arXiv:2301.00346v1 [cs.LG] 1 Jan 2023

异质因果效应联邦学习的自适应核方法

ThanhVinh Vo[1] **Arnab Bhattacharyya**[1] **Young Lee**[2] **Tze-Yun Leong**[11]新加坡国立大学计算机学院[2]Roche AG 和哈佛大学 {votv,arnabb,leongty}@nus.edu.sg

摘要

我们提出了一种新的因果推断框架，用于在联邦设置下从多个去中心化数据源中学习因果效应。我们提出了一种自适应迁移算法，利用随机傅里叶特征将损失函数分解为多个分量，每个分量对应一个数据源，从而学习各数据源之间的相似性。不同数据源的分布可能各不相同；因果效应则以独立且系统化的方式纳入模型。所提出的方法通过迁移系数估计数据源之间的相似性，因此无需预先掌握相似性度量的信息。无需在数据源之间共享原始训练数据，即可估计异质因果效应，从而最大限度地降低隐私泄露风险。我们还提供了极小极大下界，以评估从不同数据源中学习到的参数质量。实证结果表明，所提出的方法在分布不同的去中心化数据源上优于基线方法。

1 引言

自然科学和社会科学中提出的许多重要问题本质上都涉及因果关系：轻度新冠肺炎感染对肺部和脑功能有何长期影响？每日空气污染如何影响死亡率？福利政策会如何影响少数群体的就业率？因果推断已应用于多个领域，包括经济学（Finkelstein 和 Hendren，2020）、医学（Henderson 等，2016；Powers 等，2018）以及社会福利（Gutman 等，2017）。准确估计因果效应所需的大量实验数据和／或观测数据往往分布在不同地点。多数情况下，由于某些固有的组织或政策限制，无法合并这些数据源以进行集中式处理。例如，在许多国家，癌症患者的医疗或健康记录都由当地医院严格保密；医院之间不允许直接交换或共享这些记录，尤其是出于研究目的时（Gostin 等，2009）。主要研究问题是：如何安全地访问这些不同的数据源，以构建有效的全局因果效应估计量，同时平衡数据隐私和保密性遭到侵犯的风险？

当前的因果推断方法（例如，Shalit 等人，2017；Yao 等人，2018）要求将共享数据集中到一处进行处理。当前的联邦学习算法（例如，Sattler 等人，2020；Wang 等人，2020）支持基于非独立同分布（*non iid*）数据协同学习联合模型；然而，这些算法无法直接支持因果推断，因为不同数据源的数据分布可能存在差异，从而导致因果效应估计产生偏差。例如，来自两家不同医院的癌症患者在人口统计特征和平均年龄方面可能存在显著差异。如果合并这两个数据集来支持因果推断，其中一种分布可能会占据主导地位，进而导致因果效应估计产生偏差。

第36届神经信息处理系统大会（NeurIPS 2022）。

我们提出了一种新方法，用于从多个分散且分布存在差异的数据源进行联邦因果推断。我们的贡献总结如下：

• 我们提出了一种新的联邦因果推断算法，称为 CausalRFF [1]，该算法基于结构因果模型（SCM）（Pearl 2009a），利用随机傅里叶特征（Rahimi 和 Recht 2007）进行因果效应的联邦估计。随机傅里叶特征使目标函数能够拆分为多个分量，从而支持模型的联邦训练。

• 我们使用 CausalRFF，通过自适应核函数，基于具有不同分布的数据源进行联邦因果推断；推断过程无需在数据源之间共享原始数据，因此能够最大限度地降低隐私泄露风险。

• 我们给出了极小极大下界，以阐明联邦因果推断框架中估计和优化过程的局限。

我们的工作是迈向隐私保护因果推断的重要一步。本文最后探讨了将 CausalRFF 与多方差分隐私相结合的可能性。

2 相关工作

关于以保护隐私的方式将因果推断与联邦学习相结合，目前相关研究还很少。

关于因果推断： Hill（2011）；Alaa 和 van der Schaar（2017，2018）；Shalit 等人（2017）；Yoon 等人（2018）；Yao 等人（2018）；Künzel 等人（2019）；Nie 和 Wager（2020）提出了直接从本地数据源学习因果效应的方法；这些方法采用标准的可忽略性假设（Rosenbaum 和 Rubin，1983）。Louizos 等（2017）；Madras 等人（2019）对 Pearl（1995）的结构因果模型（SCM）进行了改造，以便在存在潜在混杂变量的情况下估计因果效应。

我们的工作与可迁移性概念密切相关，并对其进行了扩展。Pearl 和 Bareinboim（2011）；Bareinboim 和 Pearl（2016）；Lee 等人（2020）及相关工作对如何利用一个总体中的干预工具计算另一个总体中的因果效应进行了形式化，并提供了理论分析。Lee 等人（2020）将可迁移性推广为：根据可观测变量子集上的观察性分布和干预分布，支持识别目标域中的因果效应，从而为基于观察性数据和实验数据得出结论奠定了基础（Tsamardinos 等人，2012；Bareinboim 和 Pearl，2016）。对于来自多个去中心化且分布各异的数据源、无法在中心位置合并或处理的数据，现有研究尚未涉及其因果推断问题。近期，Aglietti 等人（2020）在源总体上开展随机实验以收集数据，随后对源总体的干预数据与目标总体的观察性数据建立联合模型并进行估计。我们的工作有所不同：我们不使用随机数据，而是仅通过观察性数据进行迁移来估计因果效应。这对应于现实生活中的一种重要场景，即只能获得回顾性观察数据，例如与新冠肺炎相关的病例与干预记录、银行与金融交易记录。

关于联邦学习： 联邦学习能够在训练数据保留在各数据源、保持分散的同时，协作学习一个共享的预测模型（McMahan 等人，2017）。一些联邦学习方法结合联邦随机梯度下降（Shokri 和 Shmatikov，2015）与联邦平均（McMahan 等人，2017），以解决回归问题 Álvarez 等人（2019）；Zhe 等人（2019）；de Wolff 等人（2020）；Joukov 和 Kuli´c（2020），以及 Hard 等人（2018）；Zhao 等人（2018）；Sattler 等人（2019）；Mohri 等人（2019）。近期的联邦学习算法支持基于非独立同分布数据协作学习联合深度神经网络模型（Sattler 等人，2020；Wang 等人，2020）。然而，所有这些算法都不直接支持因果推断，因为不同数据源可能具有不同的分布，从而导致因果效应估计产生偏差。针对联邦因果效应估计的研究仍然很少。Vo 等人（2022）提出了一种贝叶斯方法，基于高斯过程估计因果效应的后验分布，但该方法不允许各数据源具有不同的分布。Xiong 等人（2021）估计了平均处理效应（ATE）和处理组平均处理效应（ATT），并假设混杂因素均可观测。另一方面，我们的工作在存在潜在混杂因素的情况下，估计条件平均处理效应（CATE）（亦称个体处理效应，ITE）和平均处理效应（ATE）。我们利用随机傅里叶

[1] 数据源 代码： https://github.com/vothanhvinh/Causa lRFF

特征构建联邦设置下的因果推断一体化框架，该框架允许数据分布存在差异。

3 所提模型

In t本节首先详细介绍问题形式化。随后，我们将给出所关注的因果效应 t
and 及其估计方案。最后，我们将描述相关假设和结构方程 .

3.1 问题描述

问题设置与符号。假设我们有 m 个数据源，每个数据源表示为 $\mathrm{D}^{\mathrm{s}} = \{(w_i^{\mathrm{s}}, y_i^{\mathrm{s}}, \boldsymbol{x}_i^{\mathrm{s}})\}_{i=1}^{n_{\mathrm{s}}}$，其中 $\mathrm{s} \in \mathcal{S} := \{\mathrm{s}_1, \mathrm{s}2, \ldots, \mathrm{s}_m\}$，变量 w_i^{s}、y_i^{s} 和 $\boldsymbol{x}_i^{\mathrm{s}}$ 分别表示数据源 s 中个体 i 的治疗分配、与处理相关的观测结果和协变量。这些数据源 D^{s} 位于不同位置，其分布可能完全不同。所有数据源共享如图 1 所示的同一因果图，但数据分布可能不同，例如，$p_{\mathrm{s}_1}(\boldsymbol{x}, w, y) \neq p_{\mathrm{s}_2}(\boldsymbol{x}, w, y)$，其中 $p_{\mathrm{s}_1}(\cdot)$ 和 $p_{\mathrm{s}_2}(\cdot)$ 分别表示两个数据源 s_1 和 s_2，上的两种分布。类似地，相对于这些变量的边际分布和条件分布也可能不同（或相似）。我们的目标是开发一个满足以下两个条件的全局因果推断模型：（i）因果推断模型可以在隐私环境中进行训练，即每个数据源的数据不会与外部人员共享；（ii）因果推断模型能够整合多个数据源的数据，以改进每个特定数据源中的因果效应估计。

关注的因果效应。给定在上述设定下训练的因果模型，我们感兴趣的是估计条件平均处理效应（CATE）[2] 和平均处理效应（ATE）。令 Y、W、X 分别表示结果、处理和代理变量的随机变量。于是，CATE 和 ATE 定义如下（Louizos 等人，2017；Madras 等人，2019）

$$\tau(\boldsymbol{x}) := E\big[Y|\mathrm{do}(W{=}1), X{=}\boldsymbol{x}\big] - E\big[Y|\mathrm{do}(W{=}0), X{=}\boldsymbol{x}\big], \qquad \tau := E[\tau(X)], \tag{1}$$

其中，$\mathrm{do}(W{=}w)$ 表示对个体实施处理 $w \in \{0, 1\}$。此定义沿用 Louizos 等（2017）和 Madras 等人（2019）。给定一组 n 新的个体，其协变量/观测到的代理变量为 $\{\boldsymbol{x}_i\}_{i=1}^n$，该子群体中的 CATE 和 ATE 分别通过 $\tau(\boldsymbol{x}_i)$ 和 $\tau = \sum_{i=1}^n \tau(\boldsymbol{x}_i)/n$ 得出。

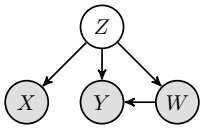


图 1：包含潜在混杂因素 Z、处理 W、结果Y、协变量/代理变量的因果图 X.

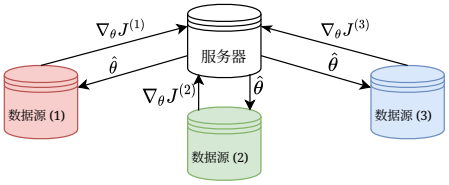


图 2：我们提出的包含三个数据源的模型示例。目标函数 $J \simeq J^{(1)} + J^{(2)} + J^{(3)}$ 被分解为 3 个分量，每个分量对应一个数据源。

估计条件平均处理效应和平均处理效应的核心任务是求得 $E[Y|\mathrm{do}(W = w)\ X = \boldsymbol{x}]$。由于每个数据源的数据分布可能彼此不同（或相似），我们使用记号$E[Y|\mathrm{do}(W = w^{\mathrm{s}}, X = \boldsymbol{x}^{\mathrm{s}}]$ 表示数据源 s 中某个个体在对 W进行干预时结果 Y的期望。考虑到潜在混杂因素 Z的存在，我们可以进一步利用 do-演算（Pearl 1995）展开这一量。具体而言，根据后门调整公式，我们有

$$E\big[Y|\mathrm{do}(W = w^{\mathrm{s}}), X = \boldsymbol{x}^{\mathrm{s}}\big] = \int E\big[Y|W = w^{\mathrm{s}}, Z = \boldsymbol{z}^{\mathrm{s}}\big] p_{\mathrm{s}}(\boldsymbol{z}^{\mathrm{s}}|\boldsymbol{x}^{\mathrm{s}}) d\boldsymbol{z}^{\mathrm{s}}. \tag{2}$$

式（2）表明，对于每个数据源 s，如果我们能够找到条件分布$p_{\mathrm{s}}(y^{\mathrm{s}}|w^{\mathrm{s}},\ \boldsymbol{z}^{\mathrm{s}})$ 和 $p_{\mathrm{s}}(\boldsymbol{z}^{\mathrm{s}}|\boldsymbol{x}^{\mathrm{s}})$，那么因果效应就是可识别的。第二个分布还可以进一步展开为 $p_{\mathrm{s}}(\boldsymbol{z}^{\mathrm{s}}|\boldsymbol{x}^{\mathrm{s}}) = \sum_{w^{\mathrm{s}}} \int p_{\mathrm{s}}(\boldsymbol{z}|\boldsymbol{x}^{\mathrm{s}},$ $y_i^{\mathrm{s}},\ w^{\mathrm{s}})p_{\mathrm{s}}(y^{\mathrm{s}}|\boldsymbol{x}^{\mathrm{s}},\ w^{\mathrm{s}})p_{\mathrm{s}}(w^{\mathrm{s}}|\boldsymbol{x}^{\mathrm{s}})\mathrm{d}y^{\mathrm{s}}$。遵循前向采样策略，接下来需要找到以下分布

$$p_{\mathrm{s}}(w^{\mathrm{s}}|\boldsymbol{x}^{\mathrm{s}}), \qquad p_{\mathrm{s}}(y^{\mathrm{s}}|\boldsymbol{x}^{\mathrm{s}}, w^{\mathrm{s}}), \qquad p_{\mathrm{s}}(\boldsymbol{z}^{\mathrm{s}}|\boldsymbol{x}^{\mathrm{s}}, y^{\mathrm{s}}, w^{\mathrm{s}}), \qquad p_{\mathrm{s}}(y^{\mathrm{s}}|w^{\mathrm{s}}, \boldsymbol{z}^{\mathrm{s}}), \tag{3}$$

[2]也称为个体处理效应（ITE）

然后从这些估计得到的分布中系统地抽取样本，以获得给定 do($W = w^s$) 和 $X = \boldsymbol{x}^s$ 时 Y的经验期望。

可识别性。如果我们能够学习式（3）中的分布，那么条件平均处理效应和平均处理效应就是可识别的；这些分布涉及潜在混杂因素 Z。Louizos 等（2017）证明，如果 Z 与观测变量 X存在关联，就可以做到这一点；在许多情况下，该模型都是可识别的，例如：Z 是类别变量且 X 是高斯混合模型（Anandkumar 等人，2014）； X 包含 Z 的三个独立视图（Goodman，1974；Allman 等人，2009；Anandkumar 等人，2012）； Z 是多元二元变量，且 X 是 Z 的含噪函数（Jernite 等人，2013；Arora 等人，2017），等等。遵循 Louizos 等人（2017）和 Madras 等人（2019）的研究，我们采用变分自编码器（VAE）思想下的变分推断来恢复潜在混杂因素，因为它能够学习丰富的潜变量模型类别，从而恢复因果效应。我们工作的可识别性与现有文献密切相关，不过我们的主要贡献在于模型的联邦设置。可识别性的证明请参见附录。

3.2 因果图与假设

由于我们的方法采用图 1 中的因果图所示的结构因果模型方法，因此结构因果模型的公理和性质蕴含了一些隐含假设：**(A1)** 一致性： $W = w \Longrightarrow Y(w) = Y$，这由结构因果模型的公理推得。**(A2)** 无干扰：对某个受试者的处理不会影响另一个受试者的结果。这是因为结果只有一个处理节点 作为其父节点。**(A3)** 正值性：每个受试者被分配到任一处理的概率都大于零。这些假设是任何因果推断算法中的标准假设。更多讨论见 Pearl (2009a,b)；Morgan 和 Winship (2015)。针对我们提出的联邦设置，我们作出以下两个额外假设：

(A4) 所有数据源中的个体 i 具有相同的共同 协变量 。
(A5) Any 个体不会同时存在于多个数据源中。

假设 **(A4)** 已在我们的设定中隐含成立，因为所有数据源共享同一因果图。这是一个合理的假设，因为我们旨在基于所有数据源构建统一模型。例如，Choudhury 等人（2019）、Vaid 等人（2020）和 Flores 等人（2020）中的去中心化数据满足联邦学习的这一假设。假设**(A5)** 旨在确保训练模型时不会有个体对其他个体产生支配性影响。例如，如果某个个体出现在所有数据源中，训练得到的模型就会受到该个体数据的偏倚影响（由于使用了比其他个体更多的该个体数据，从而造成不平衡）。因此，这一条件可以确保不存在此类偏倚。在实践中，假设 **(A5)** 有时并不成立。为了解决这一问题，我们会执行预训练步骤，以排除这类重复个体。此步骤将使用单向哈希函数执行安全匹配过程，从而识别重复个体。预训练步骤的详情见附录。

3.3 结构方程

本节介绍如何对因果关系进行建模。由于 Z 是因果图中的根节点，我们将其建模为多元正态分布：对于所有数据源， $Z \sim \mathrm{N}(\boldsymbol{\mu}, \sigma_z^2 \mathbf{I}_{d_z})$。下面详细说明 Y、 W 和 X的结构方程。令 V为一个单变量，用来表示因果图（图 1）中的一个节点或节点的一个维度，即 V可以是 Y、 W 或X的一个维度。令 $\mathrm{pa}(V)$ 表示因果图中 V的父变量集合，也就是通过有向边指向V的节点。我们按以下两种情况对 V的结构方程进行建模：

$$\text{if } V \text{ is continuous:} \quad V = f_v(\mathrm{pa}(V)) + \epsilon_v, \quad \text{if } V \text{ is binary:} \quad V = \mathbb{1}[\varphi(f_v(\mathrm{pa}(V))) > \epsilon_v], \tag{4}$$

其中，前一种情况为 $\epsilon_v \sim \mathrm{N}(0, \sigma_v^2)$，后一种情况为 $\epsilon_v \sim \mathrm{U}[0,1]$； $\varphi(\cdot)$ 是逻辑函数，$\mathbb{1}(\cdot)$ 是指示函数。后一种情况意味着，给定 $\mathrm{pa}(V)$ 后， V 服从伯努利分布，其参数为 $p(V = 1 | \mathrm{pa}(V)) = \varphi(f_v(\mathrm{pa}(V)))$。此外，如果 $W \in \mathrm{pa}(V)$，那么我们进一步建模

$$f_v(\mathrm{pa}(V)) = (1 - W) f_{v0}(\mathrm{pa}(V) \setminus \{W\}) + W f_{v1}(\mathrm{pa}(V) \setminus \{W\}). \tag{5}$$

示例。 如果 $Y \in \mathbb{R}, W \in \{0,1\}$ 和 $X_k \in \mathbb{R}$ （X_k 是 X的第 k维），那么结构方程如下：

$$Y = (1 - W) f_{y0}(Z) + W f_{y1}(Z) + \epsilon_y, \qquad W = \mathbb{1}[\varphi(f_w(Z)) > \epsilon_w], \qquad X_k = f_{x_k}(Z) + \epsilon_{X_k}.$$

在接下来的章节中，我们将介绍如何在联邦设置下学习函数 $f_v (v \in \{y0, y1, w, x\})$，并利用这些函数估计关注的因果效应。

4 CausalRFF：自适应联邦推断算法

本节将介绍一种新的联邦算法，用于学习式（3）中的分布。核心任务是将目标函数分解为多个分量，每个分量都与一个数据源相关联。

4.1 学习涉及潜在混杂因素的分布

为了估计因果效应，我们需要估计公式（3）中详述的四个量。本节介绍如何学习 $p_{\mathsf{s}}(\boldsymbol{z}^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}}, y^{\mathsf{s}}, w^{\mathsf{s}})$ 和 $p_{\mathsf{s}}(y^{\mathsf{s}}|w^{\mathsf{s}}, \boldsymbol{z}^{\mathsf{s}})$。由于边际似然没有解析形式，我们使用最大化证据下界（ELBO）的变分推断来学习上述分布

$$\mathcal{L} = \sum_{\mathsf{s}\in\mathcal{S}} \sum_{i=1}^{n_{\mathsf{s}}} \Big(E_q\big[\log p_{\mathsf{s}}(y_i^{\mathsf{s}}|w_i^{\mathsf{s}}, \boldsymbol{z}_i^{\mathsf{s}}) + \log p_{\mathsf{s}}(w_i^{\mathsf{s}}|\boldsymbol{z}_i^{\mathsf{s}}) + \log p_{\mathsf{s}}(\boldsymbol{x}_i^{\mathsf{s}}|\boldsymbol{z}_i^{\mathsf{s}})\big] - \mathrm{KL}[q(\boldsymbol{z}_i^{\mathsf{s}})\|p(\boldsymbol{z}_i^{\mathsf{s}})] \Big), \quad (6)$$

其中，$q(\boldsymbol{z}^{\mathsf{s}}) = \mathrm{N}(\boldsymbol{z}^{\mathsf{s}};\ f_q(y^{\mathsf{s}},\ w^{\mathsf{s}},\ \boldsymbol{x}^{\mathsf{s}}),\ \sigma_q^2\mathbf{I})$ 是变分后验分布。函数 $f_q(\cdot)$ 建模如下：$f_q(y^{\mathsf{s}},\ w^{\mathsf{s}},\ \boldsymbol{x}^{\mathsf{s}}) = (1-w^{\mathsf{s}})f_{q0}(y^{\mathsf{s}},\ \boldsymbol{x}^{\mathsf{s}}) + w^{\mathsf{s}} f_{q1}(y^{\mathsf{s}},\ \boldsymbol{x}^{\mathsf{s}})$，其中 f_{q0} 和 f_{q1} 是需要学习的两个函数。密度函数 $p_{\mathsf{s}}(y^{\mathsf{s}}|w^{\mathsf{s}},\ \boldsymbol{z}^{\mathsf{s}})$、$p_{\mathsf{s}}(w^{\mathsf{s}}|\boldsymbol{z}^{\mathsf{s}})$ 和 $p_{\mathsf{s}}(\boldsymbol{x}^{\mathsf{s}}|\boldsymbol{z}^{\mathsf{s}})$ 根据第 3.3 节所述的结构方程获得。有关证据下界推导的详细信息，请参阅附录。

自适应建模。 由于来自每个数据源的观测数据可能服从不同（或相似）的分布，我们将分别对其建模，并自适应地学习这些分布之间的相似性。具体而言，我们提出一种基于核的方法来学习这些分布。为此，我们首先通过使用重参数化技巧（Kingma 和 Welling，2013）为每个潜在混杂因素生成 M 个样本Z，从证据下界 $\mathcal{L}$ 的负值中得到经验$\widehat{\ }$损失函数 $\mathcal{L}$：$\boldsymbol{z}_i^{\mathsf{s}}[l] = f_q(y_i^{\mathsf{s}},\ w_i^{\mathsf{s}},\ \boldsymbol{x}_i^{\mathsf{s}}) + \sigma_q \epsilon_i^{\mathsf{s}}[l]$，其中 $\epsilon_i^{\mathsf{s}}[l]$取自标准正态分布。我们得到一个完整数据集

$$\widetilde{\mathsf{D}}^{\mathsf{s}} = \bigcup_{l=1}^{M} \big\{(w_i^{\mathsf{s}}, y_i^{\mathsf{s}}, \boldsymbol{x}_i^{\mathsf{s}}, \boldsymbol{z}_i^{\mathsf{s}}[l])\big\}_{i=1}^{n_{\mathsf{s}}}, \quad \forall \mathsf{s} \in \mathcal{S}. \quad (7)$$

利用这个完整数据集，我们最小化以下目标函数

$$J = \widehat{\mathcal{L}} + \sum_{c\in\mathcal{A}} R(f_c) \quad (8)$$

关于 f_c，其中 $\mathcal{A} = \{y0, y1, w, x, q0, q1\}$，且 $R(\cdot)$表示一个正则化项。J 的最小化结果将得到 f_c的以下形式

$$f_c(\boldsymbol{u}^{\mathsf{s}}) = \sum_{\mathsf{v}\in\mathcal{S}} \sum_{j=1}^{n_{\mathsf{v}}\times M} \kappa(\boldsymbol{u}^{\mathsf{s}}, \boldsymbol{u}_j^{\mathsf{v}})\boldsymbol{\alpha}_j^{\mathsf{v}}, \quad (9)$$

˜其中，$\boldsymbol{u}_j^{\mathsf{v}}$ 由数据集 D^{v} 中的第 j 个元组得到。详情请参见附录。由于来自不同数据源的数据可能来自完全不同（或相似）的分布，我们将使用自适应核来衡量它们的相似性。具体而言，令 $k(\boldsymbol{u}^{\mathsf{s}},\ \boldsymbol{u}^{\mathsf{v}})$ 为典型核函数，例如平方指数核、有理二次核或 Matérn 核。公式（9）中使用的核如下：$\kappa(\boldsymbol{u}^{\mathsf{s}},\ \boldsymbol{u}^{\mathsf{v}}) = \lambda^{\mathsf{s},\ \mathsf{v}} k(\boldsymbol{u}^{\mathsf{s}},\ \boldsymbol{u}^{\mathsf{v}})$，当 $\mathsf{s} \neq \mathsf{v}$ 时；否则，$\kappa(\boldsymbol{u}^{\mathsf{s}},\ \boldsymbol{u}^{\mathsf{v}}) = k(\boldsymbol{u}^{\mathsf{s}},\ \boldsymbol{u}^{\mathsf{v}})$，其中$\lambda^{\mathsf{s},\ \mathsf{v}} \in [0,1]$ 为自适应因子，由观测数据学习得到。

备注。 公式（9）表明，计算 $f_c(\boldsymbol{u}^{\mathsf{s}})$ 需要收集所有数据源中的全部数据点，因此无法在联邦设置下优化公式（8）中的目标函数。接下来，我们介绍一种称为随机傅里叶特征的方法来解决这一问题。

随机傅里叶特征。 我们将说明如何把随机傅里叶特征（Rahimi 和 Recht，2007）应用于我们的模型。设 $k(\boldsymbol{u}, \boldsymbol{u}')$ 为任意平移不变核（例如平方指数核、有理二次核或 Matérn 核）。那么，根据 Bochner 定理（Wendland，2004，定理 6.6），它可以写成如下形式：

$$k(\boldsymbol{u}, \boldsymbol{u}') = \int e^{\mathrm{i}\boldsymbol{\omega}^\top(\boldsymbol{u}-\boldsymbol{u}')} s(\boldsymbol{\omega}) d\boldsymbol{\omega} = \int \cos\left(\boldsymbol{\omega}^\top(\boldsymbol{u}-\boldsymbol{u}')\right) s(\boldsymbol{\omega}) d\boldsymbol{\omega}, \quad (10)$$

其中，$s(\boldsymbol{\omega})$ 是与该核相关的谱密度函数（一些常见核的谱密度请参见附录）。最后一个等号是因为核函数为实值且对称。此类核可以近似为

$$k(\boldsymbol{u},\boldsymbol{u}') \simeq B^{-1}\sum_{b=1}^{B}\cos(\boldsymbol{\omega}_b^\top(\boldsymbol{u}-\boldsymbol{u}')) = \phi(\boldsymbol{u})^\top\phi(\boldsymbol{u}'), \qquad \{\boldsymbol{\omega}_b\}_{b=1}^{B} \overset{i.i.d.}{\sim} s(\boldsymbol{\omega}), \tag{11}$$

其中，$\phi(\boldsymbol{u}) = B^{-\frac{1}{2}}[\cos(\boldsymbol{\omega}_1^\top\boldsymbol{u})\ldots\ldots, \cos(\boldsymbol{\omega}_B^\top\boldsymbol{u}), \sin(\boldsymbol{\omega}_1^\top\boldsymbol{u})\ldots\ldots, \sin(\boldsymbol{\omega}_B^\top\boldsymbol{u})]^\top$。最后一个等号依据三角恒等式：$\cos(u-v) = \cos u\cos v + \sin u\sin v$。将上述随机傅里叶特征代入公式（9），得到

$$f_c(\boldsymbol{u}^{\mathsf{s}}) \simeq \Big(\theta_c^{\mathsf{s}} + \sum_{\mathsf{v}\in\mathcal{S}\setminus\{\mathsf{s}\}}\lambda^{\mathsf{s},\mathsf{v}}\theta_c^{\mathsf{v}}\Big)^\top\phi(\boldsymbol{u}^{\mathsf{s}}), \tag{12}$$

其中 $\theta_c^{\mathsf{s}} = \sum_{i=1}^{n_{\mathsf{s}}}\phi(\boldsymbol{u}^{\mathsf{s}})\boldsymbol{\alpha}_i^{\mathsf{s}}$ 和 $\lambda^{\mathsf{s},\mathsf{v}}(\mathsf{s},\mathsf{v}\in\mathcal{S})$。在优化目标函数 J时，我们可以不学习 $\boldsymbol{\alpha}_i^{\mathsf{s}}$，而直接将 θ^{s} 视为待优化参数。这种做法已应用于多项研究，例如 Rahimi 和 Recht（2007）；Chaudhuri 等（2011）；Rajkumar 和 Agarwal（2012）。这一近似使我们能够将目标函数 J改写为各数据源本地目标函数的求和：

$$J \simeq \sum_{\mathsf{s}\in\mathcal{S}} J^{(\mathsf{s})}, \qquad \text{where } J^{(\mathsf{s})} = \widehat{\mathcal{L}}^{(\mathsf{s})} + m^{-1}\sum_{\mathsf{v}\in\mathcal{S}}\zeta\|\theta^{\mathsf{v}}\|_2^2, \tag{13}$$

其中 $\zeta\in\mathbb{R}^+$ 是正则化因子。每个分量 $J^{(\mathsf{s})}$ 都与数据源 s 相关联，并且可以利用该数据源中的本地数据计算得出。因此，这使得目标函数 J能够进行联邦优化。图 2 展示了我们提出的包含三个数据源的联邦因果学习算法，其中 θ 表示所有数据源中待学习的全部参数集合，包括 θ^{s} 和 $\lambda^{\mathsf{s},\mathsf{v}}$。联邦学习算法可概括如下：首先，每个数据源使用自己的数据计算局部梯度 $\nabla_\theta J^{(\mathsf{s})}$，并将其发送给服务器。随后，服务器收集所有数据源的这些梯度，并据此更新模型。接着，服务器将新模型广播给所有数据源。

极小极大下界。现在我们计算所提出模型的极小极大下界，该下界给出了随着样本量增加，估计量收敛到所关注总体量的速率。我们首先陈述以下关于式 (3) 中最后两项的结果：

引理 1 (存在潜变量时)。令 $\theta = \{\theta_c^{\mathsf{s}}: c\in\{y0,y1,x,w\}, \mathsf{s}\in\mathcal{S}\}$ 及其估计量为 $\hat{\theta}$。令 $y_i^{\mathsf{s}}\in\mathbb{R}$ 和 $\boldsymbol{x}_i^{\mathsf{s}}\in\mathbb{R}^{d_x}$。令 $\mathcal{S}_{\setminus\mathsf{s}} = \mathcal{S}\setminus\{\mathsf{s}\}$。那么，

$$\inf_{\hat{\theta}}\sup_{P\in\mathcal{P}}\mathbb{E}_P\Big[\|\hat{\theta}-\theta(P)\|_2\Big] \geq \frac{\sqrt{m(d_x+3)}\log(2\sqrt{m})}{64\sqrt{B}\sum_{\mathsf{s}\in\mathcal{S}}n_{\mathsf{s}}\big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\big)^2}. \tag{14}$$

式 (14) 的左侧可视为最优估计量的最坏情形，而右侧则描述了收敛行为。这些界不仅取决于每个数据源的样本数 (n_{s}，训练样本量)，还取决于自适应因子 $\lambda^{\mathsf{s},\mathsf{v}}$。自适应因子较小时，下界较大，因为来自数据源 s 的数据仅用于学习其自身的参数 θ^{s}。自适应因子较大时，下界较小，这表明来自某个数据源的数据有助于推断与其他数据源相关的参数。该界保证了来自所有数据源的数据如何影响调节这两个分布的学习参数$p_{\mathsf{s}}(\boldsymbol{z}^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}}, y^{\mathsf{s}}, w^{\mathsf{s}})$和 $p_{\mathsf{s}}(y^{\mathsf{s}}|w^{\mathsf{s}}, \boldsymbol{z}^{\mathsf{s}})$。引理 1 的证明见附录。

4.2 学习辅助分布

上一节展示了如何学习 $p_{\mathsf{s}}(\boldsymbol{z}^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}}, y^{\mathsf{s}}, w^{\mathsf{s}})$ 和 $p_{\mathsf{s}}(y^{\mathsf{s}}|w^{\mathsf{s}}, \boldsymbol{z}^{\mathsf{s}})$。为了计算处理效应，我们还需要学习另外两个条件分布，即 $p_{\mathsf{s}}(w^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}})$ 和 $p_{\mathsf{s}}(y^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}}, w^{\mathsf{s}})$。由于这两个分布中的所有变量都是可观测的，我们使用最大似然估计来估计它们。下面，我们介绍一种用于学习 $p_{\mathsf{s}}(w^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}})$ 的联邦设置。与上一节类似，这里的目标函数也可以分解为 m 个分量，如下所示：$J_w \simeq \sum_{\mathsf{s}\in\mathcal{S}} J_w^{(\mathsf{s})}$，其中 $J_w^{(\mathsf{s})} = \sum_{i=1}^{n_{\mathsf{s}}}\ell(w_i^{\mathsf{s}}, \varphi(g(\boldsymbol{x}_i^{\mathsf{s}}))) + m^{-1}\sum_{\mathsf{v}\in\mathcal{S}}\zeta_w\|\psi^{\mathsf{v}}\|_2^2$ 和$g(\boldsymbol{x}_i^{\mathsf{s}}) = \sum_{\mathsf{v}\in\mathcal{S}}\phi(\boldsymbol{x}_i^{\mathsf{s}})^\top(\psi^{\mathsf{s}} + \gamma^{\mathsf{s},\mathsf{v}}\psi^{\mathsf{v}})$，$\gamma^{\mathsf{s},\mathsf{v}}\in[0,1]$ 是自适应因子，ψ^{s} 是与数据源 s 关联的参数，而 $\ell(\cdot)$ 表示交叉熵损失函数，因为 w_i^{s} 是二元值。$J_w^{(\mathsf{s})}$ 的第一项由负对数似然得到。学习 $p_{\mathsf{s}}(y^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}}, w^{\mathsf{s}})$ 的过程

类似。为方便起见，在接下来的分析中，我们将该分布的参数和自适应因子分别记为 β^{s} 和 $\eta^{\mathsf{s},\mathsf{v}}$，其中 $\mathsf{s},\ \mathsf{v} \in \mathcal{S}$ 和 $\mathsf{s} \neq \mathsf{v}$。下一个引理给出了式（3）中前两组参数 $\boldsymbol{\psi}$ 和 $\boldsymbol{\beta}$ 的极小极大下界，但这次不涉及潜变量：

引理 2（不存在潜变量时）。设 $\boldsymbol{\psi} = \{\psi^{\mathsf{s}}\}_{\mathsf{s}=1}^{m}$、$\boldsymbol{\beta} = \{\beta^{\mathsf{s}}\}_{\mathsf{s}=1}^{m}$ 和 $\hat{\boldsymbol{\psi}}$、$\hat{\boldsymbol{\beta}}$ 分别为它们的估计值。设 $y_i^{\mathsf{s}} \in \mathbb{R}$。那么，

$$\textbf{(i)}\ \inf_{\hat{\boldsymbol{\psi}}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}(P)\|_2 \right] \geq \frac{m \log(2\sqrt{m})}{256 \sum_{\mathsf{s} \in \mathcal{S}} n_{\mathsf{s}} \left(1 + \sum_{\mathsf{v} \in \mathcal{S}_{\setminus \mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\right)}, \tag{15}$$

$$\textbf{(ii)}\ \inf_{\hat{\boldsymbol{\beta}}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}(P)\|_2 \right] \geq \frac{\sigma}{2^{\frac{9}{2}}} \left(\frac{m \log(2\sqrt{m})}{B \sum_{\mathsf{s} \in \mathcal{S}} n_{\mathsf{s}} \left(1 + \sum_{\mathsf{v} \in \mathcal{S}_{\setminus \mathsf{s}}} \eta^{\mathsf{s},\mathsf{v}}\right)^2} \right)^{1/2}. \tag{16}$$

引理 2 的证明见附录。引理 1 和引理 2 中给出的界有助于了解所需观测的样本数量，以及多个数据源如何通过迁移因子进行协作。由于我们使用变分推断和最大似然来学习模型中的参数，这些方法能够给出一致估计，参见 Kiefer 和 Wolfowitz（1956）；Van der Vaart（2000）；Wang 和 Blei（2019）；Yang 等人（2020）。

4.3 计算因果效应

在我们的模型中，估计因果效应的关键是计算式 (2) 中的结果。我们通过从式 (3) 中的分布抽取样本来进行。由于式 (4.1) 或第 4.2 节中已给出了这些条件分布，从中抽取样本很简单：$p_{\mathsf{s}}(w^{\mathsf{s}} | \boldsymbol{x}^{\mathsf{s}})$、$p_{\mathsf{s}}(y^{\mathsf{s}} | \boldsymbol{x}^{\mathsf{s}}, w^{\mathsf{s}})$ 和 $p_{\mathsf{s}}(y^{\mathsf{s}} | w^{\mathsf{s}}, \boldsymbol{z}^{\mathsf{s}})$。从混杂因素 $p_{\mathsf{s}}(\boldsymbol{z}^{\mathsf{s}} | \boldsymbol{x}^{\mathsf{s}}, y^{\mathsf{s}}, w^{\mathsf{s}})$ 的后验分布中抽取样本有两种方法。第一种方法是从其近似分布 $q(\boldsymbol{z}^{\mathsf{s}})$ 中抽样，因为最大化第 4.1 节中的证据下界等价于最小化 $\mathrm{KL}(q(\boldsymbol{z}^{\mathsf{s}}) \| p_{\mathsf{s}}(\boldsymbol{z}^{\mathsf{s}} | \boldsymbol{x}^{\mathsf{s}}, y^{\mathsf{s}}, w^{\mathsf{s}}))$。第二种方法是注意到，混杂因素的精确后验可以改写为 $p_{\mathsf{s}}(\boldsymbol{z}^{\mathsf{s}} | \boldsymbol{x}^{\mathsf{s}}, y^{\mathsf{s}}, w^{\mathsf{s}}) \propto p_{\mathsf{s}}(y^{\mathsf{s}} | \boldsymbol{z}^{\mathsf{s}}, w^{\mathsf{s}}) p_{\mathsf{s}}(w^{\mathsf{s}} | \boldsymbol{z}^{\mathsf{s}}) p_{\mathsf{s}}(\boldsymbol{x}^{\mathsf{s}} | \boldsymbol{z}^{\mathsf{s}}) p(\boldsymbol{z}^{\mathsf{s}})$，其右侧的分量在第 4.1 节中也都可得。因此，我们可以使用 Metropolis-Hastings（MH）算法从该分布中抽样。由于 Z 是一个多维随机变量，传统的 MH 算法需要很长的链才能收敛。我们通过使用独立提议的 MH 算法（Liu 1996）来解决这一问题，其中提议分布是第 4.1 节中学得的变分后验分布 $q(\boldsymbol{z}^{\mathsf{s}})$。第二种方法能得到更准确的样本，因为我们根据后验分布 $p_{\mathsf{s}}(\boldsymbol{z}^{\mathsf{s}} | \boldsymbol{x}^{\mathsf{s}}, y^{\mathsf{s}}, w^{\mathsf{s}})$ 的精确接受概率来筛选样本。这有助于估计给定 $\boldsymbol{x}_i^{\mathsf{s}}$ 时的条件平均处理效应。局部平均处理效应是数据源 s 中各个体的条件平均处理效应的平均值。这些量可以在本地源端机器上估计。若要计算全局平均处理效应，服务器会收集各数据源的所有局部平均处理效应，然后计算它们的加权平均。更多细节见附录。

5 实验

基线方法。在本节中，我们首先开展实验，考察 CausalRFF 相较于标准基线方法的性能，这些方法包括 BART（Hill 2011）、TARNet（Shalit 等人，2017）、CFR-wass（采用 Wasserstein 距离的 CFRNet）（Shalit 等人，2017）、CFR-mmd（采用最大均值差异距离的 CFRNet）（Shalit 等人，2017）、CEVAE（Louizos 等人，2017）、OrthoRF（Oprescu 等人，2019）、X-learner（Künzel 等人，2019）、R-learner（Nie 和 Wager，2020）以及 FedCI（Vo 等人，2022）。与 CausalRFF 不同，这些方法（FedCI 除外）未考虑在联邦设置下进行因果推断。我们将本方法与通过两种方式训练的这些基线方法进行比较：**(a)** 使用所有数据源的合并数据训练全局模型；**(b)** 使用 Breiman（1996）的自助聚合，其中 m 模型分别在各数据源的数据上训练，然后基于每个训练好的模型对处理效应的预测取平均。请注意，情况 **(a)** 违反了联邦数据设置，仅用于比较。一般而言，当所有数据源的数据分布相同时，我们预期 CausalRFF 的性能接近情况 **(a)** 下基线方法的性能。此外，我们还展示了当所有数据源的数据分布不同时，CausalRFF 的性能优于情况 **(a)** 下的基线方法。

基线方法的实现。CEVAE 的实现来自 Louizos 等（2017）。TARNet、CFR-wass 和 CFR-mmd 的实现来自 Shalit 等人（2017）。对于这些

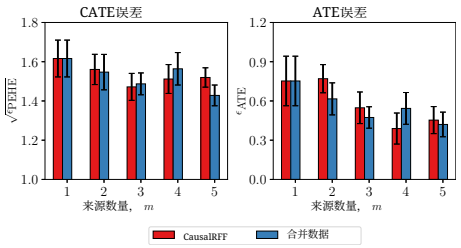


图3：DATA$^{\text{same}}$上的实验结果。

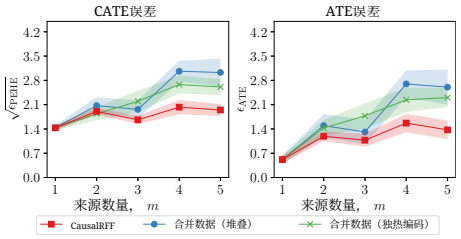


图4：DATA$^{\text{diff}}$上的实验结果。

表1：DATA$^{\text{same}}$上的样本外误差，其中性能排名前3的方法以粗体突出显示（越低越好）。'ag'（自助法聚合）中的短横线（-）表示其数值与'cb'（合并数据）相同。

方法	CATE误差，$\sqrt{\epsilon_{\text{PEHE}}}$		平均处理效应误差，ϵ_{ATE}	
	1 个数据源	3 个数据源 5 个数据源	1 个数据源	3 个数据源 5 个数据源
BART_{ag}	-	3.8±.10 3.8±.09	-	2.3±.15 2.3±.14
$\text{X-Learner}_{\text{ag}}$	-	3.2±.07 3.1±.06	-	0.6±.11 0.5±.13
$\text{R-Learner}_{\text{ag}}$	-	3.5±.17 3.9±.46	-	1.5±.35 2.0±.70
$\text{OthoRF}_{\text{ag}}$	-	5.4±.21 4.5±.12	-	0.5±.10 0.7±.16
$\text{TARNet}_{\text{ag}}$	-	3.9±.04 3.4±.03	-	2.2±.07 2.0±.02
$\text{CFR-wass}_{\text{ag}}$	-	3.0±.05 3.6±.02	-	2.1±.03 1.8±.02
$\text{CFR-mmd}_{\text{ag}}$	-	4.0±.03 3.9±.02	-	2.3±.03 2.0±.01
CEVAE_{ag}	-	**2.9**±.04	-	0.7±.08
BART_{cb}	**3.7**±.12	3.2±.07 3.1±.03	2.1±.20	1.0±.18 0.6±.13
$\text{X-Learner}_{\text{cb}}$	3.3±.06	3.4±.06 3.3±.04	**0.5**±.11	**0.4**±.0
07 3.4±.04	2.2±.72	0.6±.15 0.9±.15	$\text{OthoRF}_{\text{cb}}$	7.6±.29 4.3±.10 3.7±.07 1.4±.30 **0.4**±.
12 0.5±.10	$\text{TARNet}_{\text{cb}}$	4.2±.07 3.8±.03	3.5±.02	2.2±.13 2.1±.06
$\text{CFR-wass}_{\text{cb}}$	4.0±.11	3.8±.02 3.7±.02	2.1±.06	2.0±.03 1.9±.02 $\text{CFR-mmd}_{\text{cb}}$ 3.8±.05 3.8±.02 3.7±.
02 2.1±.04	2.1±.03	2.0±.02		
CEVAE_{cb}	**2.5**±.03			
FedCI	**2.5**±.03			
CausalRFF	**1**			

方法，我们使用指数线性单元（ELU）激活函数，并将每个隐藏层的节点数从 10 到 200 进行微调，步长为 10。对于 BART，我们使用现成的 BartPy 软件包。对于 X-Learner 和 R-Learner，我们使用 causalml 软件包（ Chen 等人，2020 ）。对于 OrthoRF，我们使用 econml 软件包（微软研究院，2019）。对于 FedCI，我们使用 Vo 等人（2022）的代码。对于所有方法，学习率都从 10^{-4} 到10^{-1} 进行微调，按乘以 10 的步长递增。同样，正则化因子也从 10^{-4} 到 10^{0} 进行微调，按乘以 10 的步长递增。我们报告两种误差指标： ϵ_{PEHE} （异质性效应估计精度）和 ϵ_{ATE} （绝对误差），用于比较各方法。我们报告 10 个数据重复样本的均值和标准误。更多细节见附录。

5.1 合成数据

数据描述。 获取用于评估因果推断算法的真实值是一项具有挑战性的任务。因此，大多数最先进的方法都使用合成数据集或半合成数据集进行评估。在本实验中，合成数据按照以下分布进行模拟：

$$\boldsymbol{z}_i^{\mathsf{s}} \sim \mathsf{Cat}(\rho), \qquad x_{ij}^{\mathsf{s}} \sim \mathsf{Bern}(\varphi(a_{j0} + (\boldsymbol{z}_i^{\mathsf{s}})^\top \mathbf{a}_{j1})), \qquad w_i^{\mathsf{s}} \sim \mathsf{Bern}(\varphi(b_0 + (\boldsymbol{z}_i^{\mathsf{s}})^\top (\mathbf{b}_1 + \Delta))),$$
$$y_i^{\mathsf{s}}(0) \sim \mathsf{N}(\mathsf{sp}(c_0 + (\boldsymbol{z}_i^{\mathsf{s}})^\top (\mathbf{c}_1 + \Delta)), \sigma_0^2), \qquad y_i^{\mathsf{s}}(1) \sim \mathsf{N}(\mathsf{sp}(d_0 + (\boldsymbol{z}_i^{\mathsf{s}})^\top (\mathbf{d}_1 + \Delta)), \sigma_1^2),$$

其中，Cat(·)、N(·) 和 Bern(·) 分别表示分类分布、正态分布和伯努利分布。 $\varphi(\cdot)$ 表示 Sigmoid函数，sp(·) 表示 SoftPlus函数，而 $\boldsymbol{x}_i = [x_{i1}, ..., x_{id_x}]^\top \in \mathbb{R}^{d_x}$ with $d_x = 30$。在此，我们将 $\boldsymbol{z}_i^{\mathsf{s}}$ 转换为独热向量。为模拟数据，我们随机设定真实值参数如下： $\rho = [.11, .17, .34, .26, .12]^\top$， (c_0, d_0) =(0.9,7.9)， $(\mathbf{c}_1, \mathbf{d}_1)$ 独立同分布地取自 N(0, 2I_5)， a_{j0} ，且 $\mathbf{a}_{j1}$ 的元素独立同分布地取自 N(0,2)。对于每个数据源，我们模拟 10 次重复实验，每次包含 $n_{\mathsf{s}} = 1000$ 条记录。我们仅保留 $\{(y_i^{\mathsf{s}}, w_i^{\mathsf{s}}, \boldsymbol{x}_i^{\mathsf{s}})\}_{i=1}^{n_{\mathsf{s}}}$ 作为观测数据，其中 $y_i^{\mathsf{s}} = y_i^{\mathsf{s}}(0)$（当 $w_i^{\mathsf{s}} = 0$ 时）以及 $y_i^{\mathsf{s}} = y_i^{\mathsf{s}}(1)$（当 $w_i^{\mathsf{s}} = 1$ 时）。每个数据源中，我们使用 50 个数据点进行训练、450 个进行测试，400 个进行验证。我们报告 10 次重复实验的评估指标及其标准误。

结果与讨论（I）。在第一个实验中，我们研究了 CausalRFF 在多个数据源上的性能，这些数据源的数据分布相同。为此，我们模拟了来自同一分布的 $m = 5$ 个数据源，即为所有数据源设置相同的真实值 $\Delta = 0.0$ 。我们将此数据集称为 DATA$^{\text{same}}$。在本实验中，我们预期，在联邦设置下训练的 CausalRFF，其结果与在合并数据上训练的效果一样好。图 3 中的结果表明，两种情况下的误差似乎以相关的方式同步变化，这验证了我们的假设。

此外，为了研究 CausalRFF 在数据分布不同的数据源上的性能，我们还模拟了 $m = 5$ 个数据源。然而，第一个数据源具有 $\Delta = 0.0$ ，而其他

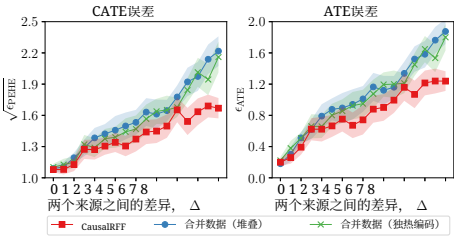


图5：不同差异水平下的实验结果， Δ。

表2：DATA$^{\text{diff}}$上的样本外误差。

方法	CATE误差， $\sqrt{\epsilon_{\text{ATE}}}$			ATE误差， ϵ_{ATE}		
	1 个数据源	3 个数据源	5 个数据源	1 个数据源	3 个数据源	5 个数据源
BART$_{\text{ag}}$	-	**3.0±.01**		-	1.3±.05	
X-Learner $_{\text{ag}}$	-	3.3±.03	3.3±.04	-	**1.2±.09**	
R-Learner $_{\text{ag}}$	-	3.2±.03	3.1±.02	-	**1.0±.07**	
OthoRF $_{\text{ag}}$	-	3.6±.05	3.6±.05	-	1.3±.09	1.6±.10
TARNet$_{\text{ag}}$	-	6.1±.19	5.7±.05	-	2.5±.06	3.0±.05
CFR-wass $_{\text{ag}}$	-	5.6±.09	5.7±.07	-	2.7±.05	2.8±.04
CFR-mmd $_{\text{ag}}$	-	5.9±.08	5.6±.05	-	2.5±.03	2.8±.02
CEVAE $_{\text{ag}}$	-	4.2±.07	3.9±.05	-	2.1±.09	1.8±.10
BART$_{\text{cb}}$	3.1±.05	4.1±.10	4.2±.10	0.8±.17	2.8±.15	2.9±.14
X-Learner$_{\text{cb}}$	3.3±.03	5.0±.08	4.6±.10	**0.5**±.12	3.3	13
R-Learner$_{\text{cb}}$	3.3±.05	3.5±.05	3.3±.05	0.7±.18	**1.1**±.10	**1.3**±.10
OthoRF$_{\text{cb}}$	3.9	**0.5**±.11	3.3±.14	3.0±.12		
TARNet$_{\text{cb}}$	4.2±.07	5.9±	6	2.2±.13	2.3±.04	2.9±.02
CFR-wass$_{\text{cb}}$	4.0±.11	5.7±.08	5.5±.08	1.9±.06	2.4±.03	2.9±.04
CFR-mmd$_{\text{cb}}$	3.8±.05	5.7±.08	5.5±.04	2.1±.04	2.4±.03	2.9±.04
CEVAE$_{\text{cb}}$	**2.4±.03**	5.0±.06	4.4±.0	:.6±.10	2.0±.07	
FedCI	**2.5±.03**			.5±.13		
CausalRFF	**1**					

四个数据源具有 $\Delta = 4.0$。我们将此数据集称为 DATA$^{\text{diff}}$。我们测试了第一个数据源上的 CATE 误差和平均处理效应。此时，我们预期 CausalRFF 的误差会低于在合并数据上训练所得的误差，因为 CausalRFF 能够学习自适应因子，从而避免其他四个数据源对第一个数据源产生负面影响。图 4 中的结果表明，与在合并数据上训练相比，CausalRFF 的误差更低（合并数据有两种方式：堆叠数据，以及添加独热向量来标识每个数据点所属的数据源），这与预期相符。在第三个实验中，我们研究了 Δ 对 CausalRFF 性能的影响。我们模拟了具有不同 Δ取值的$m = 2$ 个数据源。具体而言，第一个数据源的 $\Delta = 0.0$ 为固定值，第二个数据源的 Δ 在 0.0 到 8.0 之间变化。我们将 CausalRFF 方法与基于合并数据进行训练的方法进行了比较。同样，图 5 表明，正如预期，CausalRFF 实现了更低的误差。

结果与讨论（II）。本节旨在比较 CausalRFF 与其在两个数据集 DATA$^{\text{same}}$ 和 DATA$^{\text{diff}}$ 上的基线方法。除 FedCI（贝叶斯联邦方法）外，其他基线方法均采用前文所述的两种情形进行训练：合并数据（cb）和自助聚合（ag）。在 DATA$^{\text{same}}$ 上，我们预期所提出方法的性能与使用合并数据训练的基线方法一样好。表1中的结果表明，CausalRFF 的性能符合预期。对于 DATA$^{\text{diff}}$，我们在表2中报告了结果。数据显示，CausalRFF 在预测平均处理效应方面的性能与基线方法一样好。在预测条件平均处理效应方面，随着加入更多分布不同于第一个数据源的数据源，基线方法的性能显著下降。与此同时，CausalRFF 在预测条件平均处理效应方面的性能略有下降，但仍远优于基线方法。这是因为我们使用了自适应因子来学习各数据源之间数据分布的相似性。

5.2 大规模合成数据

数据描述。在本节中，我们针对大量数据源进行了实验。本节的设置与第5.1节类似。我们模拟了两种情况：(1) DATA-LARGE$^{\text{same}}$：包含100个数据源的数据集，其中我们为所有数据源设置 $\Delta = 0$，使其分布相同。(2) DATA-LARGE$^{\text{diff}}$：包含100个数据源的数据集，其中我们为每个数据源均匀抽取差异因子$\Delta \sim \text{U}[0, 8]$，使其分布各不相同。在这两种情况下，我们都使用前20个数据源中的测试集进行评估。

结果与讨论。表3显示，当数据源具有相同分布时，CausalRFF在估计平均处理效应和条件平均处理效应方面取得了有竞争力的结果。表4显示，当数据源的分布不同时，CausalRFF优于基线方法。这些结果与我们在第5.1节中的讨论一致。

5.3 一个真实世界数据集

数据描述。婴儿健康与发展计划（IHDP）（Hill 2011）是一项随机研究，旨在考察专科医生就诊（处理）对儿童认知发展的影响（

表3：DATA-LARGE^{same} 数据集。

方法	CATE误差，$\sqrt{\epsilon_{ATE}}$			ATE误差，ϵ_{ATE}		
	20	50	100	20	50	100
	数据源	数据源	数据源	数据源	数据源	数据源
$BART_{cb}$	3.4±.03	3.4±.01	3.3±.01	1.4±.06	1.3±.02	1.3±.01
$X\text{-}Learner_{cb}$	3.0±.01	2.9±.01	2.9±.01	**.16**±.02	**.12**±.02	**.13**±.0
$OthoRF_{cb}$	2.9±.01	**.07**±.01	**.10**±.02	**.10**±.02		
$TARNet_{cb}$	3.8±.03	3.7±.01	3.3±.01	1.1±.02	1.0±.01	.93±.01
$CFR\text{-}wass_{cb}$	3.7±.02	3.6±.01	3.2±.01	1.1±.02	.99±.01	.87±.01
$CFR\text{-}mmd_{cb}$	3.7±.02	3.6±.01	3.2±.01	1.1±.02	.98±.01	.87±.01
$CEVAE_{cb}$	**2.3**±.01	**2.2**±.01	**2.0**±.01	**.19**		
FedCI	**2.2±.02**	**2**		。04	。21±	。01
CausalRFF	**1.6±.05**			.3±.04	0.2±.02	**.1**

表4：DATA-LARGE^{diff} 数据集。

方法	CATE误差，$\sqrt{\epsilon_{PEHE}}$			ATE误差，ϵ_{ATE}		
	20	50	100	20	50	100
	数据源	数据源	数据源	数据源	数据源	数据源
$BART_{cb}$	3.4±.03	3.5±.01	3.5±.01	1.4±.06	1.5±.02	1.5±.01
$X\text{-}Learner_{cb}$	3.3±.04	**3.2**±.01	3.2±.01	1.	.02	**1.2**±.0
	3.1±.01	**.88**±.0				
$TARNet_{cb}$	5.6±.04	5.6±.02	5.7±.02	2.7		.02
$CFR\text{-}wass_{cb}$	5.4±.05	5.5±.02	5.5±.02	2.7±.05	2.7±.02	2.7±.02
$CFR\text{-}mmd_{cb}$	5.4±.05	5.4±.02	5.5±.02	2.7±.05	2.7±.02	2.7±.02
$CEVAE_{cb}$	3.4±.04	3.4±.02	3.3±.01	1.2±.06	**1.2**±.02	**1.2**±.01
FedCI	**3.2±.03**			.2±.07	**1.2**	
CausalRFF	**1.8±.03**					

结果）。该数据集包含747条记录，其中有25个协变量，用于描述儿童及其母亲的特征。处理组包括接受专科医生就诊的儿童，对照组则包括未接受专科医生就诊的儿童。该数据集通过从处理组样本中剔除母亲为非白人的儿童，进行了“去随机化”处理。随后，为每名儿童模拟处理结果和对照结果，从而使我们能够获知处理的“真实”个体因果效应。更多详情请见附录。

结果与讨论。 表5报告了IHDP 数据集上的实验结果。同样，我们看到，与基线方法相比，所提出的方法取得了具有竞争力的结果。特别是，CausalRFF 在预测平均处理效应（ATE）时的误差与基线方法一样低，这符合我们的预期。此外，CausalRFF 在预测条件平均处理效应（ CATE）时的误差低于基线方法，验证了所提出方法的有效性。最重要的是，CausalRFF 在联邦设置下进行训练，从而最大限度地降低了存储在本地数据集中的个体隐私泄露风险。

表5：IHDP 数据集上的样本外误差。

方法	CATE误差，$\sqrt{\epsilon_{PEHE}}$			ATE误差，ϵ_{ATE}		
	1 个数据源	2 个数据源	3 个数据源	1 个数据源	2 个数据源	3 个数据源
$BART_{ag}$	-	2.3±.26	2.4±.22	-	1.2±.23	1.3±.18
$X\text{-}Learner_{ag}$	-	**1.8**±.20	1.8±.22	-	**0.6±.15**	(
$R\text{-}Learner_{ag}$	-	2.4±.31	2.3±.21	-	1.3±.34	1.2±.24
$OthoRF_{ag}$	-	2.3±.21	2.1±.16	-	0.6±.22	0.7±.13
$TARNet_{ag}$	-	2.9±.13	2.7±.15	-	**0.7**±.12	0.7±.16
$CFR\text{-}wass_{ag}$	-	2.3±.31	2.2±.20	-	**0.7**±.12	0.7±.11
$CFR\text{-}mmd_{ag}$	-	2.6±.21	2.4±.15	-	0.8±.19	0.7±.18
$CEVAE_{ag}$	-	1.9±.14		-	1.2±.11	0.8±.10
$BART_{cb}$	2.2±.22	2.1±.26	2.1±.25	1.0±.16	0.8±.20	0.7±.17
$X\text{-}Learner_{cb}$	1.9±.21	1.9±.21	1.8±.18	**0.5±.21**	**0.5±.18**	**0.4±.11**
$R\text{-}Learner_{cb}$	2.8±.31	2.6±.23	2.6±.17	1.6±.25	1.6±.26	1.6±.19
$OthoRF_{cb}$	2.8±.16	2.1±.14	1.9±.14	**0.8±.15**	**0.6±.10**	**0.6±.10**
$TARNet_{cb}$	3.5±.59	2.7±.12	2.5±.15			
$CFR\text{-}wass_{cb}$	2.2±.15	2.1±.22	2.1±.23	**0.7±.23**	**0.6±.18**	
	26	2.2±.10	0.9±.30	**0.7±.17**	**0.5±.17**	
$CEVAE_{cb}$	**1.8±.22**	2.0±.11				
	07	0.9±.07				
FedCI	**1.6±.10**					
CausalRFF	**1**					

6 结论

我们提出了一种新方法，用于从具有不同分布的联邦观察性数据源中学习因果效应。我们的方法利用随机傅里叶特征，自然地将损失函数分解为各个分量。该方法允许每个分量数据组具有不同的分布，并且不需要预先了解数据源之间的数据差异。我们还证明了统计保证，展示了如何将多个数据源有效地纳入我们的因果模型。我们的工作是迈向隐私保护因果推断的重要一步。未来的工作可以包括将所提出的方法与多方差分隐私技术相结合（例如，Pathak等人，2010年；Rajkumar和Agarwal，2012年；Pettai和Laud，2015年；Hamm等人，2016年），这可能会带来更强的隐私保证模型。另一个方向是将所提出的方法与一些近期思想相结合（例如，Khemakhem等人，2020年；Sun等人，2021年），以研究模型的可识别性。

致谢与经费披露

本研究/项目由新加坡国家研究基金会和DSO国家实验室通过新加坡人工智能计划资助（AISG奖项编号：AISG2-RP-2020-016）。
AB获得了NRF人工智能研究员资助（NRFFAI1-2019-0002）和亚马逊研究奖的支持。

本工作是在YL就职于哈佛大学期间开展的，本文所表达的观点未必代表Roche AG的立场。

参考文献

Aglietti, V., Damoulas, T., Álvarez, M.和González, J.（2020）。使用高斯过程进行多任务因果学习。载于神经信息处理系统进展，第6293–6304页。Alaa, A.和van der Schaar, M.（2018）。估计异质性治疗效应的局限：实用算法设计指南。载于第*35*届国际机器学习大会论文集，第129–138页。PMLR。Alaa, A. M.和van der Schaar, M.（2017）。使用多任务高斯过程对个体化治疗效应进行贝叶斯推断。载于神经信息处理系统进展，第3424–3432页。Allman, E. S., Matias, C.和Rhodes, J. A.（2009）。具有多个观测变量的潜在结构模型中参数的可识别性。《统计年鉴》，37(6A):3099–3132。Álvarez, M. A., Ward, W.和Guarnizo, C.（2019）。多输出高斯过程的非线性过程卷积。载于第*22*届人工智能与统计国际会议，第1969–1977页。PMLR。Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M.和Telgarsky, M.（2014）。用于学习潜变量模型的张量分解。《机器学习研究杂志》，15:2773–2832。Anandkumar, A., Hsu, D.和Kakade, S. M.（2012）。混合模型和隐马尔可夫模型的矩估计法。载于第*25*届学习理论年度会议论文集，第33–1页。PMLR。Arora, S., Ge, R., Ma, T.和Risteski, A.（2017）。噪声或网络的可证明学习。载于第*49*届*ACM SIGACT*计算理论研讨会论文集，第1057–1066页。Bareinboim, E.和Pearl, J.（2016）。因果推断与数据融合问题。《美国国家科学院院刊》，113(27):7345–7352。Breiman, L.（1996）。装袋预测器。《机器学习》，24(2):123–140。Chaudhuri, K., Monteleoni, C.和Sarwate, A. D.（2011）。差分隐私经验风险最小化。《机器学习研究杂志》，12(3)。Chen, H., Harinen, T., Lee, J.-Y., Yung, M.和Zhao, Z.（2020）。CausalML：用于因果机器学习的Python软件包。Choudhury, O., Park, Y., Salonidis, T., Gkoulalas-Divanis, A., Sylla, I.等（2019）。使用联邦学习在分布式健康数据上预测药物不良反应。载于*AMIA*年度研讨会论文集，第2019卷，第313页。美国医学信息学会。de Wolff, T., Cuevas, A.和Tobar, F.（2020）。Mogptk：多输出高斯过程工具包。*arXiv*预印本*arXiv:2002.03471*。Dorie, V.（2016）。Npci：因果推断的非参数方法。网址：*https://github. com/vdorie/npci*。Finkelstein, A.和Hendren, N.（2020）。福利分析与因果推断相结合。《经济学展望杂志》，34(4):146–67。Flores, M., Dayan, I., Roth, H., Zhong, A., Harouni, A., Gentili, A., Abidin, A., Liu, A., Costa, A., Wood, B.等（2020）。用于预测SARS-COV-2患者结果的联邦学习。预印本。*medRxiv*。*2020;2020.08.11.20172809*。Goodman, L. A.（1974）。使用可识别和不可识别模型进行探索性潜在结构分析。《生物计量学》，61(2):215–231。Gostin, L. O., Levit, L. A., Nass, S. J.等（2009）。超越*HIPAA*隐私规则：通过研究增强隐私、改善健康。美国国家科学院出版社。

Gutman, R., Intrator, O. 和 Lancaster, T.（2017）。一种用于估计养老院床位保留政策因果效应的贝叶斯程序。《生物统计学》，19(4):444–460。Hamm, J., Cao, Y. 和 Belkin, M.（2016）。从多方数据中进行隐私保护学习。载于第*33*届国际机器学习大会论文集，第555–563页。PMLR。Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C. 和 Ramage, D.（2018）。用于移动键盘预测的联邦学习。*arXiv*预印本 *arXiv:1811.03604*。Henderson, N. C., Louis, T. A., Wang, C. 和 Varadhan, R.（2016）。面向以患者为中心的结局研究的异质性治疗效应贝叶斯分析。《卫生服务与结局研究方法学》，16(4):213–233。Hill, J. L.（2011）。用于因果推断的贝叶斯非参数建模。《计算与图形统计学杂志》，20(1):217–240。Jernite, Y., Halpern, Y. 和 Sontag, D.（2013）。使用四元组检验发现噪声或网络中的隐变量。神经信息处理系统进展，26:2355–2363。Joukov, V. 和 Kuli´c, D.（2020）。快速近似多输出高斯过程。*arXiv*预印本 *arXiv:2008.09848*。Khemakhem, I., Kingma, D., Monti, R. 和 Hyvarinen, A.（2020）。变分自编码器与非线性独立成分分析：一个统一框架。载于第*23*届国际人工智能与统计学会议论文集，第2207–2217页。PMLR。Kiefer, J. 和 Wolfowitz, J.（1956）。存在无限多个附带参数时最大似然估计量的一致性。《数学统计年刊》，第887–906页。Kingma, D. P. 和 Welling, M.（2013）。自编码变分贝叶斯。载于第*2*届国际学习表征会议论文集。Künzel, S. R., Sekhon, J. S., Bickel, P. J. 和 Yu, B.（2019）。使用机器学习估计异质性治疗效应的元学习器。《美国国家科学院院刊》，116(10):4156–4165。Lee, S., Correa, J. 和 Bareinboim, E.（2020）。广义可迁移性：整合来自异质领域的实验。载于第*34*届*AAAI*人工智能大会论文集，纽约州纽约市。AAAI出版社。Liu, J. S.（1996）。Metropolis化独立抽样，并与拒绝抽样和重要性抽样进行比较。《统计与计算》，6(2):113–119。Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R. 和 Welling, M.（2017）。使用深度潜变量模型进行因果效应推断。载于神经信息处理系统进展，第6446–6456页。Madras, D., Creager, E., Pitassi, T. 和 Zemel, R.（2019）。通过因果认知实现公平性：为有偏数据学习因果潜变量模型。载于公平、问责与透明度大会论文集，第349–358页。美国计算机协会。McMahan, B., Moore, E., Ramage, D., Hampson, S. 和 y Arcas, B. A.（2017）。从去中心化数据中通信高效地学习深度网络。载于第*20*届国际人工智能与统计学会议论文集，第1273–1282页。PMLR。微软研究院（2019）。EconML：用于基于机器学习的异质性治疗效应估计的 Python 软件包。https://github.com/microsoft/EconML。版本0.x。Milton, P., Coupland, H., Giorgi, E. 和 Bhatt, S.（2019）。利用线性回归和核函数轻松进行空间分析。《流行病学》，29:100362。Mohri, M., Sivek, G. 和 Suresh, A. T.（2019）。无关先验的联邦学习。载于第*36*届国际机器学习大会论文集，第4615–4625页。PMLR。

Morgan, S. L. 和 Winship, C.（2015）。《反事实与因果推断》。剑桥大学出版社。Nie, X. 和 Wager, S.（2020）。异质性治疗效应的准预言估计。《生物计量学》。Oprescu, M., Syrgkanis, V. 和 Wu, Z. S.（2019）。用于因果推断的正交随机森林。载于第*36*届国际机器学习大会论文集，第4932–4941页。PMLR。Pathak, M., Rane, S. 和 Raj, B.（2010）。通过聚合本地训练的分类器实现多方差分隐私。神经信息处理系统进展，23。Pearl, J.（1995）。用于实证研究的因果图。《生物计量学》，82(4):669–688。Pearl, J.（2009a）。统计学中的因果推断：概述。《统计学调查》，3:96–146。Pearl, J.（2009b）。《因果性：模型、推理与推断》。剑桥大学出版社。Pearl, J. 和 Bareinboim, E.（2011）。因果关系与统计关系的可迁移性：一种形式化方法。载于第*25*届*AAAI*人工智能大会论文集。Pettai, M. 和 Laud, P.（2015）。结合差分隐私与安全多方计算。载于第*31*届年度计算机安全应用大会论文集，第421–430页。Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T. 和 Tibshirani, R.（2018）。高维情况下异质性处理效应估计的一些方法。《医学统计学》，37(11):1767–1787。Rahimi, A. 和 Recht, B.（2007）。用于大规模核机器的随机特征。神经信息处理系统进展，20。Rajkumar, A. 和 Agarwal, S.（2012）。用于多方分类的差分隐私随机梯度下降算法。载于第*15*届国际人工智能与统计学会议论文集，第933–941页。PMLR。Rosenbaum, P. R. 和 Rubin, D. B.（1983）。倾向评分在因果效应观察性研究中的核心作用。《生物计量学》，70(1):41–55。Sattler, F., Wiedemann, S., Müller, K.-R. 和 Samek, W.（2019）。基于非独立同分布数据的稳健且通信高效的联邦学习。*IEEE*《神经网络与学习系统汇刊》，31(9):3400–3413。Sattler, F., Wiedemann, S., Müller, K.-R. 和 Samek, W.（2020）。基于非独立同分布数据的稳健且通信高效的联邦学习。*IEEE*《神经网络与学习系统汇刊》，31(9):3400–3413。Shalit, U., Johansson, F. D. 和 Sontag, D.（2017）。估计个体处理效应：泛化界与算法。载于第*34*届国际机器学习大会论文集，第3076–3085页。JMLR.org。Shokri, R. 和 Shmatikov, V.（2015）。隐私保护深度学习。载于*ACM SIGSAC*计算机与通信安全大会，第1310–1321页。Sun, X., Wu, B., Zheng, X., Liu, C., Chen, W., Qin, T. 和 Liu, T.-Y.（2021）。恢复潜在因果因素，以实现对分布偏移的泛化。神经信息处理系统进展，34:16846–16859。Tsamardinos, I., Triantafillou, S. 和 Lagani, V.（2012）。迈向异质数据集与研究的整合性因果分析。《机器学习研究杂志》，13:1097–1157。Vaid, A., Jaladanki, S. K., Xu, J., Teng, S., Kumar, A. 和 Lee, S.（2020）。电子健康记录的联邦学习提升了患者死亡率预测效果。《族裔》，52(77.6):0–001。Van der Vaart, A. W.（2000）。《阿

*ymp*渐近统计学，第 3 卷。剑桥 g大学 y 出版社。

Vo, T. V., Lee, Y., Hoang, T. N., 和 Leong, T.-Y.（2022）。基于观察性数据的因果效应贝叶斯联邦估计。载于第*38*届人工智能不确定性会议论文集。Wang, H., Kaplan, Z., Niu, D., 和 Li, B.（2020）。利用强化学习优化非独立同分布数据上的联邦学习。载于*IEEE INFOCOM 2020 - IEEE*计算机通信会议，第1698–1707页。Wang, Y. 和 Blei, D. M.（2019）。变分贝叶斯的频率学派一致性。美国统计学会会刊，114(527):1147–1161。Wendland, H.（2004）。散乱数据逼近，第17卷。剑桥大学出版社。Xiong, R., Koenecke, A., Powell, M., Shen, Z., Vogelstein, J. T., 和 Athey, S.（2021）。异质观察性数据中的联邦因果推断。*arXiv*预印本 *arXiv:2107.11732*。Yang, Y., Pati, D., 和 Bhattacharya, A.（2020）。 α-具有统计保证的变分推断。《统计年鉴》，48(2):886–905。Yao, L., Li, S., Li, Y., Huai, M., Gao, J., 和 Zhang, A.（2018）。基于观察性数据的治疗效应估计表示学习。载于神经信息处理系统进展，第2633–2643页。Yoon, J., Jordon, J., 和 van der Schaar, M.（2018）。GANITE：利用生成对抗网络估计个体化治疗效应。载于第*6*届国际学习表征会议论文集。Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., 和 Chandra, V.（2018）。非独立同分布数据上的联邦学习。*arXiv*预印本 *arXiv:1806.00582*。Zhe, S., Xing, W., 和 Kirby, R. M.（2019）。可扩展的高阶高斯过程回归。载于第*22*届人工智能与统计国际会议，第2611–2620页。PMLR。

附录：用于异质因果效应联邦学习的自适应核方法

一种用于移除重复个体的预训练步骤

如正文所述，我们作出以下五项假设：

(A1) 一致性： $W = w \Longrightarrow Y(w) = Y$，这由结构因果模型的公理推得。**(A2)** 无干扰：对一个个体的处理不会影响另一个个体的结果。这是因为结果只有一个处理节点作为其父节点。**(A3)** 正值性（也称为重叠）：每个个体接受每种处理的概率都大于零。**(A4)** 每个数据源中的个体必须具有相同的一组共同协变量。**(A5)** 不存在数据分布在多个数据源中的个体。

假设**(A1)**、**(A2)** 和**(A3)** 是任何因果推断算法中的标准假设。

由于所有数据源都共享相同的因果图，因此假设**(A4)** 已在我们的设置中隐含成立。鉴于我们打算基于所有数据源构建统一模型，这是一个合理的假设。例如，Choudhury 等人（2019）、Vaid 等人（2020）和 Flores 等人（2020）中的去中心化数据（仅举几例）满足联邦学习的这一假设。

假设 **(A5)** 旨在确保训练模型时，没有个体会对其他个体产生主导影响。例如，如果某个个体出现在所有数据源中，训练得到的模型就会受到该个体数据的偏倚影响（由于使用了比其他个体更多的该个体数据，因此产生了不平衡）。因此，这一条件可以确保不存在这种偏倚。

在实践中，假设 **(A5)** 有时并不成立。为了解决这一问题，我们提出一个预训练步骤，用于排除这类重复个体。预训练步骤概述如下：

> **(1)** 假设可以通过一组特征唯一识别某个个体。例如，国民身份证号码和国籍可以组合起来唯一识别一个人。**(2)** 为了识别重复个体，我们首先使用 MD5、SHA256 等哈希函数对上述特征进行编码。**(3)** 然后，我们将编码后的序列发送到中央服务器。**(4)** 服务器会收集所有数据源的编码序列，并检查其中是否存在重复的编码序列。**(5)** 所有重复序列都对应重复个体。因此，我们通知各数据源在训练过程中排除这些个体。

我们在图6中总结了使用三个数据源进行的预训练步骤。

B 可识别性

如果混杂因素未被观测到，因果效应将无法识别。然而，Louizos 等（2017）证明，如果能够恢复联合分布 $p_s(\boldsymbol{x}^s,\ y^s,\ w^s,\ \boldsymbol{z}^s)$，那么因果效应就是可识别的。接下来，我们将说明如何识别这些因果效应。

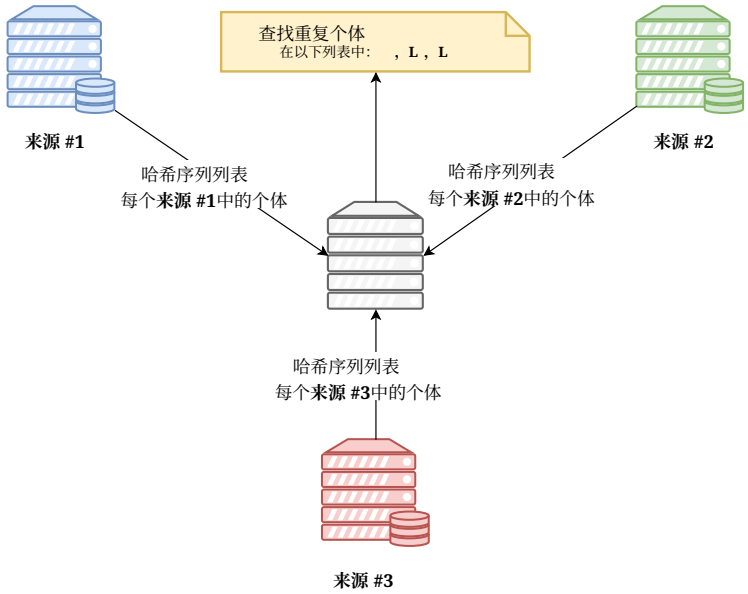


图6：预训练步骤的示意图。此步骤旨在识别数据源中的重复个体。此外，此步骤可保护隐私，因为每个数据源仅发送个体序列的哈希值。

证明。 证明改编自 Louizos 等人（定理 1，2017）。我们需要证明，分布 $p_{\mathsf{s}}(y^{\mathsf{s}}|\text{do}(W=\boldsymbol{w}^{\mathsf{s}})\ \boldsymbol{x}^{\mathsf{s}})$ 可以从观察性数据中识别。我们有

$$p_{\mathsf{s}}(y^{\mathsf{s}}|\text{do}(W=\boldsymbol{w}^{\mathsf{s}}),\boldsymbol{x}^{\mathsf{s}}) = \int p_{\mathsf{s}}(y^{\mathsf{s}}|\text{do}(W=\boldsymbol{w}^{\mathsf{s}}),\boldsymbol{x}^{\mathsf{s}},\boldsymbol{z}^{\mathsf{s}})p_{\mathsf{s}}(\boldsymbol{z}^{\mathsf{s}}|\text{do}(W=\boldsymbol{w}^{\mathsf{s}}),\boldsymbol{x}^{\mathsf{s}})d\boldsymbol{z}^{\mathsf{s}}$$
$$= \int p_{\mathsf{s}}(y^{\mathsf{s}}|\boldsymbol{w}^{\mathsf{s}},\boldsymbol{x}^{\mathsf{s}},\boldsymbol{z}^{\mathsf{s}})p_{\mathsf{s}}(\boldsymbol{z}^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}})d\boldsymbol{z}^{\mathsf{s}}.$$

其中，最后一个等式是通过应用 do 演算得到的。最后一个表达式 $\int p_{\mathsf{s}}(y^{\mathsf{s}}|\boldsymbol{w}^{\mathsf{s}},\boldsymbol{x}^{\mathsf{s}},\boldsymbol{z}^{\mathsf{s}})p_{\mathsf{s}}(\boldsymbol{z}^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}})$ $d\boldsymbol{z}^{\mathsf{s}}$，可由联合分布 $p_{\mathsf{s}}(\boldsymbol{x}^{\mathsf{s}},y^{\mathsf{s}},w^{\mathsf{s}},\boldsymbol{z}^{\mathsf{s}})$ 确定。在我们的工作中，$p_{\mathsf{s}}(\boldsymbol{x}^{\mathsf{s}},y^{\mathsf{s}},w^{\mathsf{s}},\boldsymbol{z}^{\mathsf{s}})$ 通过对分布 $p_{\mathsf{s}}(w^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}})$、$p_{\mathsf{s}}(y^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}},w^{\mathsf{s}})$、$p_{\mathsf{s}}(\boldsymbol{z}^{\mathsf{s}}|\boldsymbol{x}^{\mathsf{s}},y^{\mathsf{s}},w^{\mathsf{s}})$ $p_{\mathsf{s}}(y^{\mathsf{s}}|w^{\mathsf{s}},\boldsymbol{z}^{\mathsf{s}})$ 和 $p(\boldsymbol{z}^{\mathsf{s}})$ 进行因子分解来恢复。在联邦设置中自适应地学习这些分布是我们工作的主要任务。至此，证明完成。 □

C 计算条件平均处理效应、局部平均处理效应和全局平均处理效应

本节详细介绍训练模型后如何计算条件平均处理效应、局部平均处理效应和全局平均处理效应。

C.1 计算条件平均处理效应和局部平均处理效应

训练模型后，每个数据源都可以计算其自身数据源的条件平均处理效应和局部平均处理效应，并将其用于自身。

$$E[y_i^{\mathsf{s}}|\text{do}(w_i^{\mathsf{s}}=w),\boldsymbol{x}_i^{\mathsf{s}}] = \int E[y_i^{\mathsf{s}}|w_i^{\mathsf{s}}=w,\boldsymbol{z}_i^{\mathsf{s}}]p(\boldsymbol{z}_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}})d\boldsymbol{z}_i^{\mathsf{s}} \simeq \frac{1}{N}\sum_{l=1}^{N} f_y(w_i^{\mathsf{s}}=w,\boldsymbol{z}_i^{\mathsf{s}}[l])$$

其中，$f_y(w_i^{\mathsf{s}}=w,\boldsymbol{z}_i^{\mathsf{s}}[l])$ 是 $p_{\mathsf{s}}(y_i^{\mathsf{s}}|w_i^{\mathsf{s}},\boldsymbol{z}_i^{\mathsf{s}})$ 和 $\{z_i^{\mathsf{s}}[l]\}_{l=1}^{N} \overset{i.i.d.}{\sim} p_{\mathsf{s}}(\boldsymbol{z}_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}})$ 的均值函数。

问题是从 $p_{\mathsf{s}}(\boldsymbol{z}_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}})$ 中抽取 $\{\boldsymbol{z}_i^{\mathsf{s}}[l]\}_{l=1}^{N}$。我们观察到

$$p_{\mathsf{s}}(\boldsymbol{z}_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}}) = \sum_{w_i^{\mathsf{s}}\in\{0,1\}} \int p_{\mathsf{s}}(\boldsymbol{z}_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}},y_i^{\mathsf{s}},w_i^{\mathsf{s}})p_{\mathsf{s}}(y_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}},w_i^{\mathsf{s}})p_{\mathsf{s}}(w_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}})\,dy_i^{\mathsf{s}}.$$

因此，若要抽取样本，我们按以下步骤进行：

(1) 从 $p_{\mathsf{s}}(w_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}})$ 中抽取一个 w_i^{s} 样本。(2) 将上述 w_i^{s} 样本代入 $p_{\mathsf{s}}(y_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}}, w_i^{\mathsf{s}})$。(3) 从 $p_{\mathsf{s}}(y_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}}, w_i^{\mathsf{s}})$ 中抽取一个 y_i^{s} 样本。(4) 将上述 y_i^{s} 样本代入 $p_{\mathsf{s}}(\boldsymbol{z}_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}}, y_i^{\mathsf{s}}, w_i^{\mathsf{s}})$。(5) 从 $p_{\mathsf{s}}(\boldsymbol{z}_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}}, y_i^{\mathsf{s}}, w_i^{\mathsf{s}})$ 中抽取一个 $\boldsymbol{z}_i^{\mathsf{s}}$ 样本。

训练模型后，即可获得 $p_{\mathsf{s}}(y_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}},\ w_i^{\mathsf{s}})$ 和 $p_{\mathsf{s}}(w_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}})$ 的概率密度函数。正如正文所述，有两种方法可以从 $p_{\mathsf{s}}(\boldsymbol{z}_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}},\ y_i^{\mathsf{s}},\ w_i^{\mathsf{s}})$ 中抽样。第一种方法是从 $q(\boldsymbol{x}_i^{\mathsf{s}})$ 中抽样，因为它可以近似 $p_{\mathsf{s}}(\boldsymbol{z}_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}},\ y_i^{\mathsf{s}},\ w_i^{\mathsf{s}})$。第二种方法是使用带独立采样器的 Metropolis-Hastings 算法（Liu 1996）。对于第二种方法，我们有

$$p_{\mathsf{s}}(\boldsymbol{z}_i^{\mathsf{s}}|\boldsymbol{x}_i^{\mathsf{s}}, y_i^{\mathsf{s}}, w_i^{\mathsf{s}}) \propto p_{\mathsf{s}}(y_i^{\mathsf{s}}|\boldsymbol{z}_i^{\mathsf{s}}, w_i^{\mathsf{s}}) p_{\mathsf{s}}(w_i^{\mathsf{s}}|\boldsymbol{z}_i^{\mathsf{s}}) p_{\mathsf{s}}(\boldsymbol{x}_i^{\mathsf{s}}|\boldsymbol{z}_i^{\mathsf{s}}) p(\boldsymbol{z}_i^{\mathsf{s}}).$$

因此，它可用于计算所关注的接受概率。请注意，第二种方法会基于精确的接受概率进一步筛选样本，因此能够得到更精确的样本。

上述方法有助于估计给定 $\boldsymbol{x}_i^{\mathsf{s}}$ 时的条件平均处理效应。局部平均处理效应是数据源 s 中各个体的条件平均处理效应的平均值。这些量可以在本地数据源的机器上估计。下一节将介绍如何计算全局平均处理效应。

C.2 根据各数据源的局部平均处理效应计算全局平均处理效应

要计算全局平均处理效应，服务器会收集每个数据源的局部平均处理效应，然后计算它们的加权平均值。例如，假设有三个数据源，其局部平均处理效应值分别为 7.0、8.5 和 6.8。这些局部平均处理效应分别基于 10、5 和 12 个个体计算得出。那么，全局平均处理效应如下所示：

$$\text{global ATE} = \frac{10 \times 7.0 + 8 \times 8.5 + 12 \times 6.8}{10 + 8 + 12} = 7.32.$$

由于每个数据源仅共享其局部平均处理效应和个体数量，因此不会泄露有关这些个体的任何敏感信息。

D 比较指标

我们在实验中报告了两个误差指标：

- 异质性效应估计精度（PEHE）：

$$\epsilon_{\text{PEHE}} = \sum_{i=1}^{n} (\tau(\boldsymbol{x}_i) - \hat{\tau}(\boldsymbol{x}_i))^2 / n, \tag{17}$$

- 绝对误差：

$$\epsilon_{\text{ATE}} = |\tau - \hat{\tau}|, \tag{18}$$

其中，$\tau(\boldsymbol{x}_i)$ τ 分别是真实的个体处理效应和平均处理效应，$\hat{}\tau(\boldsymbol{x}_i)$，$\hat{}\tau$ 则是它们的估计值。我们报告了采用不同随机初始化的训练算法对数据进行 10 次重复实验的均值和标准误。

E 损失函数的推导

在本节中，我们将介绍损失函数以及用于调节目标分布的函数形式。

E.1 涉及潜在混杂因素的分布学习

对数边际似然的证据下界表达式如下

$$\begin{aligned}\log p(\mathbf{x},\mathbf{y},\mathbf{w}) &= \log \int p(\mathbf{x},\mathbf{y},\mathbf{w},\mathbf{z})d\mathbf{z} \\ &\geq \int q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{y},\mathbf{w},\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\ &= \sum_{\mathsf{s}\in\mathcal{S}} \sum_{i=1}^{n_{\mathsf{s}}} \Big(E_q\big[\log p_{\mathsf{s}}(y_i^{\mathsf{s}}|w_i^{\mathsf{s}}, \boldsymbol{z}_i^{\mathsf{s}}) + \log p_{\mathsf{s}}(w_i^{\mathsf{s}}|\boldsymbol{z}_i^{\mathsf{s}}) + \log p_{\mathsf{s}}(\boldsymbol{x}_i^{\mathsf{s}}|\boldsymbol{z}_i^{\mathsf{s}})\big] - \mathrm{KL}[q(\boldsymbol{z}_i^{\mathsf{s}})\|p(\boldsymbol{z}_i^{\mathsf{s}})] \Big) =: \mathcal{L}.\end{aligned}$$

使用完整数据集 $\tilde{\mathbf{D}}^{\mathsf{s}} = \bigcup_{l=1}^{M} \{(w_i^{\mathsf{s}},\ y_i^{\mathsf{s}},\ \boldsymbol{x}_i^{\mathsf{s}},\ \boldsymbol{z}_i^{\mathsf{s}}[l])\}_{i=1}^{n_{\mathsf{s}}},\ \forall \mathsf{s} \in \mathcal{S}$，我们最小化以下损失函数 J：

$$J = \widehat{\mathcal{L}} + \sum_{c\in\mathcal{A}} R(f_c), \qquad \mathcal{A} = \{y_0, y_1, q_0, q_q, x, w\},$$

其中，$\mathcal{L}$ 是由 $\mathcal{L}$取负得到的经验损失函数。接下来，我们根据表示定理推导 f_c 的形式。

我们进一步定义 $f_x = [f_{x,1},...,f_{x,d_x}]$，其中 $f_{x,\ d}$ 是一个以 $\boldsymbol{z}_i^{\mathsf{s}}$ 为输入并将其映射为 $\mathbb{R}$ 中实数值的函数。类似地，$f_{q_0} = [f_{q_0,1},...,f_{q_0,d_z}]$ 和$f_{q_1} = [f_{q_1,1},...,f_{q_1,d_z}]$。
令 $\mathcal{H}_c$ $(c \in \mathcal{A})$ 为再生核希尔伯特空间（RKHS），$\kappa_c(\cdot,\ \cdot)$ 为与 $\mathcal{H}_c$相关联的核函数。我们定义 $\mathcal{B}_c$ 如下:

$$\begin{aligned}
\mathcal{B}_{y_0} &= \mathsf{span}\big\{\kappa_{y0}(\cdot, \boldsymbol{z}_i^{\mathsf{s}}[l]),\ \text{where } \mathsf{s}\in\mathcal{S}; i = 1,...,n_{\mathsf{s}}; l = 1,...,M\big\}, \\
\mathcal{B}_{y_1} &= \mathsf{span}\big\{\kappa_{y1}(\cdot, \boldsymbol{z}_i^{\mathsf{s}}[l]),\ \text{where } \mathsf{s}\in\mathcal{S}; i = 1,...,n_{\mathsf{s}}; l = 1,...,M\big\}, \\
\mathcal{B}_{x} &= \mathsf{span}\left\{\kappa_{x}(\cdot, \boldsymbol{z}_i^{\mathsf{s}}[l]),\ \text{where } \mathsf{s}\in\mathcal{S}; i = 1,...,n_{\mathsf{s}}; l = 1,...,M\right\}, \\
\mathcal{B}_{w} &= \mathsf{span}\left\{\kappa_{w}(\cdot, \boldsymbol{z}_i^{\mathsf{s}}[l]),\ \text{where } \mathsf{s}\in\mathcal{S}; i = 1,...,n_{\mathsf{s}}; l = 1,...,M\right\}, \\
\mathcal{B}_{q_0} &= \mathsf{span}\left\{\kappa_{q0}(\cdot, [\boldsymbol{x}_i^{\mathsf{s}}, y_i^{\mathsf{s}}]),\ \text{where } \mathsf{s}\in\mathcal{S}; i = 1,...,n_{\mathsf{s}}\right\}, \\
\mathcal{B}_{q_1} &= \mathsf{span}\left\{\kappa_{q1}(\cdot, [\boldsymbol{x}_i^{\mathsf{s}}, y_i^{\mathsf{s}}]),\ \text{where } \mathsf{s}\in\mathcal{S}; i = 1,...,n_{\mathsf{s}}\right\}.
\end{aligned}$$

我们提出以下正则化项:

$$R(f_{y0}) = \mathsf{reg_factor}_{y0} \times \|f_{y0}\|^2_{\mathcal{H}_{y0}}, \quad R(f_x) = \sum_{d=1}^{d_x} \mathsf{reg_factor}_{x,d} \times \|f_{x,d}\|^2_{\mathcal{H}_x} \quad (d = 1,...,d_x).$$

正则化项 $R(f_{y1})$ 和 $R(f_w)$ 与 $R(f_{y0})$ 类似，而 $R(f_{q0})$、$R(f_{q1})$ 与 $R(f_x)$ 类似。

我们可以看到，$\mathcal{B}_c$ 是 $\mathcal{H}_c$的子空间。我们分别将 f_{y0}、f_{y1}、f_w、$f_{x,\ d}$ $(d = 1,...,d_x)$、$f_{q0,\ d}$ $(d = 1,...,d_z)$以及$f_{q1,\ d}$ $(d = 1,...,d_z)$投影到子空间 $\mathcal{B}_{y0}$、$\mathcal{B}_{y1}$、$\mathcal{B}_w$、$\mathcal{B}_x$、$\mathcal{B}_{q0}$ 和 $\mathcal{B}_{q1}$上，得到f'_{y0}、f'_{y1}、f'_w、$f'_{x,\ d}$、$f'_{q0,\ d}$以及$f'_{q1,\ d}$。接下来，我们还将它们投影到 $\mathcal{B}_{(\cdot)}$的正交补空间上，得到 $f^{\perp}_{y0}$、$f^{\perp}_{y1}$、$f^{\perp}_w$、$f^{\perp}_{x,\ d}$、$f^{\perp}_{q0,\ d}$ 以及$f^{\perp}_{q1,\ d}$。
注意到 $f_{(\cdot)} = f'_{(\cdot)} + f^{\perp}_{(\cdot)}$。因此，$\|f_{(\cdot)}\|^2_{\mathcal{H}_{(\cdot)}} = \|f'_{(\cdot)}\|^2_{\mathcal{H}_{(\cdot)}} + \|f^{\perp}_{(\cdot)}\|^2_{\mathcal{H}_{(\cdot)}} \geq \|f'_{(\cdot)}\|^2_{\mathcal{H}_{(\cdot)}}$，这意味着当 $f_{(\cdot)}$ 位于其子空间 $\mathcal{B}_{(\cdot)}$ 中时，$\mathsf{reg_factor}_{(\cdot)} \times \|f_{(\cdot)}\|^2_{\mathcal{H}_{(\cdot)}}$ 取得最小值。 **(I)**
此外，根据再生性质，我们有

$$f_{y_0}(\boldsymbol{z}_i^{\mathsf{s}}[l]) = \big\langle f_{y_0}, \kappa_{y_0}(\cdot, \boldsymbol{z}_i^{\mathsf{s}}[l])\big\rangle_{\mathcal{H}_y} = \big\langle f'_{y_0}, \kappa_{y_0}(\cdot, \boldsymbol{z}_i^{\mathsf{s}}[l])\big\rangle_{\mathcal{H}_y} + \big\langle f^{\perp}_{y_0}, \kappa_{y_0}(\cdot, \boldsymbol{z}_i^{\mathsf{s}}[l])\big\rangle_{\mathcal{H}_y} = f'_{y_0}(\boldsymbol{z}_i^{\mathsf{s}}[l]).$$

类似地，我们还有 $f_{y1}(\boldsymbol{z}_i^{\mathsf{d}}[l]) = f'_{y1}(\boldsymbol{z}_i^{\mathsf{d}}[l])$、$f_w(\boldsymbol{z}_i^{\mathsf{d}}[l]) = f'_w(\boldsymbol{z}_i^{\mathsf{d}}[l])$、$f_{x,\ d}(\boldsymbol{z}_i^{l}) = f'_{x,\ d}(\boldsymbol{z}_i^{\mathsf{d}}[l])$、$f_{q0,\ d}(y_i^{\mathsf{d}},\ \boldsymbol{x}_i^{\mathsf{d}}) = f'_{q0,\ d}(y_i^{\mathsf{d}},\ \boldsymbol{x}_i^{\mathsf{d}})$ 以及 $f_{q_1,\ d}(y_i^{\mathsf{d}},\ \boldsymbol{x}_i^{\mathsf{d}}) = f'_{q_1,\ d}(y_i^{\mathsf{d}},\ \boldsymbol{x}_i^{\mathsf{d}})$。因此，

$$\widehat{\mathcal{L}}(f_{y0}, f_{y1}, f_{q0}, f_{q1}, f_x, f_w) = \widehat{\mathcal{L}}(f'_{y0}, f'_{y1}, f'_{q0}, f'_{q1}, f'_x, f'_w). \qquad \textbf{(II)}$$

(I) 并且（**II**）意味着 f_{y0}，f_{y1}，$f_{q0,\ d}$，$f_{q1,\ d}$，$f_{x,\ d}$，f_w是对应<style id="">子空间中元素的加权和 ir

相应的子空间。因此，

$$\boxed{f_c(\boldsymbol{u}^{\mathsf{s}}) = \sum_{\mathsf{v}\in\mathcal{S}} \sum_{j=1}^{n_{\mathsf{v}}\times M} \kappa(\boldsymbol{u}^{\mathsf{s}}, \boldsymbol{u}_j^{\mathsf{v}})\boldsymbol{\alpha}_j^{\mathsf{v}}.}$$

将这种形式与正文（第 4.1 节）中介绍的自适应核和随机傅里叶特征结合使用，即可得到所需的模型。

E.2 学习辅助分布

J_w、 J_y 的推导以及调制辅助分布的函数形式，与第 E.1 节中详细介绍的 J 类似。不同之处在于，经验损失函数是由负对数似然而非证据下界得出的。

F 一些常用核函数的谱分布

表 6（摘自 Milton 等人（2019））列出了一些常用核函数及其对应的谱密度 $s(\boldsymbol{\omega})$。这些密度函数用于抽取正文第 4 节所介绍的随机傅里叶特征所需的 $\boldsymbol{\omega}$ 样本。在我们的实验中，使用了高斯核。

表 6：一些常用核函数及其对应的谱密度。注意， $K_\nu(\cdot)$ 表示第二类修正贝塞尔函数， $\Gamma(\cdot)$ 表示伽马函数。

核	**核函数**， $k(\boldsymbol{x}_1 - \boldsymbol{x}_2)$	**谱密度**， $s(\boldsymbol{\omega})$
高斯	$\exp\left(-\frac{\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2}{2\ell^2}\right)$	$\left(\frac{2\pi}{\ell^2}\right)^{\frac{-d}{2}} \exp\left(-\frac{\ell^2\|\boldsymbol{\omega}\|_2^2}{2}\right)$
拉普拉斯	$\exp\left(-\ell\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_1\right)$	$\left(\frac{2}{\pi}\right)^{\frac{d}{2}} \prod_{i=1}^{d} \frac{\ell}{\ell^2+\omega_i^2}$
Matérn	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2}{\ell}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2}{\ell}\right)$	$\frac{2^d\pi^{\frac{d}{2}}\Gamma(\nu+\frac{d}{2})(2\nu)^\nu}{\Gamma(\nu)\ell^{2\nu}} \left(\frac{2\nu}{\ell^2}+4\pi^2\|\boldsymbol{\omega}\|_2^2\right)^{-\left(\nu+\frac{d}{2}\right)}$

G 引理 1 的证明

设 $\mathcal{S}_{\setminus \mathsf{s}} := \mathcal{S} \setminus \{\mathsf{s}\}$。模型总结如下：

$$\begin{aligned}
p(\boldsymbol{z}_i^{\mathsf{s}}) &= \mathsf{N}(0, \sigma_z^2 \mathbf{I}_{d_z}),\\
p(w_i^{\mathsf{s}}|\boldsymbol{z}_i^{\mathsf{s}}) &= \mathsf{Bern}\Big(\varphi\Big(\Big(\theta_w^{\mathsf{s}} + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}\theta_w^{\mathsf{v}}\Big)^\top \phi(\boldsymbol{z}_i^{\mathsf{s}})\Big)\Big),\\
p(y_i^{\mathsf{s}}|w_i^{\mathsf{s}}, \boldsymbol{z}_i^{\mathsf{s}}) &= \mathsf{N}\Big(\Big(w_i^{\mathsf{s}}\Big(\theta_{y1}^{\mathsf{s}} + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}\theta_{y1}^{\mathsf{v}}\Big) + (1-w_i^{\mathsf{s}})\Big(\theta_{y0}^{\mathsf{s}} + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}\theta_{y0}^{\mathsf{v}}\Big)\Big)^\top \phi(\boldsymbol{z}_i^{\mathsf{s}}), \sigma_y^2\Big),\\
p(\boldsymbol{x}_i^{\mathsf{s}}|\boldsymbol{z}_i^{\mathsf{s}}) &= \mathsf{N}\Big(\Big(\theta_x^{\mathsf{s}} + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}\theta_x^{\mathsf{v}}\Big)^\top \phi(\boldsymbol{z}_i^{\mathsf{s}}), \sigma_x^2 \mathbf{I}_{d_x}\Big),
\end{aligned}$$

其中 $\boldsymbol{z}_i^{(\cdot)} \in \mathbb{R}^{d_z}$, $y_i^{(\cdot)} \in \mathbb{R}$, $w_i^{(\cdot)} \in \{0,1\}$, $\boldsymbol{x}_i^{(\cdot)} \in \mathbb{R}^{d_x}$, $\lambda > 0$。

设 $\theta = \{\theta_w^{\mathsf{s}}, \theta_{y0}^{\mathsf{s}}, \theta_{y1}^{\mathsf{s}}, \theta_x^{\mathsf{s}}\}_{\mathsf{s}\in\mathcal{S}}$。设 $\mathcal{V}_w$, $\mathcal{V}_{y0}$, $\mathcal{V}_{y1}$, $\mathcal{V}_x$ 为单位 $\|\cdot\|_2$-球的 $1/(2\sqrt{m})$-打包，其基数分别至少为 $(2\sqrt{m})^{2B}$、$(2\sqrt{m})^{2B}$、$(2\sqrt{m})^{2B}$、$(2\sqrt{m})^{2B d_x}$。设 $\mathcal{V}^{\mathsf{s}} = \overline{\delta}(\mathcal{V}_w \times \mathcal{V}_{y0} \times \mathcal{V}_{y1} \times \mathcal{V}_x)$ 和 $\mathcal{V} = \mathcal{V}^{\mathsf{s}_1} \times \mathcal{V}^{\mathsf{s}_2} \times \ldots \times \mathcal{V}^{\mathsf{s}_m}$。我们看到

$$|\mathcal{V}| \geq (2\sqrt{m})^{2mB(d_x+3)}.$$

下面我们推导极小极大界：

证明。 我们有

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 = \sqrt{\sum_{\mathsf{s}\in\mathcal{S}}\sum_{c\in\mathcal{A}} \|(\theta_c^{\mathsf{s}})_1 - (\theta_c^{\mathsf{s}})_2\|_2^2} \geq \sqrt{\sum_{\mathsf{s}\in\mathcal{S}} 4\left(\frac{\delta}{2\sqrt{m}}\right)^2} = \delta.$$

边缘分布

$$p_{\boldsymbol{\theta}}(w, y, \boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(w, y, \boldsymbol{x}, \boldsymbol{z})d\boldsymbol{z} = \int p_{\boldsymbol{\theta}}(y|w, \boldsymbol{z})p_{\boldsymbol{\theta}}(w|\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}.$$

此外，我们有

$$D_{\mathrm{KL}}(p^n_{\boldsymbol{\theta}_1} \,\|\, p^n_{\boldsymbol{\theta}_2}) = \sum_{\mathsf{s}\in\mathcal{S}} D_{\mathrm{KL}}(p^{n_{\mathsf{s}}}_{\boldsymbol{\theta}_1} \,\|\, p^{n_{\mathsf{s}}}_{\boldsymbol{\theta}_2}).$$

我们将证明分为三部分**(I)**、**(II)**和**(III)**：

(I) $D_{\mathbf{KL}}(p^{n_{\mathsf{s}}}_{\theta_1} \,\|\, p^{n_{\mathsf{s}}}_{\theta_2})$**的上界**

由于数据相互独立，我们有

$$\begin{aligned}
&D_{\mathrm{KL}}(p^{n_{\mathsf{s}}}_{\boldsymbol{\theta}_1} \,\|\, p^{n_{\mathsf{s}}}_{\boldsymbol{\theta}_2}) = n_{\mathsf{s}} D_{\mathrm{KL}}(p^1_{\boldsymbol{\theta}_1} \,\|\, p^1_{\boldsymbol{\theta}_2}) \\
&\leq n_{\mathsf{s}} \int D_{\mathrm{KL}}\left(p_{\boldsymbol{\theta}_1}(y|w, \boldsymbol{z})p_{\boldsymbol{\theta}_1}(w|\boldsymbol{z})p_{\boldsymbol{\theta}_1}(\boldsymbol{x}|\boldsymbol{z}) \middle\| p_{\boldsymbol{\theta}_2}(y|w, \boldsymbol{z}')p_{\boldsymbol{\theta}_2}(w|\boldsymbol{z}')p_{\boldsymbol{\theta}_2}(\boldsymbol{x}|\boldsymbol{z}')\right) p(\boldsymbol{z})p(\boldsymbol{z}')d\boldsymbol{z}d\boldsymbol{z}' \\
&= n_{\mathsf{s}} \int \Big[p_{\boldsymbol{\theta}_1}(w=0|\boldsymbol{z}) D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(y|w=0, \boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(y|w=0, \boldsymbol{z}')\big] \\
&\qquad + p_{\boldsymbol{\theta}_1}(w=1|\boldsymbol{z}) D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(y|w=1, \boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(y|w=1, \boldsymbol{z}')\big] \\
&\qquad + D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(w|\boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(w|\boldsymbol{z}')\big] + D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(\boldsymbol{x}|\boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(\boldsymbol{x}|\boldsymbol{z}')\big] \Big] p(\boldsymbol{z})p(\boldsymbol{z}')d\boldsymbol{z}d\boldsymbol{z}'.
\end{aligned}$$

接下来，我们求出每个分量的上界。

◇ 第一项和第二项的上界

$$\begin{aligned}
&p_{\boldsymbol{\theta}_1}(w=0|\boldsymbol{z}) D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(y|w=0, \boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(y|w=0, \boldsymbol{z}')\big] \\
&\qquad \leq \frac{1}{2\sigma_y^2}\left(\left((\theta^{\mathsf{s}}_{y0})_1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}(\theta^{\mathsf{v}}_{y0})_1\right)^\top \phi(\boldsymbol{z}) - \left((\theta^{\mathsf{s}}_{y0})_2 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}(\theta^{\mathsf{v}}_{y0})_2\right)^\top \phi(\mathsf{z}')\right)^2 \\
&\qquad \leq \frac{8B^2\delta^2(1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}})^2}{\sigma_y^2}.
\end{aligned}$$

类似地，我们也有

$$p_{\boldsymbol{\theta}_1}(w=1|\boldsymbol{z}) D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(y|w=1, \boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(y|w=1, \boldsymbol{z}')\big] \leq \frac{8B^2\delta^2(1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}})^2}{\sigma_y^2}.$$

◇ 第三项的上界

$$\begin{aligned}
&D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(w|\boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(w|\boldsymbol{z}')\big] \\
&= \varphi\left(\left((\theta^{\mathsf{s}}_w)_1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}(\theta^{\mathsf{v}}_w)_1\right)^\top \phi(\boldsymbol{z})\right) \log \frac{\varphi\left(\left((\theta^{\mathsf{s}}_w)_1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}(\theta^{\mathsf{v}}_w)_1\right)^\top \phi(\boldsymbol{z})\right)}{\varphi\left(\left((\theta^{\mathsf{s}}_w)_2 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}(\theta^{\mathsf{v}}_w)_2\right)^\top \phi(\boldsymbol{z}')\right)} \\
&\quad + \varphi\Big(-\left((\theta^{\mathsf{s}}_w)_1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}(\theta^{\mathsf{v}}_w)_1\right)^\top \phi(\boldsymbol{z})\Big) \log \frac{\varphi\Big(-\left((\theta^{\mathsf{s}}_w)_1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}(\theta^{\mathsf{v}}_w)_1\right)^\top \phi(\boldsymbol{z})\Big)}{\varphi\Big(-\left((\theta^{\mathsf{s}}_w)_2 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}(\theta^{\mathsf{v}}_w)_2\right)^\top \phi(\boldsymbol{z}')\Big)}.
\end{aligned}$$

对于第一项，

$$\varphi\left(\left((\theta^{\mathsf{s}}_w)_1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}(\theta^{\mathsf{v}}_w)_1\right)^\top \phi(\mathbf{z})\right) \log \frac{\varphi\left(\left((\theta^{\mathsf{s}}_w)_1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}(\theta^{\mathsf{v}}_w)_1\right)^\top \phi(\mathbf{z})\right)}{\varphi\left(\left((\theta^{\mathsf{s}}_w)_2 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}(\theta^{\mathsf{v}}_w)_2\right)^\top \phi(\mathbf{z}')\right)}$$

$$
\begin{aligned}
&\le \Big|\log\Big(1+e^{-\left((\theta_w^{\mathsf{s}})_2+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}(\theta_w^{\mathsf{v}})_2\right)^\top\phi(\mathbf{z})}\Big)-\log\Big(1+e^{-\left((\theta_w^{\mathsf{s}})_1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}(\theta_w^{\mathsf{v}})_1\right)^\top\phi(\mathbf{z}')}\Big)\Big|\\
&\le \Big\|(\theta_w^{\mathsf{s}})_1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}(\theta_w^{\mathsf{v}})_1\Big\|_2\|\phi(\mathbf{z})\|_2+\Big\|(\theta_w^{\mathsf{s}})_2+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}(\theta_w^{\mathsf{v}})_2\Big\|_2\|\phi(\mathbf{z}')\|_2\\
&\le \Big(\delta+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\delta\Big)\|\phi(\mathbf{z})\|_2+\Big(\delta+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\delta\Big)\|\phi(\mathbf{z}')\|_2\\
&\le 4B\delta\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\Big).
\end{aligned}
$$

同样，我们也有

$$
\begin{aligned}
&\varphi\Big(-\Big((\theta_w^{\mathsf{s}})_1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}(\theta_w^{\mathsf{v}})_1\Big)^\top\phi(\boldsymbol{z})\Big)\log\frac{\varphi\Big(-\Big((\theta_w^{\mathsf{s}})_1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}(\theta_w^{\mathsf{v}})_1\Big)^\top\phi(\boldsymbol{z})\Big)}{\varphi\Big(-\Big((\theta_w^{\mathsf{s}})_2+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}(\theta_w^{\mathsf{v}})_2\Big)^\top\phi(\boldsymbol{z}')\Big)}\\
&\qquad\le 4B\delta\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\Big).
\end{aligned}
$$

因此，

$$
D_{\mathrm{KL}}\big[p_{\Theta_1}(w|\boldsymbol{z})\big\|p_{\Theta_2}(w|\boldsymbol{z}')\big]\le 8B\delta\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\Big).
$$

◇ 第四个分量的上界

$$
\begin{aligned}
&D_{\mathrm{KL}}\big[p_{\Theta_1}(\boldsymbol{x}|\boldsymbol{z})\big\|p_{\Theta_2}(\boldsymbol{x}|\boldsymbol{z}')\big]\\
&\quad=\frac{1}{2\sigma_x^2}\Big\|\Big((\theta_x^{\mathsf{s}})_1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}(\theta_x^{\mathsf{v}})_1\Big)^\top\phi(\boldsymbol{z})-\Big((\theta_x^{\mathsf{s}})_2+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}(\theta_x^{\mathsf{v}})_2\Big)^\top\phi(\boldsymbol{z}')\Big\|_2^2\\
&\quad\le\frac{1}{2\sigma_x^2}\Big(\Big\|\Big((\theta_x^{\mathsf{s}})_1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}(\theta_x^{\mathsf{v}})_1\Big)^\top\phi(\boldsymbol{z})\Big\|_2+\Big\|\Big((\theta_x^{\mathsf{s}})_2+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}(\theta_x^{\mathsf{v}})_2\Big)^\top\phi(\boldsymbol{z}')\Big\|_2\Big)^2\\
&\quad\le\frac{8B^2\delta^2\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\Big)^2}{\sigma_x^2}.
\end{aligned}
$$

(II) 合并结果

根据上述各分量的上界，我们得到

$$
\begin{aligned}
D_{\mathrm{KL}}(p_{\Theta_1}^{n_{\mathsf{s}}}\,\|\,p_{\Theta_2}^{n_{\mathsf{s}}})&\le n_{\mathsf{s}}\int\Big[\frac{16B^2\delta^2(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}})^2}{\sigma_y^2}+8B\delta\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\Big)\\
&\qquad+\frac{8B^2\delta^2(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}})^2}{\sigma_x^2}\Big]p(\boldsymbol{z})p(\boldsymbol{z}')d\boldsymbol{z}d\boldsymbol{z}'\\
&=n_{\mathsf{s}}\Big[\Big(\frac{1}{\sigma_y^2}+\frac{1}{2\sigma_x^2}\Big)16B^2\delta^2\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\Big)^2+8B\delta\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\Big)\Big].
\end{aligned}
$$

(III) 极小极大下界

我们有

$$
\begin{aligned}
D_{\mathrm{KL}}(p_{\Theta_1}^{n}\,\|\,p_{\Theta_2}^{n})&=\sum_{\mathsf{s}\in\mathcal{S}}D_{\mathrm{KL}}(p_{\Theta_1}^{n_{\mathsf{s}}}\,\|\,p_{\Theta_2}^{n_{\mathsf{s}}})\\
&\le\sum_{\mathsf{s}\in\mathcal{S}}n_{\mathsf{s}}\Big[\Big(\frac{1}{\sigma_y^2}+\frac{1}{2\sigma_x^2}\Big)16B^2\delta^2\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\Big)^2+8B\delta\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\Big)\Big].
\end{aligned}
$$

因此,

$$\begin{aligned}
&\inf_{\hat{\boldsymbol{\theta}}_n}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}(P)\|_2\right]\\
\geq&\frac{\delta}{2}\left(1-\frac{\sum_{\mathsf{s}\in\mathfrak{s}}n_{\mathsf{s}}\left[\left(\frac{1}{\sigma_y^2}+\frac{1}{2\sigma_x^2}\right)16B^2\delta^2\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)^2+8B\delta\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)\right]+\log 2}{\log|\mathcal{V}|}\right)\\
\geq&\frac{\delta}{2}\left(1-\frac{\sum_{\mathsf{s}\in\mathfrak{s}}n_{\mathsf{s}}\left[\left(\frac{1}{\sigma_y^2}+\frac{1}{2\sigma_x^2}\right)16B^2\delta^2\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)^2+8B\delta\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)\right]+\log 2}{2mB(d_x+3)\log(2\sqrt{m})}\right).
\end{aligned}$$

We choose $\delta=\frac{\sqrt{mB(d_x+3)}\log(2\sqrt{m})}{4B\sum_{\mathsf{s}\in\mathfrak{s}}n_{\mathsf{s}}\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)^2}$, then

$$\begin{aligned}
&1-\frac{\sum_{\mathsf{s}\in\mathfrak{s}}n_{\mathsf{s}}\left[\left(\frac{1}{\sigma_y^2}+\frac{1}{2\sigma_x^2}\right)16B^2\delta^2\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)^2+8B\delta\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)\right]+\log 2}{2mB(d_x+3)\log(2\sqrt{m})}\\
&\geq 1-\left(\frac{1}{\sigma_y^2}+\frac{1}{2\sigma_x^2}\right)\frac{\log(2\sqrt{m})}{2\sum_{\mathsf{s}\in\mathfrak{s}}n_{\mathsf{s}}\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)^2}-\frac{1}{\sqrt{mB(d_x+3)}}-\frac{1}{2mB(d_x+3)}\\
&\geq 1-\left(\frac{1}{\sigma_y^2}+\frac{1}{2\sigma_x^2}\right)\frac{\log(2\sqrt{m})}{2\sum_{\mathsf{s}\in\mathfrak{s}}n_{\mathsf{s}}\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)^2}-\frac{1}{2}-\frac{1}{8}.
\end{aligned}$$

如果 $\sum_{\mathsf{s}\in\mathfrak{s}}n_{\mathsf{s}}\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)^2\geq 2\left(\frac{1}{\sigma_y^2}+\frac{1}{2\sigma_x^2}\right)\log(2\sqrt{m})$，则

$$\begin{aligned}
\inf_{\hat{\boldsymbol{\theta}}_n}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}(P)\|_2\right]&\geq\frac{1}{2}\times\frac{\sqrt{mB(d_x+3)}\log(2\sqrt{m})}{4B\sum_{\mathsf{s}\in\mathfrak{s}}n_{\mathsf{s}}\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)^2}\times\left(1-\frac{1}{4}-\frac{1}{2}-\frac{1}{8}\right)\\
&=\frac{\sqrt{m(d_x+3)}\log(2\sqrt{m})}{64\sqrt{B}\sum_{\mathsf{s}\in\mathfrak{s}}n_{\mathsf{s}}\left(1+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)^2}.
\end{aligned}$$

证明完毕。 □

H 引理 2 的证明

引理 2 的证明分为两部分（**i**）和（**ii**）。我们分别进行计算:

H.1 第（i）部分的证明

我们将模型总结如下

$$w^{\mathsf{s}}\sim\mathsf{Bern}\left(\varphi\left(\left(\psi^{\mathsf{s}}+\sum_{\mathsf{v}\in\mathfrak{s}_{\setminus\mathsf{s}}}\gamma^{\mathsf{s},\mathsf{v}}\psi^{\mathsf{v}}\right)^\top\phi(\boldsymbol{x}^{\mathsf{s}})\right)\right).$$

设 $\boldsymbol{\psi}=\{\psi^{\mathsf{s}}\}_{\mathsf{s}\in\mathfrak{s}}$。设 $\mathcal{V}_{\mathsf{s}}$ 为单位 $\|\cdot\|_2$球的 $1/(2\sqrt{m})$-打包，其基数至少为 $(2\sqrt{m})^{2B}$。现在我们选择一个集合 $\mathcal{V}=\delta(\mathcal{V}_{\mathsf{s}_1}\times\mathcal{V}_{\mathsf{s}_2}\times\ldots\times\mathcal{V}_{\mathsf{s}_m})$。我们看到

$$|\mathcal{V}|\geq(2\sqrt{m})^{2mB}.$$

证明。 我们有

$$\|\boldsymbol{\psi}_1-\boldsymbol{\psi}_2\|_2=\sqrt{\sum_{\mathsf{s}\in\mathfrak{s}}\|\psi_1^{\mathsf{s}}-\psi_2^{\mathsf{s}}\|_2^2}\geq\delta/2.$$

此外，

$$D_{\mathrm{KL}}(p^n_{\boldsymbol{\psi}_1} \,\|\, p^n_{\boldsymbol{\psi}_2}) = \sum_{\mathsf{s}\in\boldsymbol{\mathcal{S}}} D_{\mathrm{KL}}(p^{n_{\mathsf{s}}}_{\boldsymbol{\psi}_1} \,\|\, p^{n_{\mathsf{s}}}_{\boldsymbol{\psi}_2}).$$

我们首先求出 $D_{\mathrm{KL}}(p^{n_{\mathsf{s}}}_{\boldsymbol{\psi}_1} \,\|\, p^{n_{\mathsf{s}}}_{\boldsymbol{\psi}_2})$ 的上界。由于数据相互独立，我们有

$$\begin{aligned}
D_{\mathrm{KL}}(p^{n_{\mathsf{s}}}_{\boldsymbol{\psi}_1} \,\|\, p^{n_{\mathsf{s}}}_{\boldsymbol{\psi}_2}) &= n_{\mathsf{s}} D_{\mathrm{KL}}(p^1_{\boldsymbol{\psi}_1} \,\|\, p^1_{\boldsymbol{\psi}_2}) \\
&= n_{\mathsf{s}}\Bigg[\varphi\Big(\Big(\psi_1^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_1^{\mathsf{v}}\Big)^\top \phi(\boldsymbol{x}^{\mathsf{s}})\Big) \log \frac{\varphi\Big(\Big(\psi_1^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_1^{\mathsf{v}}\Big)^\top \phi(\boldsymbol{x}^{\mathsf{s}})\Big)}{\varphi\Big(\Big(\psi_2^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_2^{\mathsf{v}}\Big)^\top \phi(\boldsymbol{x}^{\mathsf{s}})\Big)} \\
&\qquad + \varphi\Big(-\Big(\psi_1^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_1^{\mathsf{v}}\Big)^\top \phi(\boldsymbol{x}^{\mathsf{s}})\Big) \log \frac{\varphi\Big(-\Big(\psi_1^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_1^{\mathsf{v}}\Big)^\top \phi(\boldsymbol{x}^{\mathsf{s}})\Big)}{\varphi\Big(-\Big(\psi_2^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_2^{\mathsf{v}}\Big)^\top \phi(\boldsymbol{x}^{\mathsf{s}})\Big)}\Bigg].
\end{aligned}$$

第一项:

$$\begin{aligned}
&\varphi\Big(\Big(\psi_1^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_1^{\mathsf{v}}\Big)^\top \phi(\mathbf{x}^{\mathsf{s}})\Big) \log \frac{\varphi\Big(\Big(\psi_1^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_1^{\mathsf{v}}\Big)^\top \phi(\mathbf{x}^{\mathsf{s}})\Big)}{\varphi\Big(\Big(\psi_2^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_2^{\mathsf{v}}\Big)^\top \phi(\mathbf{x}^{\mathsf{s}})\Big)} \\
&\quad\le \left|\log\Big(1 + e^{-\left(\psi_2^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_2^{\mathsf{v}}\right)^\top \phi(\mathbf{x}^{\mathsf{s}})}\Big) - \log\Big(1 + e^{-\left(\psi_1^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_1^{\mathsf{v}}\right)^\top \phi(\mathbf{x}^{\mathsf{s}})}\Big)\right| \\
&\quad\overset{(\star)}{\le} \left|\Big(\psi_2^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_2^{\mathsf{v}}\Big)^\top \phi(\mathbf{x}^{\mathsf{s}}) - \Big(\psi_1^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_1^{\mathsf{v}}\Big)^\top \phi(\mathbf{x}^{\mathsf{s}})\right| \\
&\quad\le 4B\delta\Big(1 + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\Big),
\end{aligned}$$

where $(\star)$ 源于 SoftPlus函数 $\log(1+e^x)$ 是 1-Lipschitz 的这一事实。特别地,

$$\big|\log(1+e^{x_1}) - \log(1+e^{x_2})\big| = \left|\int_{x_1}^{x_2} \frac{e^x}{1+e^x} dx\right| \le \left|\int_{x_1}^{x_2} 1 dx\right| = |x_1 - x_2|.$$

同样地，对于第二项，我们也有

$$\begin{aligned}
&\varphi\Big(-\Big(\psi_1^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_1^{\mathsf{v}}\Big)^\top \phi(\boldsymbol{x}^{\mathsf{s}})\Big) \log \frac{\varphi\Big(-\Big(\psi_1^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_1^{\mathsf{v}}\Big)^\top \phi(\boldsymbol{x}^{\mathsf{s}})\Big)}{\varphi\Big(-\Big(\psi_2^{\mathsf{s}} + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\psi_2^{\mathsf{v}}\Big)^\top \phi(\boldsymbol{x}^{\mathsf{s}})\Big)} \\
&\qquad\qquad\le 4B\delta\Big(1 + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\Big).
\end{aligned}$$

因此，

$$D_{\mathrm{KL}}(p^{n_{\mathsf{s}}}_{\boldsymbol{\psi}_1} \,\|\, p^{n_{\mathsf{s}}}_{\boldsymbol{\psi}_2}) \le 8B\delta\Big(1 + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\Big) n_{\mathsf{s}}.$$

因此，

$$D_{\mathrm{KL}}(p^n_{\boldsymbol{\psi}_1} \,\|\, p^n_{\boldsymbol{\psi}_2}) \le 8B\delta \sum_{\mathsf{s}\in\boldsymbol{\mathcal{S}}} n_{\mathsf{s}}\Big(1 + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\Big).$$

因此，我们有

$$\inf_{\hat{\boldsymbol{\psi}}_n} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\|\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}(P)\|_2\right] \ge \frac{\delta}{4}\left(1 - \frac{8B\delta\sum_{\mathsf{s}\in\boldsymbol{\mathcal{S}}} n_{\mathsf{s}}\Big(1 + \sum_{\mathsf{v}\in\boldsymbol{\mathcal{S}}_{\setminus\mathsf{s}}} \gamma^{\mathsf{s},\mathsf{v}}\Big) + \log 2}{\log|\mathcal{V}|}\right)$$

$$\geq \frac{\delta}{4}\left(1-\frac{8B\delta\sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}}\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\gamma^{\mathsf{s},\mathsf{v}}\Big)+\log 2}{2mB\log(2\sqrt{m})}\right).$$

我们选择 $\delta = \frac{m}{\log(2\sqrt{m})16\sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}}\left(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\gamma^{\mathsf{s},\mathsf{v}}\right)}$，然后

$$1-\frac{8B\delta\sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}}\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\gamma^{\mathsf{s},\mathsf{v}}\Big)+\log 2}{2mB\log(2\sqrt{m})}\geq\frac{1}{4}.$$

因此，

$$\begin{aligned}\inf_{\hat{\boldsymbol{\psi}}_n}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\|\hat{\boldsymbol{\psi}}_n-\boldsymbol{\psi}(P)\|_2\right] &\geq \frac{1}{4}\times\frac{mB\log(2\sqrt{m})}{16B\sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}}\big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\gamma^{\mathsf{s},\mathsf{v}}\big)}\times\frac{1}{4}\\ &=\frac{m\log(2\sqrt{m})}{256\sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}}\big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\gamma^{\mathsf{s},\mathsf{v}}\big)}.\end{aligned}$$

这就完成了第 (i) 部分的证明。 □

H.2 第(ii)部分的证明

证明。我们将模型总结如下

$$y^{\mathsf{s}}=\Big((1-w^{\mathsf{s}})\big(\beta_0^{\mathsf{s}}+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}\beta_0^{\mathsf{v}}\big)+w^{\mathsf{s}}\big(\beta_1^{\mathsf{s}}+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}\beta_1^{\mathsf{v}}\big)\Big)^\top\phi(\mathbf{x}^{\mathsf{s}})+\epsilon_{\mathsf{s}},\qquad \epsilon_{\mathsf{s}}\sim\mathsf{N}(0,\sigma^2).$$

设 $\boldsymbol{\beta}=\{\beta_0^{\mathsf{s}},\beta_1^{\mathsf{s}}\}_{\mathsf{s}\in\mathcal{S}}$。设 $\mathcal{V}_{0\mathsf{s}}$ 和 $\mathcal{V}_{1\mathsf{s}}$ 是单位 $\|\cdot\|_2$-球的 $1/(2\sqrt{m})$-打包，其基数至少为 $(2\sqrt{m})^{2B}$。设 $\mathcal{V}_{\mathsf{s}}=\mathcal{V}_{0\mathsf{s}}\times\mathcal{V}_{1\mathsf{s}}$。现在我们选取一个集合 $\mathcal{V}=\delta(\mathcal{V}_{\mathsf{s}_1}\times\mathcal{V}_{\mathsf{s}_2}\times\ldots\times\mathcal{V}_{\mathsf{s}_m})$。我们可以看到

$$|\mathcal{V}|\geq(2\sqrt{m})^{4mB}.$$

我们有

$$\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_2=\sqrt{\sum_{\mathsf{s}\in\mathcal{S}}\Big(\|(\beta_0^{\mathsf{s}})_1-(\beta_0^{\mathsf{s}})_2\|_2^2+\|(\beta_1^{\mathsf{s}})_1-(\beta_1^{\mathsf{s}})_2\|_2^2\Big)}\geq\delta/\sqrt{2}.$$

此外，

$$D_{\mathrm{KL}}(p^n_{\boldsymbol{\beta}_1}\,\|\,p^n_{\boldsymbol{\beta}_2})=\sum_{\mathsf{s}\in\mathcal{S}}D_{\mathrm{KL}}(p^{n_{\mathsf{s}}}_{\boldsymbol{\beta}_1}\,\|\,p^{n_{\mathsf{s}}}_{\boldsymbol{\beta}_2})=\sum_{\mathsf{s}\in\mathcal{S}}n_{\mathsf{s}}D_{\mathrm{KL}}(p^1_{\boldsymbol{\beta}_1}\,\|\,p^1_{\boldsymbol{\beta}_2}).$$

此外，

$$\begin{aligned}&D_{\mathrm{KL}}(p^1_{\boldsymbol{\beta}_1}\,\|\,p^1_{\boldsymbol{\beta}_2})\\ &\quad=\frac{1}{2\sigma^2}\Bigg(\Big((1-w^{\mathsf{s}})\big((\beta_0^{\mathsf{s}})_1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}(\beta_0^{\mathsf{v}})_1\big)+w^{\mathsf{s}}\big((\beta_1^{\mathsf{s}})_1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}(\beta_1^{\mathsf{v}})_1\big)\Big)^\top\phi(\mathbf{x}^{\mathsf{s}})\\ &\qquad-\Big((1-w^{\mathsf{s}})\big((\beta_0^{\mathsf{s}})_2+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}(\beta_0^{\mathsf{v}})_2\big)+w^{\mathsf{s}}\big((\beta_1^{\mathsf{s}})_2+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}(\beta_1^{\mathsf{v}})_2\big)\Big)^\top\phi(\mathbf{x}^{\mathsf{s}})\Bigg)^2\\ &\quad\leq\frac{1}{2\sigma^2}\Bigg(\Big((1-w^{\mathsf{s}})\big(2\delta+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}2\delta\big)+w^{\mathsf{s}}\big(2\delta+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}2\delta\big)\Big)\|\phi(\mathbf{x}^{\mathsf{s}})\|_2\Bigg)^2\\ &\quad\leq\frac{8B^2\delta^2}{\sigma^2}\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}\Big)^2,\end{aligned}$$

因此，

$$D_{\mathrm{KL}}(p^n_{\boldsymbol{\beta}_1} \,\|\, p^n_{\boldsymbol{\beta}_2}) \le \frac{8B^2\delta^2}{\sigma^2} \sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}} \Big(1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \eta^{\mathsf{s},\mathsf{v}}\Big)^2.$$

因此，

$$\begin{aligned}\inf_{\hat{\boldsymbol{\beta}}_n} \sup_{P\in\mathcal{P}} \mathbb{E}_P \Big[\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}(P)\|_2\Big] &\ge \frac{\delta}{2\sqrt{2}} \left(1 - \frac{\frac{8B^2\delta^2}{\sigma^2} \sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}} \Big(1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \eta^{\mathsf{s},\mathsf{v}}\Big)^2 + \log 2}{\log|\mathcal{V}|}\right) \\ &\ge \frac{\delta}{2\sqrt{2}} \left(1 - \frac{\frac{8B^2\delta^2}{\sigma^2} \sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}} \Big(1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \eta^{\mathsf{s},\mathsf{v}}\Big)^2 + \log 2}{4mB\log(2\sqrt{m})}\right).\end{aligned}$$

我们选择 $\delta^2 = \frac{mB\log(2\sqrt{m})}{4\frac{B^2}{\sigma^2}\sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}} \Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}}\eta^{\mathsf{s}}_{\mathsf{v}}\Big)^2}$，那么

$$1 - \frac{\frac{8B^2\delta^2}{\sigma^2} \sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}} \Big(1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \eta^{\mathsf{s},\mathsf{v}}\Big)^2 + \log 2}{4mB\log(2\sqrt{m})} = 1 - \frac{2mB\log(2\sqrt{m}) + \log 2}{4mB\log(2\sqrt{m})} \ge \frac{1}{4}.$$

因此，

$$\begin{aligned}\inf_{\hat{\boldsymbol{\beta}}_n} \sup_{P\in\mathcal{P}} \mathbb{E}_P \Big[\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}(P)\|_2\Big] &\ge \frac{1}{2\sqrt{2}} \sqrt{\frac{4mB\log(2\sqrt{m})}{2\frac{8B^2}{\sigma^2}\sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}} \Big(1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \eta^{\mathsf{s},\mathsf{v}}\Big)^2}} \times \frac{1}{4} \\ &= \frac{\sigma}{16\sqrt{2}} \sqrt{\frac{m\log(2\sqrt{m})}{B\sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}} \Big(1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \eta^{\mathsf{s},\mathsf{v}}\Big)^2}}.\end{aligned}$$

至此，(ii) 部分的证明完成。 □

I 极小极大下界的其他情形

在引理 1 和引理 2 中，我们给出了 $y_i^{\mathsf{s}} \in \mathbb{R}$ 和 $\boldsymbol{x}_i^{\mathsf{s}} \in \mathbb{R}^{d_x}$ 时的极小极大下界。这里，我们简要描述其他情形。

I.1 引理 1 的其他情形

在本节中，我们将进一步详细说明二元结果和二元代理变量的下界。在这种情况下，我们需要重新推导

$$p_{\boldsymbol{\theta}_1}(w=j|\boldsymbol{z}) D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(y|w=j,\boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(y|w=j,\boldsymbol{z}')\big] \quad \text{and} \quad D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(\boldsymbol{x}|\boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(\boldsymbol{x}|\boldsymbol{z}')\big],$$

其中 $j = 1,2$。利用之前对量 $D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(w|\boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(w|\boldsymbol{z}')\big]$ 所采用的类似推导，我们有

$$p_{\boldsymbol{\theta}_1}(w=j|\boldsymbol{z}) D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(y|w=j,\boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(y|w=j,\boldsymbol{z}')\big] \le 8B\delta\Big(1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}\Big),$$

and

$$D_{\mathrm{KL}}\big[p_{\boldsymbol{\theta}_1}(\boldsymbol{x}|\boldsymbol{z}) \big\| p_{\boldsymbol{\theta}_2}(\boldsymbol{x}|\boldsymbol{z}')\big] \le d_x 8B\delta\Big(1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}\Big).$$

合并结果后，我们得到

$$D_{\mathrm{KL}}(p^n_{\boldsymbol{\theta}_1} \,\|\, p^n_{\boldsymbol{\theta}_2}) = \sum_{\mathsf{s}\in\mathcal{S}} D_{\mathrm{KL}}(p^{n_{\mathsf{s}}}_{\boldsymbol{\theta}_1} \,\|\, p^{n_{\mathsf{s}}}_{\boldsymbol{\theta}_2}) \le \sum_{\mathsf{s}\in\mathcal{S}} n_{\mathsf{s}} 8(d_x+3)B\delta\Big(1 + \sum_{\mathsf{v}\in\mathcal{S}_{\setminus \mathsf{s}}} \lambda^{\mathsf{s},\mathsf{v}}\Big).$$

因此，我们有

$$\inf_{\hat{\boldsymbol{\theta}}_n}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}(P)\|_2\right]\geq\frac{\delta}{2}\left(1-\frac{\sum_{\mathsf{s}\in\mathcal{S}}n_{\mathsf{s}}8(d_x+3)B\delta\left(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)+\log 2}{2mB(d_x+3)\log(2\sqrt{m})}\right).$$

We choose $\delta=\frac{m\log(2\sqrt{m})}{8\sum_{\mathsf{s}\in\mathcal{S}}n_{\mathsf{s}}\left(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)}$, then

$$1-\frac{\sum_{\mathsf{s}\in\mathcal{S}}n_{\mathsf{s}}8(d_x+3)B\delta\left(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)+\log 2}{2mB(d_x+3)\log(2\sqrt{m})}\geq\frac{3}{8}.$$

因此，

$$\boxed{\inf_{\hat{\boldsymbol{\theta}}_n}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}(P)\|_2\right]\geq\frac{3mB\log(2\sqrt{m})}{128\sum_{\mathsf{s}\in\mathcal{S}}n_{\mathsf{s}}B\left(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\lambda^{\mathsf{s},\mathsf{v}}\right)}.}$$

注 1. 请注意，本节和H.1 节中的推导为我们提供了足够的工具，可用于计算任何其他情形下的极小极大下界，即结果和代理变量（二元或连续）的任意组合。关键是首先根据所构造的打包，求出 $D_{KL}(p^n_{\boldsymbol{\theta}_1}\,\|\,p^n_{\boldsymbol{\theta}_2})$ 的上界。然后，利用 *Fano*方法得到极小极大下界。

I.2 引理 2 的其他情形

请注意，引理 2 第**(i)** 部分的下界只有一种情形，因为我们只关注二元处理，且该情形已在正文中给出。对于**(ii)**部分，考虑 $y_i^{\mathsf{s}}\in\{0,1\}$，则结果的模型将服从伯努利分布。沿用H.2节中的方案，我们需要求出 $D_{\mathrm{KL}}(p^n_{\boldsymbol{\beta}_1}\,\|\,p^n_{\boldsymbol{\beta}_2})$ 的新上界。具体而言，

$$D_{\mathrm{KL}}(p^n_{\boldsymbol{\beta}_1}\,\|\,p^n_{\boldsymbol{\beta}_2})=\sum_{\mathsf{s}\in\mathcal{S}}n_{\mathsf{s}}\left[\varphi(v_1)\log\frac{\varphi(v_1)}{\varphi(v_2)}+\varphi(-v_1)\log\frac{\varphi(-v_1)}{\varphi(-v_2)}\right],$$

其中 $v_j=\Big((1-w^{\mathsf{s}})\big((\beta_0^{\mathsf{s}})j+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\eta^{\mathsf{s},\,\mathsf{v}}(\beta_0^{\mathsf{v}})_j\big)+w^{\mathsf{s}}\big((\beta_1^{\mathsf{s}})j+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\eta^{\mathsf{s},\,\mathsf{v}}(\beta_1^{\mathsf{v}})_j\big)\Big)^\top\phi(\mathbf{x}^{\mathsf{s}})$。我们有

$$\begin{aligned}\varphi(v_1)\log\frac{\varphi(v_1)}{\varphi(v_2)}&\leq\Big\|(1-w^{\mathsf{s}})\big((\beta_0^{\mathsf{s}})_1-(\beta_0^{\mathsf{s}})_2+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}[(\beta_0^{\mathsf{v}})_1-(\beta_0^{\mathsf{v}})_2]\big)\\&\qquad+w^{\mathsf{s}}\big((\beta_1^{\mathsf{s}})_1-(\beta_1^{\mathsf{s}})_2+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}[(\beta_1^{\mathsf{v}})_1-(\beta_1^{\mathsf{v}})_2]\big)\Big\|_2\|\phi(\mathbf{x}^{\mathsf{s}})\|_2\\&\leq 4B\delta\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\gamma^{\mathsf{s},\mathsf{v}}\Big),\end{aligned}$$

同理，$\varphi(-v_1)\log\frac{\varphi(-v_1)}{\varphi(-v_2)}\leq 4B\delta\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\gamma^{\mathsf{s},\,\mathsf{v}}\Big)$。因此，

$$D_{\mathrm{KL}}(p^n_{\boldsymbol{\beta}_1}\,\|\,p^n_{\boldsymbol{\beta}_2})\leq 8B\delta\sum_{\mathsf{s}\in\mathcal{S}}n_{\mathsf{s}}\Big(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}\Big).$$

使用 H.2节中的类似技巧，我们得到

$$\boxed{\inf_{\hat{\boldsymbol{\beta}}_n}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}(P)\|_2\right]\geq\frac{m\log(2\sqrt{m})}{32\sqrt{2}\sum_{\mathsf{s}\in\mathcal{S}}n_{\mathsf{s}}\left(1+\sum_{\mathsf{v}\in\mathcal{S}_{\setminus\mathsf{s}}}\eta^{\mathsf{s},\mathsf{v}}\right)}.}$$

我们观察到，该下界与引理 2 的第**(i)** 部分相似，因为它们都是二元响应变量的下界。此下界中的常数大于引理 2 的第**(i)** 部分中的常数，具体为 $1/(32\sqrt{2})$ 和 $1/256$。这是意料之中的，因为与引理 2 的第**(i)** 部分中的模型（$\{\psi^{\mathsf{s}}\}_{\mathsf{s}\in\mathcal{S}}$）相比，该模型包含更多参数，即 $\{\beta_0^{\mathsf{s}},\beta_1^{\mathsf{s}}\}_{\mathsf{s}\in\mathcal{S}}$。

J IHDP 数据集描述

本节介绍 IHDP 数据的详细信息，由于篇幅有限，这些内容未在正文中展开。

婴儿健康与发展计划（IHDP）是一项随机研究，旨在研究专科医生就诊（处理）对儿童认知发展的影响（结果）。该数据集包含 747 条记录，涵盖 25 个描述儿童及其母亲特征的协变量。处理组包括接受过专科医生就诊的儿童，对照组则包括未接受此类就诊的儿童。更多详情见附录。我们使用 NPCI 软件包（Dorie 2016）提供的数值方案，为每名儿童模拟处理结果和对照结果，因此可以得知真实的 个体处理效应。在本实验中，我们使用该数据集的 10 个重复样本。对于每个重复样本，我们将数据分为三个数据源，每个数据源包含 249 个数据点。对于每个数据源，我们使用前 50 个数据点进行训练，接下来的 100 个用于测试，其余 99 个用于验证。我们报告了 10 个数据重复样本中各项评估指标的均值和标准误。

—— 结束 ——