和 SAMWORTH, R. J.（2021）。自适应迁移学习。《统计年鉴》49，3618–3649。[25] ROJAS-CARULLA, M., SCHÖLKOPF, B., TURNER, R. 和 PETERS, J.（2018）。用于因果迁移学习的不变模型。《机器学习研究杂志》19，1309–1342。[26] ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. 和 PETERS, J.（2021）。锚点回归：异质数据与因果性的交汇。《英国皇家统计学会杂志B辑：统计方法学》83，215–246。[27] SALEMI, A. 等（2024）。LaMP：当大语言模型遇上个性化。载于2024年计算语言学协会会议论文集（ACL）。[28] SHENFELD, I., FALTINGS, F., AGRAWAL, P. 和 PACCHIANO, A.（2025）。通过奖励分解实现语言模型个性化。arXiv 预印本 arXiv:2503.06358。[29] STEINFELDT, J., BUERGEL, T., LOOCK, L., KITTNER, P., RUYOGA, G., ZU BELZEN, J. U., SASSE, S., STRANGALIES, H., CHRISTMANN, L., HOLLMANN, N. 等（2022）。基于神经网络整合多基因与临床信息：开发并验证英国生物样本库队列中主要不良心血管事件10年风险预测模型。《柳叶刀·数字健康》4，e84–e94。[30] TIAN, Y. 和 FENG, Y.（2023）。高维广义线性模型下的迁移学习。《美国统计学会杂志》118，2684–2697。

[31] TSYBAKOV, A. B. (2008). 非参数估计导论。Springer，纽约，纽约州。[32] ZHANG, Z., ROSSI, R. A., KVETON, B., SHAO, Y., YANG, D., ZAMANI, H., DERNONCOURT, F., BAR- ROW, J., YU, T., KIM, S., ZHANG, R., GU, J., DERR, T., CHEN, H., WU, J., CHEN, X., WANG, Z., MITRA, S., LIPKA, N., AHMED, N. 和 WANG, Y. (2024). 大型语言模型的个性化：综述。arXiv 预印本 arXiv:2411.00027。[33] ZHAO, Y., YU, G., WANG, J., DOMENICONI, C., GUO, M., ZHANG, X. 和 CUI, L. (2022). 个性化联邦少样本学习。IEEE神经网络与学习系统汇刊 35 2534–2544。[34] ZRNIC, T. 和 CANDÈS, E. J. (2024). 主动统计推断。国际机器学习会议 62993–63010。