arXiv:2305.02219v1 [cs.LG] 3 May 2023

LESS-VFL：面向纵向联邦学习的通信高效特征选择

Timothy Castiglia [1] **Yi Zhou** [2] **Shiqiang Wang** [2] **Swanand Kadhe** [2] **Nathalie Baracaldo** [2] **Stacy Patterson** [1]

摘要

我们提出了 LESS-VFL，这是一种面向具有纵向切分数据的分布式系统的低通信成本特征选择方法。我们考虑由一个服务器和多个参与方组成的系统，这些参与方拥有本地数据集，共享样本 ID 空间，但具有不同的特征集。参与方希望协作训练一个用于预测任务的模型。作为模型训练的一部分，参与方希望移除系统中的不重要特征，以提升泛化、效率和可解释性。在 LESS-VFL 中，经过一段较短的预训练阶段后，服务器优化全局模型中属于自身的部分，以确定参与方模型中的相关输出。随后，服务器将这些信息共享给参与方，使其能够在无需通信的情况下进行本地特征选择。我们通过解析证明，LESS-VFL 能够从模型训练中移除伪特征。大量实验证据表明，LESS-VFL 能够以其他特征选择方法一小部分的通信成本，实现高准确率并移除伪特征。

1. 引言

联邦学习近年来逐渐受到研究界的关注，并已在多个应用领域展现出良好前景，例如医疗保健、智能交通和预测能源系统（Sun 等，2019年；Kairouz 等，2021年；Zhou 等，2021年）。联邦学习算法支持参与方之间进行分布式模型训练，而无需直接共享本地私有数据。

纵向联邦学习（VFL）是联邦学习算法中的一个重要类别，近来受到了广泛关注（Yang 等人，2019年；Cha 等人，2021年；Castiglia 等人，2022年）。纵向联邦学习的研究考虑的是以下情况

[1]伦斯勒理工学院 [2]IBM Research。通讯作者：Timothy Castiglia <castigliatim@gmail.com>。

《 39^{th} 国际机器学习会议》，美国夏威夷檀香山。PMLR 202，2023年。版权所有 2023年，归作者所有。

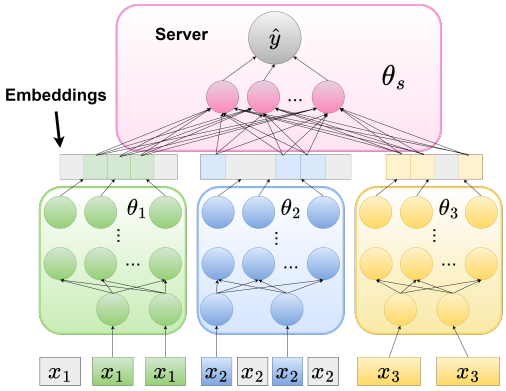


图*1*。VFL 模型架构示例。使用 LESS-VFL 训练后，将不显著的特征和嵌入分量（以灰色表示）移除。

参与方拥有共同的样本 ID 空间，但各自持有不同的私有特征集。例如，医疗服务提供商、可穿戴技术公司和保险公司希望协同训练一个疾病预测模型。这些参与方掌握着同一批个人的信息（样本 ID 空间），但每个参与方存储的健康信息（特征空间）各不相同。在纵向联邦学习中，参与方通常使用本地特征提取模型（例如深度神经网络）来生成本地特征集的低维 嵌入（Hu 等人，2019年；Ceballos 等人，2020年）。服务器将嵌入作为输入传入融合模型进行预测。我们在图1中给出了一个纵向联邦学习模型示例。

特征选择是机器学习任务的重要组成部分。数据集往往包含许多与当前预测任务无关的伪特征。例如，医疗保健提供者可能会使用电子病历（EMR）训练模型，其中包含临床文档、常规就诊的结果，以及许多可能与疾病诊断无关的特征（Canino 等人，2016年）。未能移除伪特征会对泛化产生严重影响。在图2中，我们比较了使用原始数据集进行垂直联邦学习训练与使用该数据集及额外一组高斯噪声特征进行训练时的测试准确率。仅仅加入这些伪特征，就会导致纵向联邦学习的测试准确率大幅下降。除了改进模型

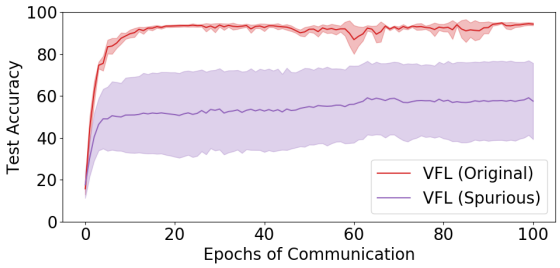


图 2。 Activity数据集上的测试准确率（详情见第 5 节）。VFL（原始）表示使用原始数据集运行算法 2 所得的测试准确率。VFL（伪特征）表示在添加了伪特征的数据集上得到的测试准确率。实线表示5 次运行的平均值，阴影区域表示标准差。

泛化、特征选择常用于模型可解释性(Clinciu& Hastie,2019年)。

大多数集中式特征选择算法无法直接应用于纵向联邦学习，因为它们要么需要直接共享数据，要么通信效率低下。纵向联邦学习中的参与方可能分布在全球各地，因此通信在时间、金钱和资源方面的成本都很高。因此，设计通信高效的纵向联邦学习算法十分重要。尽管已有少数研究提出了纵向联邦学习特征选择算法（总结见表 1），但尚无研究对纵向联邦学习中的特征选择问题进行形式化分析，并提出一种能够以理论保证去除伪特征的方法。在本研究中，我们试图回答以下问题：我们能否设计一种通信高效的纵向联邦学习特征选择算法，并通过形式化验证证明它可以去除伪特征并实现高准确率？

相关工作。 特征选择算法通常可分为三大类：过滤法、包装法和嵌入法（Chandrashekar & Sahin， 2014年）。过滤法利用数据的统计指标，在模型训练之前预先确定特征重要性。这些方法需要直接访问特征以计算相关指标，因此在不共享原始数据的情况下，无法直接应用于纵向联邦学习（VFL）场景。包装法通常涉及多次重新训练模型，以确定不同特征子集的重要性。这对于纵向联邦学习场景并不实际，因为模型训练需要参与方与服务器之间进行大量通信。嵌入法在训练模型的同时确定所有特征的重要性。这些方法可以在执行特征选择之前完整地训练模型，也可以在训练过程中逐步移除不重要的特征。在训练过程中移除不重要特征的嵌入法似乎非常适合纵向联邦学习，但必须对其进行调整，以支持分布式训练并降低通信开销。

已有少数研究提出了嵌入式纵向联邦学习特征选择方法（Feng &Yu，2020年；Chen 等人，2021年；Hou 等人，2022年；Zhang 等人，2022a；b；Chen 等人，2022年；Feng，2022年；Li 等人，2023年）。然而，这些方法大多不支持深度神经网络，或要求模型经过充分训练后才能开始进行特征选择（见表1）。更为关键的是，这些研究均未提供理论证据证明其提出的方法能够移除伪特征，仅提供了实证证据。一个重要的开放问题是，如何在纵向联邦学习场景中形式化特征选择问题，并为证明不重要的特征会被移除提供理论框架。

表1. 纵向联邦学习特征选择算法。

纵向联邦学习特征选择算法	支持神经网络。网络、】【	已选特征训练期间	可证明地移除伪特征
MMVFL（Feng& Yu，2020年）	✗	✓	✗
Fed-EINI（Chen 等人，2021年）	✗	✓	✗
Hou 等人（2022年）	✗	✓	✗
Zhang 等人(2022年b)	✗	✓	✗
Zhang 等人(2022年a)	✓	✗	✗
EVFL（Chen 等人，2022年）	✓	✗	✗
VFLFS（Feng，2022年）	✓	✓	✗
FedSDG-FS（Li 等人，2023年）	✓	✓	✗
LESS-VFL（我们的方法）	✓	✓	✓

贡献。 在这项工作中，我们对纵向联邦学习中的特征选择问题进行了形式化，并提出了纵向联邦学习的本地通信高效组套索方法（LESS-VFL）。这是一种适用于纵向联邦学习的嵌入式特征选择方法，能够以通信高效的方式可靠地去除伪特征。我们以一种新颖的方式利用组套索正则化（Zhao 等人，2015；Zhang 等人，2020年；Wang 等人，2021年），从而减少参与方之间的通信量。经过短暂的预训练后，服务器会从每个参与方确定一组“重要”的嵌入分量。利用这些信息，每个参与方都可以在本地使用组套索进行特征选择，无需通信。尽管已有研究证明，集中式的组套索实现能够去除伪特征（Dinh 和 Ho，2020年），但我们的方法能否在纵向联邦学习中提供类似保证尚不明确。我们的分析证明，在给定重要嵌入分量的情况下，随着样本量增大，各参与方渐近地解决特征选择问题。在实验中，我们将 LESS-VFL 与直接在纵向联邦学习中应用组套索正则化的方法进行比较，发现 LESS-VFL 能大幅降低特征选择的通信成本。

我们总结如下贡献：

- 我们在第 2 节对纵向联邦学习中的特征选择问题进行了形式化。
- 我们提出了一种三阶段方法 LESS-VFL，并在第 3 节介绍了其实用实现。
- 我们在第4节通过解析证明，LESS-VFL 能够移除伪特征并实现高准确率。

• 我们在第5节提供了实验证据，表明 LESS-VFL 能够以较低的通信成本实现高准确率。

2. 问题定义与背景

我们考虑一个包含 M 个参与方和一个服务器的系统。每个参与方 m 为 N 个训练样本存储 d_m 个特征。我们将参与方 m的数据集记为$\mathbf{X}_m \in \mathbb{R}^{N\times d_m}$。我们将$\mathbf{X}_m$ 中的第 i个样本记为$\mathbf{x}_m^{(i)}$。我们假设每个参与方的数据集均已对齐，即对于所有参与方$m \neq j$， $\mathbf{x}_m^{(i)}$和$\mathbf{x}_j^{(i)}$ 是数据样本 i的不同特征。我们将组合数据样本记为$\mathbf{x}^{(i)} = [\mathbf{x}_1^{(i)}, .., \mathbf{x}_M^{(i)}]$，并将 $\mathcal{X}$记为样本$\mathbf{x}^{(i)}$所有可能取值的集合。我们将组合数据集记为$\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_M]$。我们假设服务器存储训练标签$\mathbf{y}$。请注意，如果某个参与方存储了标签，则该参与方也可以充当服务器。

每个参与方使用参数 $\boldsymbol{\theta}_m$训练本地模型$\mathbf{h}_m(\cdot)$，服务器使用参数$\boldsymbol{\theta}_s$训练服务器模型$\mathbf{h}_s(\cdot)$。参与方模型的输出称为嵌入。所有参与方嵌入均作为输入传入服务器融合模型$\mathbf{h}_s(\cdot)$。完整的纵向联邦学习模型 $f(\cdot)$定义如下：

$$f(\boldsymbol{\Theta}; \mathbf{x}^{(i)}) := \mathbf{h}_s(\boldsymbol{\theta}_s, \mathbf{h}_1(\boldsymbol{\theta}_1; \mathbf{x}_1^{(i)}), \ldots, \mathbf{h}_M(\boldsymbol{\theta}_M; \mathbf{x}_M^{(i)})).$$

where $\boldsymbol{\Theta} = [\boldsymbol{\theta}_s^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_M^\top]^\top$.

我们将特征选择问题形式化。回顾第 1 节可知，并非所有输入特征都可能对当前预测任务有显著作用。对于给定的任意输入 $\mathbf{w}$ 和模型 $u(\boldsymbol{\theta})$，我们将显著性这一概念形式化，并将 Dinh 和 Ho Dinh 和 Ho (2020年) 中的定义 2.2 扩展到纵向联邦学习系统。

定义 2.1. 对于给定的模型 u 及其参数 $\boldsymbol{\theta}$，设 $\mathbf{w}^j$ 为模型的第 j 个输入，模型表示为 $u(\boldsymbol{\theta};\mathbf{w})$； $g^j(\mathbf{w}\ s)$ 是一个将$\mathbf{w}^j$ 替换为值 s 的函数。若对于所有 $s \in \mathbb{R}$，都有 $u(\boldsymbol{\theta};\mathbf{w}) = u(\boldsymbol{\theta}; g^j(\mathbf{w}\ s))$，则输入$\mathbf{w}^j$ 在该模型 $u(\boldsymbol{\theta};\mathbf{w})$ 中是非显著的。否则，$\mathbf{w}^j$ 是显著的。

我们想强调，定义 2.1 中，显著输入的集合取决于模型参数 $\boldsymbol{\theta}$。本文中，我们将定义 2.1 中的显著性概念具体应用于以下两种情形：

1. 当输入为训练特征集且模型为分类器时，即定义 2.1 中的 $\mathbf{w} = \mathbf{x}$，$u(\cdot) = f(\cdot)$ 和$\boldsymbol{\theta} = \boldsymbol{\Theta}$； 2. 当输入为基于参与方本地模型生成的嵌入分量集且模型为服务器模型时，即 $\mathbf{w} = [\mathbf{h}_1(\cdot); \ldots; \mathbf{h}_M(\cdot)]$, $u(\cdot) = \mathbf{h}_s(\cdot)$ 和 $\boldsymbol{\theta} = \boldsymbol{\theta}_s$。

我们将 $\boldsymbol{\Theta}^\diamond = [(\boldsymbol{\theta}_s^\diamond)^\top, (\boldsymbol{\theta}_1^\diamond)^\top, \ldots, (\boldsymbol{\theta}_M^\diamond)^\top]^\top$ 定义为生成训练标签的生成模型：$y^{(i)} = f(\boldsymbol{\Theta}^\diamond; \mathbf{x}^{(i)}) + \epsilon^{(i)}$，其中 $\mathbf{x}^{(i)}$ 取自分布 $\mathcal{P}_{\mathbf{x},y}$，且 $\epsilon \sim \mathcal{N}(0, \sigma^2)$（形式化定义见第 4 节）。我们进行特征选择的主要目标是确定属于 $\boldsymbol{\Theta}^\diamond$ 的显著特征集合。我们将参与方 m 的生成模型 $\boldsymbol{\theta}_m^\diamond$ 的显著特征集合记为 $\mathbf{s}_m$，将非显著特征集合记为 $\mathbf{z}_m$，其中任意数据样本$\mathbf{x}_m$均适用。我们可以考虑与显著特征和非显著特征对应的输入层权重：

$$\mathbf{x}_m = (\mathbf{s}_m, \mathbf{z}_m) \text{ and } \pi(\boldsymbol{\theta}_m^\diamond) = (\mathbf{U}_m, \mathbf{V}_m)$$

其中， $\pi(\boldsymbol{\theta})$用于提取模型 $\boldsymbol{\theta}$中的输入权重，$\mathbf{U}_m$ 和$\mathbf{V}_m$ 分别是 $\boldsymbol{\theta}_m^\diamond$中显著特征和非显著特征对应的输入层权重。请注意，在训练期间，$\mathbf{U}_m$ 和$\mathbf{V}_m$ 之间的划分并不已知，这只是为了数学表达上的便利。

嵌入式纵向联邦学习特征选择的目标是找到一个模型，使其预测结果与生成模型 $\boldsymbol{\Theta}^\diamond$ 相似，同时将非显著特征权重置为零：

$$\min_{\boldsymbol{\Theta}} R(\boldsymbol{\Theta}; \mathcal{P}_{\mathbf{X},\mathbf{y}}) \text{ s.t. } \mathbf{U}_m^k \neq \mathbf{0} \ \ \forall m \in [M],\ \forall k \in [d_m^s] \quad \mathbf{V}_m^l = \mathbf{0} \ \ \forall m \in [M],\ \forall l \in [d_m^z] \tag{1}$$

其中 $R(\cdot)$ 是数据分布 $\mathcal{P}_{\mathbf{X}}$ 上的某种泛化风险，$\mathbf{y}$（例如，期望平方损失、交叉熵），而d_m^s 和 d_m^z 分别是参与方 m 的显著特征和非显著特征的数量。将非显著特征的输入权重设为零，可消除它们对网络的影响，因此实质上是从模型中移除这些特征（如图 1 所示）。注意，$\mathbf{V}_m$ 在 $\boldsymbol{\Theta}^\diamond$中不一定为零，因为非显著特征的影响可能在模型的任意一层消失 $f(\boldsymbol{\Theta}^\diamond)$。

一种流行的用于求解（1）中神经网络的集中式方法是组套索（Zhao 等人，2015年；Zhang 等人，2020年；Wang 等人，2021年）。如果我们将组套索直接应用于纵向联邦学习场景，则可以将估计量定义如下：

$$\bar{\boldsymbol{\Theta}} := \arg\min_{\boldsymbol{\Theta}} R_N(\boldsymbol{\Theta}; \mathbf{X}; \mathbf{y}) + \textstyle\sum_{m=1}^M \lambda_m G(\boldsymbol{\theta}_m) \tag{2}$$

其中， $R_N(\cdot)$ 是在 N 训练样本上近似 $R(\cdot)$ 的经验风险，而 $G(\cdot)$ 是 $L_{2,\ 1}$ 正则化项：

$$G(\boldsymbol{\theta}_m) := \textstyle\sum_{j=1}^{d_m} \|\pi(\boldsymbol{\theta}_m)^j\|_2$$

其中，投影 $\pi(\cdot)$提取 $\boldsymbol{\theta}_m$ 的输入层权重， d_m 是输入特征的数量。正则化器 $G(\cdot)$对每个特征对应的输入层权重进行稀疏化，将无关特征的权重推至零。

为什么不采用标准组套索? 使用标准纵向联邦学习训练最小化组套索目标函数（2）（Hu 等人，2019年；Ceballos 等人，2020年）需要参与方和服务器在每次训练迭代中交换嵌入和偏导数（见算法2）。能否不必在每次迭代中都传输嵌入，而是利用服务器提供的辅助信息，在每个客户端本地进行特征

算法1使用 P-SGD 实现的 LESS-VFL

输入：预训练模型参数 $\hat{\boldsymbol{\theta}}_s, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_M$
对于 $m \leftarrow 1, \ldots, M$，并行**执行**
  发送 $\mathbf{h}_m(\hat{\boldsymbol{\theta}}_m; \mathbf{X}_m)$ 到服务器
结束循环
初始化：$\boldsymbol{\theta}_s^0 \leftarrow \hat{\boldsymbol{\theta}}_s$
对于 $t \leftarrow 0, \ldots, T_1 - 1$ **执行**
  随机采样 $\mathcal{B} \subset [N]$
  $\hat{\Phi} \leftarrow \{\boldsymbol{\theta}_s^t, \mathbf{h}_1(\hat{\boldsymbol{\theta}}_1; \mathbf{X}_m^{(\mathcal{B})}) \ldots, \mathbf{h}_M(\hat{\boldsymbol{\theta}}_M; \mathbf{X}_m^{(\mathcal{B})})\}$
  $\boldsymbol{\theta}_s^{t+1} \leftarrow \text{prox}_{\lambda_s, \eta_s^t}\left(\boldsymbol{\theta}_s^t - \eta_s^t \nabla_s R_{\mathcal{B}}(\hat{\Phi}; \mathbf{y}^{(\mathcal{B})})\right)$
结束循环
for $m \leftarrow 1, \ldots, M$ 并行执行 **do**
  $\mathcal{K}_m = \{k \mid \|\pi(\boldsymbol{\theta}_s^{T_1})^k\|_2 > 0\}$
  初始化：$\boldsymbol{\theta}_m^0 \leftarrow \hat{\boldsymbol{\theta}}_m$
结束循环
循环 $t \leftarrow 0, \ldots, T_2 - 1$ **执行**
  对于 $m \leftarrow 1, \ldots, M$ 并行**执行**
    随机采样 $\mathcal{B} \subset [N]$
    $\boldsymbol{\theta}_m^{t+1} = \text{近端}_{\lambda_m, \eta_m^t}\left(\boldsymbol{\theta}_m^t - \eta_m^t \nabla H_{\mathcal{B}}(\boldsymbol{\theta}_m^t; \hat{\boldsymbol{\theta}}_m; \mathcal{K}_m)\right)$
  结束循环
结束循环
$\bar{\boldsymbol{\Theta}} \leftarrow [\boldsymbol{\theta}_s^{T_1}, \boldsymbol{\theta}_1^{T_2}, \ldots, \boldsymbol{\theta}_M^{T_2}]$

选择？在下一节中，我们提出一种特征选择方法来求解（1），该方法通过参与方与服务器之间的最少通信，利用本地训练。

3. 算法

我们提出了 LESS-VFL，这是一种通信高效的方法，可在纵向联邦学习中进行特征选择。我们对 LESS-VFL 的三个阶段进行了形式化描述，并给出了一种实用的实现方案。

阶段1 – 预训练。参与方和服务器首先共同求解以下经验风险最小化问题：

$$\hat{\boldsymbol{\Theta}} := \arg\min_{\boldsymbol{\Theta}} R_N(\boldsymbol{\Theta}; \mathbf{X}; \mathbf{y}). \tag{3}$$

算法2 附录 A (Hu 等人，2019年；Ceballos 等人，2020年) 中描述的标准纵向联邦学习训练，是求解（3）近似解的一种实用方法。

阶段2——嵌入分量选择。在此阶段，服务器确定重要嵌入分量的集合。每个参与方针对每个样本 i 发送其当前预训练嵌入 $\mathbf{h}_m(\hat{\boldsymbol{\theta}}_m; \mathbf{x}^{(i)})$。这些嵌入是固定的，并在此阶段作为服务器模型的输入。稍微滥用一下符号表示法，我们令 $\hat{} \hat{} R_N(\boldsymbol{\theta}_s, \mathbf{h}_1(\boldsymbol{\theta}_1) \ldots, \mathbf{h}_M(\boldsymbol{\theta}_M))$ 表示服务器模型使用预训练嵌入作为输入时的经验风险。The 服务器解决以下问题：

$$\bar{\boldsymbol{\theta}}_s := \arg\min_{\boldsymbol{\theta}_s} R_N(\boldsymbol{\theta}_s, \mathbf{h}_1(\hat{\boldsymbol{\theta}}_1), \ldots, \mathbf{h}_M(\hat{\boldsymbol{\theta}}_M)) + \lambda_s G(\boldsymbol{\theta}_s) \tag{4}$$

其中，$G(\cdot)$ 是 $L_{2,1}$ 是 $\boldsymbol{\theta}_s$ 输入层上的正则化项，而 $\boldsymbol{\theta}_m$ 是参与方 m 预训练后的预训练参数。请注意，服务器将预训练嵌入作为输入，无需与参与方通信即可计算 $R_N(\cdot)$。求解（4）可在最小化风险的同时，使嵌入分量权重稀疏化。图1对此进行了说明，其中不显著的嵌入分量（灰色部分）不再向服务器模型提供输入。

在算法1 （第 1–10 行）中，我们提供了一种实用方法，用于求解式（4）的近似解。参与方使用预训练模型为所有数据样本生成嵌入，并将其发送至服务器（第 1–4 行）。随后，服务器开始执行嵌入分量选择（第 5–10 行）。服务器随机采样一个索引小批量 $\mathcal{B}$，然后计算风险相对于服务器模型的偏导数：$\nabla_s R_{\mathcal{B}}(\cdot)$。接着，服务器采用近端随机梯度下降（P-SGD）。我们用 $\text{prox}_{\lambda,\eta}(\boldsymbol{\theta})$ 表示参数为 λ、步长为 η 时，针对 $L_{2,1}$ 正则化项的近端算子的闭式解：

$$\text{prox}_{\lambda,\eta}(\mathbf{P}^j) = \begin{cases} \mathbf{P}^j - \lambda\eta \frac{\mathbf{P}^j}{\|\mathbf{P}^j\|_2} & \|\mathbf{P}^j\|_2 > \lambda\eta \\ \mathbf{0} & \|\mathbf{P}^j\|_2 \leq \lambda\eta \end{cases}$$

其中，$\mathbf{P}^j := \pi(\boldsymbol{\theta})^j$ 是第 j 组输入权重。训练完成后，任何输入权重非零的嵌入分量都被视为显著分量，并将显著嵌入分量的索引列表 $\mathcal{K}_m$ 发送给每个参与方 m（第 11–14 行）。

阶段3 – 特征选择。在此阶段，每个参与方的目标是匹配显著嵌入分量的值，同时从其模型中移除非显著特征。我们将参与方的嵌入值与预训练嵌入值在显著分量集合 $\mathcal{K}_m$ 上的平方差表示为：

$$e(\boldsymbol{\theta}_m; \hat{\boldsymbol{\theta}}_m; \mathcal{K}_m; \mathbf{x}_m^{(i)}) := \sum_{k \in \mathcal{K}_m} (\mathbf{h}_m(\boldsymbol{\theta}_m; \mathbf{x}_m^{(i)})^k - \mathbf{h}_m(\hat{\boldsymbol{\theta}}_m; \mathbf{x}_m^{(i)})^k)^2$$

其中，$\mathbf{h}_m(\boldsymbol{\theta}_m; \mathbf{x}_m^{(i)})^k$ 是第 k 个嵌入分量。每个参与方在对其输入层权重进行稀疏化的同时，针对每个数据样本最小化 $e(\cdot)$：

$$\bar{\boldsymbol{\theta}}_m := \arg\min_{\boldsymbol{\theta}_m} H_N(\boldsymbol{\theta}_m; \hat{\boldsymbol{\theta}}_m; \mathcal{K}_m; \mathbf{X}_m) + \lambda_m G(\boldsymbol{\theta}_m) \tag{5}$$

其中，

$$H_N(\boldsymbol{\theta}_m; \hat{\boldsymbol{\theta}}_m; \mathcal{K}_m; \mathbf{X}_m) := \frac{1}{N} \sum_{i=1}^N e(\boldsymbol{\theta}_m; \hat{\boldsymbol{\theta}}_m; \mathcal{K}_m; \mathbf{x}_m^{(i)}).$$

求解（5）的近似解的一种实用方法见算法 1 （第 15–20 行）。每个参与方都使用小批量梯度 $\nabla H_{\mathcal{B}}(\cdot)$ 更新其模型，并应用 $\text{prox}_{\lambda_m,\ \eta_m^t}(\boldsymbol{\theta})$，其中正则化参数为 λ_m。

最小化（5）后，所有置零的输入特征权重都被视为非显著特征，并从模型中移除。如图1所示，非显著特征（灰色）的输入权重会被移除。完成特征选择后，如有需要，可以使用算法2进一步利用剩余特征优化网络。

算法成本。LESS-VFL 的阶段1与标准纵向联邦学习相同，因此每次迭代具有相同的通信成本。阶段2和阶段3仅需要一轮通信，即参与方将当前嵌入发送给服务器。算法 T_1 和 T_2 中的迭代次数1分别控制服务器和参与方的计算成本，这两个参数均可进行调整。

隐私。 LESS-VFL 使用纵向联邦学习训练期间已共享的信息来进行特征选择。因此，它提供与标准纵向联邦学习相同的隐私保障。尽管参与方之间不会共享原始数据，但纵向联邦学习仍可能受到重构攻击和标签泄露的威胁。已有一些技术被应用于纵向联邦学习，以防御此类攻击（Qiu 等人，2022年；Zou 等人，2022年），这些技术同样可以应用于 LESS-VFL。在采用这些方法时，我们在第 4 节中的分析仍然成立。

理论与实践。 我们注意到，算法 1 和 2在固定迭代次数下运行时，会为各阶段的优化问题提供近似解。通过使用 P-SGD，即使尚未收敛、所选特征集合并非最优，也能将输入权重设为零并完成特征选择。第 4节对 LESS-VFL 的分析考虑了（3）、（4）和（5）均得到精确求解的理想情况。然而，我们在第 5节的实验中发现，LESS-VFL 仅需少量用于预训练的通信轮次，并通过算法1得到近似解，就能移除伪特征并达到高准确率。

超参数调优。 可以高效地确定 LESS-VFL 的最佳超参数（例如正则化参数、预训练轮数）。参与方和服务器可以针对阶段1，使用不同的迭代次数生成多个预训练模型。在阶段2中，服务器随后可以在无需通信的情况下探索服务器模型正则化参数的空间 λ_s。对于阶段3，服务器可以与参与方共享由此得到的重要嵌入分量索引集合，之后每个参与方都可以在无需通信的情况下调优其本地正则化参数 λ_m。为选择阶段2中的迭代次数 T_1，服务器可以在本地最小化其优化问题，直到训练损失趋于平稳。同样，为选择阶段3中的 T_2，每个参与方可以将其本地特征选择问题最小化，直至其代理训练损失趋于平稳。

4. 理论分析

我们分析 LESS-VFL，并证明该算法在何种条件下能够最小化风险并移除伪特征。我们假设每个参与方 m 的网络结构如下：

- 输入层： $\boldsymbol{\ell}_m^1(\mathbf{x}_m) = \mathbf{P}_m \cdot \mathbf{x}_m + \mathbf{p}_m$
- 隐藏层： $\boldsymbol{\ell}_m^j(\mathbf{x}_m) = \zeta_m^j(\mathbf{S}_m^j, \boldsymbol{\ell}_m^{j-1}(\mathbf{x}_m) \ldots, \boldsymbol{\ell}_m^1(\mathbf{x}_m))$
- 输出层： $\mathbf{h}_m(\boldsymbol{\theta}_m; \mathbf{x}_m) = \mathbf{Q}_m \cdot \boldsymbol{\ell}_m^{L-1}(\mathbf{x}_m) + \mathbf{q}_m$

其中， d_m^j 是参与方 m的第 j个隐藏层中的神经元数量 m，$\mathbf{P}_m \in \mathbb{R}^{d_m^1 \times d_m}$，$\mathbf{Q}_m \in \mathbb{R}^{d_m^L \times d_m^{L-1}}$，$\mathbf{p}_m \in \mathbb{R}^{d_m^1}$，$\mathbf{q}_m \in \mathbb{R}^{d_m^L}$ 以及 $\zeta_m^j(\cdot)$ 是由参数$\mathbf{S}_m^j$决定的前一层函数。我们将服务器网络结构定义为相同的形式，并用下标 s表示。这种结构适用于多种类型的神经网络，包括前馈网络、卷积网络以及许多残差网络（Dinh 和 Ho，2020年）。

我们作出以下假设，这些假设是基于模型的特征选择中的标准假设（Huang 等人，2010年；吴和 Liu，2009；Dinh 和 Ho，2020年）：

假设 4.1。训练数据 $\{(\mathbf{x}^{(i)},\ y^{(i)})\}_{i=1}^N$ 是从分布 $\mathcal{P}_{\mathbf{X},\ \mathbf{y}}$ 中独立同分布抽取的样本，并且输入密度 $p_{\mathbf{X}}$ 在其开域 $\mathcal{X}$ 上为正且连续，且$y^{(i)} = f(\boldsymbol{\Theta}^\diamond;\ \mathbf{x}^{(i)}) + \epsilon^{(i)}$，其中 $\epsilon \sim \mathcal{N}(0,\ \sigma^2)$。

假设4.1指出，存在一个生成模型f（$\boldsymbol{\Theta}^\diamond$），该模型在一定的高斯噪声下生成训练标签。这确保了特征选择是可行的[-]，因为所学习的模型 f（$\boldsymbol{\Theta}$）与生成模型的结构相匹配。关于输入密度的假设 $p_{\mathbf{X}}$确保输入特征之间不存在完全相关。注意，由于 $\mathcal{X}$是一个开域，我们假设在此解析中所有潜在特征都是连续的。

假设 4.2。所有模型中的隐藏层函数 $\zeta_m^j(\cdot)$都是解析的。经验风险是均方误差： $R_N(\boldsymbol{\Theta};\ \mathbf{X};\ \mathbf{y}) := \frac{1}{N}\sum_{i=1}^N (f(\boldsymbol{\Theta};\ \mathbf{x}^{(i)}) - y^{(i)})^2$，而泛化风险函数是期望平方误差：$R(\boldsymbol{\Theta};\ \mathcal{P}_{\mathbf{X},\ \mathbf{y}}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_{\mathbf{X},\ \mathbf{y}}}[(f(\boldsymbol{\Theta};\ \mathbf{x}) - y)^2]$.

假设 4.2 确保风险函数是解析的，从而我们可以分析所学习模型与生成模型之间的距离。请注意，根据假设 4.1，$\boldsymbol{\Theta}^\diamond$ 最小化期望平方误差 $R(\cdot)$。

[-]接下来，我们将形式化定义目标，即寻找参数 $\boldsymbol{\Theta}$ 使其能够求解（1），也就是说，在移除非显著特征的同时，实现与生成模型 $\boldsymbol{\Theta}^\diamond$ 相同的效果。我们将集合 $\mathcal{T}^*$ 定义为能够达到与生成模型相同风险的参数集合

模型：

$$\mathcal{T}^* := \{\boldsymbol{\Theta} : R(\boldsymbol{\Theta}) = R(\boldsymbol{\Theta}^\diamond)\}.$$

回顾第2 节可知，对于 $\boldsymbol{\Theta}^\diamond$，非显著特征上的输入权重不一定为零。对于 $\mathcal{T}^*$中的任意模型，情况也相同。因此，我们将 $\mathcal{T}^*$ 中的一组参数定义为：其中非显著特征上的权重也设为零：

$$\mathcal{T}_\phi^* := \{\boldsymbol{\Theta} : \boldsymbol{\Theta} \in \mathcal{T}^* \text{ and } \mathbf{V}_m = \mathbf{0}\},$$

其中，$\mathbf{V}_m$ 表示生成模型 $\boldsymbol{\Theta}^\diamond$ （定义见第2节）中非显著特征上的输入权重。我们将向量 $\boldsymbol{\theta}$ 到向量集合 $\mathcal{S}$ 的距离定义为：

$$d(\boldsymbol{\theta}, \mathcal{S}) = \inf_{\boldsymbol{\theta}' \in \mathcal{S}} \| \boldsymbol{\theta} - \boldsymbol{\theta}' \|_2.$$

¯若对于我们学习得到的参数 $\boldsymbol{\Theta}$，特征选择问题（1）得到解决，当且仅当 d（$\boldsymbol{\Theta}$，$\mathcal{T}_\phi^*$）$\to 0$¯。Dinh&Ho（2020年）证明，在集中式机器学习问题中，可以通过组套索实现这一点。我们的目标是证明，我们的三阶段方法在纵向联邦学习场景中也能达到相同效果。

4.1. 主要结果

下面给出我们的主要结果。

定理4.3。 令 $\tilde{\boldsymbol{\Theta}} = \text{argmin}_{\boldsymbol{\Theta} \in \mathcal{T}^*} \|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}\|_2$。令$\mathcal{K}_m$ 和 $\mathcal{Z}_m$ 分别为模型 m’中参与方的显著和非˜显著嵌入分量 $\boldsymbol{\Theta}$，分别。¯令 $\boldsymbol{\Theta}$ 为依次求解 (3)、(4)和 (5) 后得到的模型。在假设 *4.1*和*4.2*下，可得：

(i) 求解 (4) 并得到 $\bar{\boldsymbol{\theta}}_s$后，对于所有 m，$\pi(\bar{\boldsymbol{\theta}}_s)^k \neq \mathbf{0}$对于所有 $k \in \mathcal{K}_m$ 且 $\pi(\bar{\boldsymbol{\theta}}_s)^l \to \mathbf{0}$ 对于所有 $l \in \mathcal{Z}_m$，其中$\pi(\boldsymbol{\theta}_s)$为嵌入层权重。

(ii) 当服务器对所有 m 找到 $\mathcal{K}_m$ 时，对于任意 $\delta > 0$ 和 $\nu > 1$，若 $\lambda_s \sim N^{-1/4}$ 且对于所有 m 均有 $\lambda_m \sim N^{-1/4}$，

$$d(\bar{\boldsymbol{\Theta}}, \mathcal{T}_\phi^*) = O\left(\sqrt{M}\left(\frac{\log N}{N}\right)^{\frac{1}{4(\nu-1)}}\right) \tag{6}$$

以概率 $1 - \delta$。

定理 4.3 的(i)指出，服务器找到了一组˜在 $\boldsymbol{\Theta}$中具有显著性的嵌入分量，即与生成模型的风险相匹配的最接近预训练模型ˆ。 $\boldsymbol{\Theta}$ 该结果确保，服务器提供给各参与方的嵌入分量列表可以作为损失函数的准确代理。

定理4.3 (ii) 指出，如果我们将 LESS-VFL 的每个阶段运行至收敛，那么随着训练样本数量 N的增加，我们将以多项式速率趋近于一个使风险最小化并移除非显著特征的模型。由于参与方无法在本地计算风险 $R_N(\cdot)$，因此每个参与方都使用$H_N(\cdot)$，即生成的嵌入与作为预训练嵌入的显著分量的代理。显然，在无法访问服务器模型来计算风险的情况下，能否在每个参与方执行特征选择并非一目了然。定理 4.3 指出，无论服务器模型的深度或复杂程度如何，只要给定通过求解（3）获得的预训练嵌入，以及通过求解（$\mathcal{K}_m$）找到的显著嵌入分量集合，通过求解（5），每个参与方都能够成功移除其非显著特征。这强调了 LESS-VFL 的所有阶段都是必不可少的。

式(6)中的界与集中式组套索的界（Dinh 和 Ho，2020年）相似，只是增加了取决于参与方数量 M的次线性误差增长项。 M 在许多纵向联邦学习应用中通常很小（Kairouz 等，2021年），因此该项对界的影响很小。

4.2. 证明概要

为简洁起见，我们针对 $M = 1$ （一个参与方和服务器）的情况给出证明概要，并使用下标m 表示该参与方。我们在附录中给出了4.3 定理在 $M > 1$ 情况下的完整证明。$M > 1$ 的证明与 $M = 1$ 的证明类似，因为服务器端组套索将嵌入视为输入，并且与参与方数量无关；参与方端组套索则并行运行，仅使用自身的显著嵌入分量作为损失函数的代理。扩展到 $M > 1$ 时的关键挑战在于显著嵌入分量与显著参与方特征之间的关系（见附录B.9 中的引理B.5）。

我们先给出一些定义和补充记号。我们将 $H(\cdot)$定义为在完整数据分布$\mathcal{P}_{\mathbf{x},\mathbf{y}}$上两个嵌入之间的期望平方差：

$$H(\boldsymbol{\theta}_m; \boldsymbol{\theta}'_m; \mathcal{K}_m) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_{\mathbf{x},\mathbf{y}}}\left[e(\boldsymbol{\theta}_m; \boldsymbol{\theta}'_m; \mathcal{K}_m; \mathbf{x}_m)\right]$$

其中 $\boldsymbol{\theta}_m, \boldsymbol{\theta}'_m$ 是参与方的模型参数，$\mathcal{K}_m$ 是嵌入分量的集合，而 $e(\cdot)$ 是 $\mathcal{K}_m$中嵌入分量之间的平方差。回想一下，定义 2.1中给出的显著性概念适用于任意输入和模型。我们总结证明 $d(\boldsymbol{\Theta}, \mathcal{T}_\phi^*) \to 0$的步骤：

(a) 证明，若 $H(\cdot)$ 最小化，则当 $\mathcal{K}_m$为显著嵌入分量的集合时，$R(\cdot)$ 也会最小化。 (b) 证明，当 $\mathcal{K}_m$ 为显著嵌入分量的集合时，$e(\cdot\,;\,\mathcal{K}_m)$ 与 $f(\cdot)$ 具有相同的显著特征和非显著特征。(c) 证明，(4) 能找到最优服务器参数，并找到显著嵌入分量的集合。(d) 证明，给定显著嵌入分量的集合后，(5) 能找到最优参与方参数并移除非显著特征。

我们首先证明（a）和（b）。在以下命题中，

我们讨论完整网络中特征的重要性与参与方子网络中各特征对嵌入分量的重要性之间的关系。

命题 4.4。考虑一个模型 $\Theta = [\theta_s^\top, \theta_m^\top]^\top$。设 s 和 z 分别是 $f(\Theta)$ 的显著特征和非显著特征集合。设 $f(\theta_s; h(\theta_m; x))$ 的显著嵌入分量集合为 s_s。令 $g^j(x\ r)$ 将输入 x^j 替换为值 r。对于每个显著嵌入分量 $k \in s_s$，对于所有 $j \in z$ 以及任意 $r \in \mathbb{R}$，$h(\theta_m; x)^k = h(\theta_m; g^j(x\ r))^k$。

通俗地说，命题 4.4 指出，显著嵌入分量的取值不会因非显著特征而改变，而只会因显著特征而改变。

证明。假设对于某个显著嵌入分量 $k \in \mathbf{s}_s$、非显著特征 $j \in \mathbf{z}$ 以及 $r \in \mathbb{R}$，有 $\mathbf{h}(\boldsymbol{\theta}_m;\mathbf{x})^k \neq \mathbf{h}(\boldsymbol{\theta}_m; g^j(\mathbf{x}\ r))^k$。根据这一假设，且由于分量 k 是显著的，对于某个值 $r \in \mathbb{R}$，有 $f(\boldsymbol{\theta}_s; \mathbf{h}(\boldsymbol{\theta}_m; \mathbf{x})) \neq f(\boldsymbol{\theta}_s; \mathbf{h}(\boldsymbol{\theta}_m; g^j(\mathbf{x}; r)))$。这与 j 是非显著特征这一事实矛盾。 □

利用命题4.4，我们可以证明 (a) 和 (b)。

引理 4.5. 设 $\tilde{\Theta} = [\tilde{\theta}_s^\top, \tilde{\theta}_m^\top]^\top \in \mathcal{T}^*$。设 $\mathcal{K}_m$ 是 $f(\theta_s; h(\theta_m))$ 的显著嵌入分量。设 $\theta_m =$ 为 $arg\ min_{\theta'_m} H(\theta'_m; \tilde{\theta}_m; \mathcal{K}_m)$。设 s 和 z 分别是 $f(\Theta^\diamond)$ 的显著特征和非显著特征。设 s_h 和 z_h 分别是带有参数 θ_m 的 $e(\theta_m; \theta_m; \mathcal{K}_m)$ 的显著特征和非显著特征。那么：

$$[\tilde{\boldsymbol{\theta}}_s, \boldsymbol{\theta}_m] \in \mathcal{T}^*,\ \boldsymbol{s}_h = \boldsymbol{s},\ and\ \boldsymbol{z}_h = \boldsymbol{z}.$$

引理 4.5 的证明见附录 B.2。引理 4.5 证明了 $H(\cdot)$ 可以作为 $R(\cdot)$ 的代理，条件是预训练模型参数 Θ 属于最优集合 $\mathcal{T}^*$，且选定的嵌入分量集合 $\mathcal{K}_m$只包含显著嵌入分量。

接下来，我们考察目标 (c)，并证明服务器能够找到 (a) 和 (b) 所需的显著嵌入分量。首先，我们将 $\tilde{\Theta} = \arg\min_{\Theta \in \mathcal{T}^*} \|\Theta - \hat{\Theta}\|_2$ 定义为 $\mathcal{T}^*$ 中最接近预训练参数 Θ的模型。服务器的目标是找到 Θ中的显著嵌入分量集合，随后参与方可利用该集合移除其非显著特征。我们将 $\mathcal{T}^*$ 中非显著嵌入分量权重设为零的服务器模型参数集合定义如下：

$$\mathcal{S}_\phi^* = \{\boldsymbol{\theta}_s : \exists\, \boldsymbol{\theta}_m \text{ s.t. } [\boldsymbol{\theta}_s, \boldsymbol{\theta}_m] \in \mathcal{T}^* \text{ and } \mathbf{V}_s = \mathbf{0}\}$$

其中，$\mathbf{V}_s$ 是 Θ 中非显著嵌入分量的权重。若 $d(\boldsymbol{\theta}_s, \mathcal{S}_\phi^*) \to 0$，服务器便能找到显著嵌入分量的集合，从而完成目标 (c)。我们在以下定理中界定这一距离。

定理 4.6。给定由 Θ 定义的预训练模型 (3)，对于任意 $\delta > 0$，存在 $N \geq N_0(\delta)$使得：

$$d(\bar{\boldsymbol{\theta}}_s, \mathcal{S}_\phi^*) = O\left(\frac{\log N}{\lambda_s \sqrt{N}} + \left(\frac{\log N}{\sqrt{N}} + \lambda_s^{\nu/(\nu-1)}\right)^{1/\nu}\right)$$

以概率 $1 - \delta$。

定理4.6的证明见附录B.3。定理4.6 中的界表明，若 $\lambda_s \sim N^{-1/4}$，则非显著权重将随着训练样本数量 N 的增加以多项式速率趋近于零；若正则化参数设置得当，即可找到显著嵌入分量集合。

最后，由于服务器找到了显著嵌入分量集合，使得 $H(\cdot)$ 可以作为 $R(\cdot)$ 的代理，因此我们可以证明目标 (d)。我们将 $\mathcal{T}^*$ 中非显著特征权重设为零的参与方模型集合定义如下：

$$\mathcal{C}_\phi^* = \{\boldsymbol{\theta}_m : \exists\, \boldsymbol{\theta}_s \text{ s.t. } [\boldsymbol{\theta}_s, \boldsymbol{\theta}_m] \in \mathcal{T}^* \text{ and } \mathbf{V}_m = \mathbf{0}\}$$

其中，$\mathbf{V}_m$ 是生成模型 $\Theta^\diamond$中非显著特征的输入权重。如果 $d(\boldsymbol{\theta}_m, \mathcal{C}_\phi^*) \to 0$，那么参与方就会移除非显著特征，从而完成目标 (d)。我们将在下述定理中界定这一距离。

定理4.7。令 $[\tilde{\theta}_s^\top, \tilde{\theta}_m^\top]^\top = \text{argmin}_{\Theta \in \mathcal{T}^*} \|\Theta - \hat{\Theta}\|_2$其中 Θ 是式 (3)中定义的预训练模型参数。如果 $\mathcal{K}_m$ 在式 (5) 中是显著嵌入分量集合对于 $f(\boldsymbol{\theta}_s;\ \boldsymbol{h}(\boldsymbol{\theta}_m))$，那么对于任意 $\delta > 0$，都存在某个样本数量 $N \geq N_0(\delta)$使得：

$$d(\bar{\boldsymbol{\theta}}_m, \mathcal{C}_\phi^*) \leq O\left(\frac{\log N}{\lambda_m \sqrt{N}} + \left(\frac{\log N}{\sqrt{N}} + \lambda_m^{\nu/(\nu-1)}\right)^{1/\nu}\right)$$

以概率 $1 - \delta$。

定理 4.7的证明见附录 B.4。与服务器的情况类似，当 $\lambda_m \sim N^{-1/4}$ 时，该界以多项式速率趋于零。注意，要使定理 4.7 成立，服务器必须向参与方提供重要嵌入的集合。否则，无法保证该界。在第 5 节中，我们探讨嵌入分量选择阶段在实践中的重要性。在附录 B.5 中，我们将定理 4.6 和 4.7 扩展到 $M > 1$ 的情况，并可结合这些结果证明定理 4.3。

5. 实验

我们通过运行固定次数的算法2（标准纵向联邦学习算法，见附录A），然后运行算法1来移除非显著特征，之后继续使用算法2进行训练，从而实现 LESS-VFL。我们在多个数据集上评估了 LESS-VFL。

• **MIMIC-III**（Johnson 等，2016年；Harutyunyan 等，2019年）：由匿名患者的时间序列医疗信息构成的医院数据集。用于预测住院期间死亡率。包含14，681 个样本，每个样本包含712 个特征。

• **活动**（Anguita 等人，2013）：记录人类进行各种活动的时间序列位置数据。用于对当前活动（行走、坐下、跑步等）进行多分类。包含7，352 个样本，每个样本包含560 个特征。

• **网络钓鱼**（Dua 和 Graff，2017）：提供用于判断网站是否为网络钓鱼网站的相关特征的数据集（HTTP、TinyURL、转发等）。包含11，055 个样本，每个样本包含30 个特征。

• **Gina**（Guyon，2007年）：手写的两位数图像。用于对奇数和偶数进行二分类，这意味着只有第一位数字是分类所必需的，其余特征均为干扰项。包含3，468 个样本，每个样本包含968 个特征。

• **Sylva**（Guyon，2007年）：森林覆盖类型信息。用于二分类（黄松与其他所有类型）。与 Gina 类似，一半的特征是干扰项；每个样本包含两条与目标相关的信息记录，另外两条为随机选取。包含14，395 个样本，每个样本包含216 个特征。

对于每个数据集，我们额外添加了 50% 的高斯噪声特征。这些伪特征充当非显著特征，使我们能够衡量 LESS-VFL 的特征选择效果。请注意，原始数据集中的特征不一定全都显著。我们唯一确定为非显著的特征，是添加到每个数据集中的高斯噪声特征。因此，最终测试准确率可作为指标，反映我们是否正确选出了数据集中的显著特征，并训练出了泛化能力良好的模型。

我们将 LESS-VFL 与以下纵向联邦学习基线进行比较。

•**VFL（原始）**：如算法2所述的纵向联邦学习，且数据集中不含伪特征。

• **VFL（伪特征）**：算法2 包含数据集中的伪特征。

• **组套索**：通过使用 P-SGD 近似求解（2），将组套索直接应用于纵向联邦学习模型。

• **局部套索**：该算法与 LESS-VFL 相同，但移除了阶段 2 的嵌入分量选择。

我们将评估范围限定为无需以训练完成的纵向联邦学习模型作为输入的方法，因此比较中不包括 Zhang 等人（2022a年）和Chen 等人（2022年）。VFLFS（Feng，2022年）中的特征选择部分采用组套索，我们将其纳入评估。

训练细节。对于每个数据集，我们将原始特征和高斯噪声特征均匀地分配给一组参与方（网络钓鱼任务中有三方，其他任务中有四方）。每个参与方的模型都是三层全连接神经网络，服务器训练一个线性模型，该模型将参与方嵌入的拼接结果作为输入。我们通过网格搜索确定 LESS-VFL、局部套索和组套索的正则化参数，以及 LESS-VFL 和局部套索的预训练轮数。我们选取能达到最高训练准确率且至少移除80% 伪特征的参数。对于在纵向联邦学习中使用算法2，以及局部套索和 LESS-VFL 的预训练和特征选择后阶段，我们使用学习率为 0.01 的 ADAM 优化器。我们运行150轮 P-SGD，用于 LESS-VFL 中的嵌入分量选择，以及 LESS-VFL 和局部套索中的特征选择；我们发现，这样的迭代次数足以使训练损失趋于平稳。

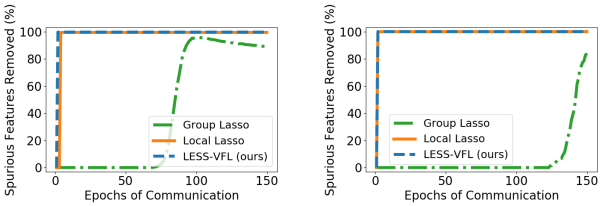
表2。 达到基线测试准确率的90%并移除至少80%的伪特征所需的通信成本。所示数值为5次运行的平均值 ± 标准差。

数据集	通信成本（MB）		
	组套索	局部套索	LESS-VFL（我们的方法）
MIMIC-III	57.35 ± 0.00	30.47 ± 1.79	**7.17 ± 0.00**
活动	322.73 ± 61.32	26.56 ± 5.83	**21.17 ± 0.00**
网络钓鱼	95.22 ± 1.89	8.10 ± 3.40	**3.99 ± 0.75**
Gina	13.55 ± 0.00	1.90 ± 0.27	**1.48 ± 0.26**
Sylva	22.49 ± 0.00	**5.62 ± 0.00**	**5.62 ± 0.00**

通信成本。在表2中，我们比较了在至少移除 80% 的伪特征的同时，达到目标测试准确率所需的通信成本。我们将目标准确率设定为 VFL（原始）所达到的最高准确率的 90% 。在所有情况下，LESS-VFL 均以最低的通信成本满足这些条件，与组套索相比降低了通信成本。在钓鱼数据集的情况下，LESS-VFL 的通信成本比组套索 ∼20× 更低。LESS-VFL 仅在预训练期间以及特征选择开始时进行一次通信，从而大幅降低了相较于组套索的特征选择成本。此外，LESS-VFL 的通信成本始终不高于局部套索。局部套索放弃了嵌入分量选择，而在大多数数据集中，这导致了更高的通信成本。我们将在下一组实验中进一步探讨这一点。

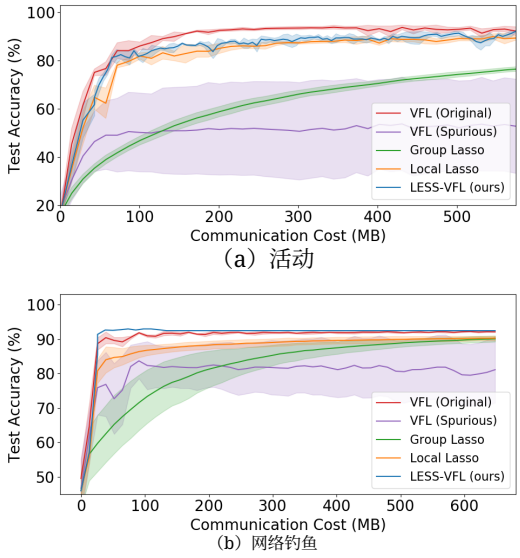
在剩余的实验中，我们旨在展示 LESS-VFL 在整个训练过程中的表现。我们重点关注两个具有代表性的数据集（活动和网络钓鱼）。所有数据集的实验结果均见附录 C。

特征移除。 在图 3中，我们比较了组套索、局部套索和 LESS-VFL 随通信周期推进而移除的伪特征百分比。组套索在训练过程中逐渐移除特征，



（a）活动　（b）网络钓鱼

图3。 移除伪特征所需的通信轮次。所示数值为 5 次运行的平均值。预训练后，LESS-VFL 和局部套索移除的伪特征比例相近，但局部套索达到高准确率所需时间更长（见表 5）。组套索逐步移除特征，而局部套索和 LESS-VFL 在预训练后仅需少数几个通信轮次即可移除特征。



（a）活动

（b）网络钓鱼

图4. 测试准确率随通信成本的变化。纵向联邦学习（原始）在不含伪特征的情况下训练，而其他所有方法均在含有伪特征的情况下训练。实线表示5次运行的平均值，阴影区域表示标准差。

而局部套索和 LESS-VFL 会在预训练的若干轮通信后移除特征。LESS-VFL 从利用无需通信的本地训练来执行特征选择中获益良多。局部套索在大致相同的通信周期内移除的特征比例与 LESS-VFL 相近。然而，我们会在下一个实验中看到，局部套索可能需要更多通信，才能同时达到高准确率并移除伪特征。

准确率。 在图4中，我们绘制了测试准确率与通信成本之间的关系。两个数据集中 VFL（伪特征）的测试准确率表明，在未移除伪特征的情况下进行垂直联邦学习训练，可能会严重影响泛化能力。在这两种情况下，LESS-VFL 都比组套索更快达到高准确率，并且达到了与不含伪特征的基线 VFL 算法相近的准确率。事实上，在

表3. 异构特征划分下的实验结果。达到基准测试准确率的90%，并移除至少80%的伪特征所需的通信成本。所示数值为5次运行的平均值 ± 标准差。

数据集	通信成本（MB）		
	组套索	局部套索	LESS-VFL（本文方法）
MIMIC-III	57.35 ± 0.00	**7.17 ± 0.00**	**7.17 ± 0.00**
活动	187.39 ± 52.61	24.77 ± 6.75	**15.76 ± 3.09**
网络钓鱼	94.57 ± 1.94	5.51 ± 0.79	**3.98 ± 0.74**
Gina	13.55 ± 0.00	1.63 ± 0.33	**1.35 ± 0.00**
Sylva	21.93 ± 1.12	**5.62 ± 0.00**	**5.62 ± 0.00**

钓鱼数据集中，LESS-VFL 的表现优于 VFL（原始）基线。这是因为 LESS-VFL 移除了不显著的嵌入分量，从而降低了特征选择后的通信成本。在 Activity数据集中，局部套索的表现与 LESS-VFL 相近，但在钓鱼数据集中，局部套索需要高得多的通信成本才能达到高准确率。在图4b中，我们可以看到，在特征选择后，局部套索的模型准确率低于 LESS-VFL（在 ∼25 MB 时）。这进一步表明，嵌入分量选择既可以通过降低优化服务器模型参数的风险来提高模型准确率，也可以为参与方提供进行局部特征选择所需的重要信息。

不均匀特征。 在之前的实验中，我们考虑了所有参与方拥有相同比例高斯噪声特征的情况。现在，我们考虑各参与方的高斯噪声特征分布不均匀的情况：一个参与方拥有80% 的额外高斯噪声特征，一个拥有 25%，一个拥有 10%，还有一个不含高斯噪声特征。表 3 展示了组套索、局部套索和 LESS-VFL 在移除总伪特征的 80% 后，达到基线纵向联邦学习 VFL（原始）测试准确率90% 所需的通信成本。我们发现，在这种异构设置下，LESS-VFL 在移除伪特征的同时仍能实现高准确率，并且通信成本较低。

6. 结论

在这项工作中，我们提出了 LESS-VFL，这是一种用于纵向联邦学习中特征选择的高通信效率方法。我们通过解析证明，LESS-VFL 可以移除伪特征。实验结果表明，LESS-VFL 在降低通信成本的同时，能够实现相当的准确率，并移除相近比例的伪特征。未来，我们计划将分析扩展到非解析神经网络和自适应组套索。

致谢

本材料部分基于美国国家科学基金会资助的项目（Grant CNS-1553340），以及伦斯勒-IBM人工智能研究合作项目（http://airc.rpi.edu），该项目隶属于IBM人工智能视界网络。

参考文献

Anguita, D., Ghio, A., Oneto, L., Parra, X. 和 Reyes-Ortiz, J. L. 使用智能手机进行人体活动识别的公有领域数据集。载于 欧洲人工神经网络研讨会，2013 年。

Canino, G., Suo, Q., Guzzi, P. H., Tradigo, G., Zhang, A. 和 Veltri, P. 用于诊断、电子病历和地理数据相关性的特征选择模型。载于 *ACM*生物信息学、计算生物学与健康信息国际会议论文集，第 616–621 页，2016年。Castiglia, T., Das, A., Wang, S. 和 Patterson, S. Compressed-VFL：针对纵向切分数据的通信高效学习。载于 第*39*届国际机器学习会议论文集，第 2738–2766 页，2022年。Ceballos, I., Sharma, V., Mugica, E., Singh, A., Roman, A., Vepakomma, P. 和 Raskar, R. 由 SplitNN 驱动的纵向切分。 *arXiv:2008.04137*，2020年。Cha, D., Sung, M. 和 Park, Y.-R. 使用自编码器实现纵向联邦学习：实际应用、泛化能力与效用研究。 *JMIR*医学信息学，9(6):e26598，2021年。Chandrashekar, G. 和 Sahin, F. 特征选择方法综述。 计算机与电气工程，40(1):16–28，2014年。Chen, P., Du, X., 陆 Z., 吴 J. 和 Hung, P. C. K. EVFL：面向数据导向型人工智能系统的可解释纵向联邦学习。 系统架构学报，126:102474，2022年。Chen, X., Zhou, S., Guan, B., Yang, K., Fao, H., Wang, H. 和 Wang, Y. Fed-EINI：用于纵向联邦学习中决策树集成的高效可解释推理框架。载于 *IEEE*大数据国际会议，第 1242–1248 页，2021年。Clinciu, M. 和 Hastie, H. 可解释人工智能术语综述。载于交互式自然语言技术与可解释人工智能研讨会论文集，第 8–13 页，2019年。Dinh, V. C. 和 Ho, L. S. T. 解析型深度神经网络的一致性特征选择。载于 神经信息处理系统进展，2020年。Dua, D. 和 Graff, C. UCI机器学习知识库，2017年。网址http://archive.ics.uci.edu/ml。

Feng, S. 基于纵向联邦学习且利用非重叠样本的特征选择。*Expert Syst.Appl.*，208:118097，2022年。Feng, S. 和 Yu, H. 多参与方多分类纵向联邦学习。*arXiv*，arbs/2001.11154，2020年。Guyon, I. 无先验学习与先验知识。IJCNN 无先验学习与先验知识研讨会，2007年。URL http://www.agnostic.inf.ethz.ch/datasets.php。访问日期：2023年4月27日。Harutyunyan, H.、Khachatrian, H.、Kale, D. C.、Ver Steeg, G. 和 Galstyan, A. 基于临床时间序列数据的多任务学习与基准测试。*Scientific Data*，6(1)：96，2019年。Hou, J.、Su, M.、Fu, A. 和 Yu, Y. 基于纵向联邦随机森林的可验证隐私保护方案。*IEEE Internet Things*，9(22)：22158–22172，2022年。Hu, Y.、Niu, D.、Yang, J. 和 Zhou, S. FDML：面向分布式特征的协作式机器学习框架。*Proc. ACM Int. Conf.Knowl. Discov. Data Min.*，第2232–2240页，2019年。Huang, J.、Horowitz, J. L. 和 Wei, F. 非参数加性模型中的变量选择。*The Annals ofStatistics*，38(4)：2282–2313，2010年。Ji, S.、Koll´ar, J. 和 Shiffman, B. 代数簇的全局Łojasiewicz 不等式。*Trans. AmericanMath.Soc.*，329(2)：813–818，1992年。Johnson, A. E.、Pollard, T. J.、Shen, L.、Lehman, L.-w. H.、Feng, M.、Ghassemi, M.、Moody, B.、Szolovits, P.、Anthony Celi, L. 和 Mark, R. G. MIMIC-III：一个免费开放的重症监护数据库。*Nature*，2016年。Kairouz, P.、McMahan, H. B.、Avent, B.、Bellet, A.、Bennis, M.、Bhagoji, A. N.、Bonawitz, K. A.、Charles, Z.、Cormode, G.、Cummings, R.、D'Oliveira, R. G. L.、Eichner, H.、Rouayheb, S. E.、Evans, D.、Gardner, J.、Garrett, Z.、Gasc´on, A.、Ghazi, B.、Gibbons, P. B.、Gruteser, M.、Harchaoui, Z.、He, C.、He, L.、Huo, Z.、Hutchinson, B.、Hsu, J.、Jaggi, M.、Javidi, T.、Joshi, G.、Khodak, M.、Kone˘cn´y, J.、Korolova, A.、Koushanfar, F.、Koyejo, S.、Lepoint, T.、¨Liu, Y.、Mittal, P.、Mohri, M.、Nock, R.、Ozg¨ur, A.、Pagh, R.、Qi, H.、Ramage, D.、Raskar, R.、Raykova, M.、Song, D.、Song, W.、Stich, S. U.、Sun, Z.、Suresh, A. T.、Tram`er, F.、Vepakomma, P.、Wang, J.、Xiong, L.、Xu, Z.、Yang, Q.、Yu, F. X.、Yu, H. 和 Zhao, S. 联邦学习的进展与开放性问题。*Found. TrendsMach. Learn.*，14(1-2)：1–210，2021年。doi：10.1561/2200000083。Li, A.、Peng, H.、Zhang, L.、Huang, J.、Guo, Q.、Yu, H. 和 Liu, Y. FedSDG-FS：面向纵向联邦学习的高效安全特征选择。载于*IEEE Int. Conf. Comp.Comm.*，2023年。

Qiu, P.、Zhang, X.、Ji, S.、Pu, Y. 和 Wang, T. 你所需要的只是哈希：防御纵向联邦学习中的数据重构攻击。*arXiv:2212.00325*，2022年。Sun, C.、Ippel, L.、van Soest, J.、Wouters, B.、Malic, A.、Adekunle, O.、van den Berg, B.、Mussmann, O.、Koster, A.、van der Kallen, C.、van Oppen, C.、Townend, D.、Dekker, A. 和 Dumontier, M. 用于分析垂直分区场景下个人健康数据的隐私保护基础设施。264：373–377，2019年。Wang, J.、Zhang, H.、Wang, J.、Pu, Y. 和 Pal, N. R. 使用带有组套索正则化和受控冗余的神经网络进行特征选择。*IEEETrans.NeuralNetworks Learn. Syst.*，32(3)：1110–1123，2021年。Wu, Y. 和 Liu, Y. 分位数回归中的变量选择。*Statistica Sinica*，第801–817页，2009年。Yang, Q.、Liu, Y.、Chen, T. 和 Tong, Y. 联邦机器学习：概念与应用。*ACM*智能系统与技术汇刊，10(2)：12:1–12:19，2019年。Zhang, H.、Wang, J.、Sun, Z.、Zurada, J. M. 和 Pal, N. R. 使用组套索正则化进行神经网络特征选择。*IEEE*知识与数据工程汇刊，32(4)：659–673，2020年。Zhang, R.、Li, H.、Hao, M.、Chen, H. 和 Zhang, Y. 电子健康系统中的安全特征选择：面向纵向联邦学习。载于 *IEEE*国际通信会议，第1257–1262页。IEEE，2022a年。Zhang, Y.、Hu, Y.、Gao, X.、Gong, D.、Guo, Y.、Gao, K. 和 Zhang, W. 一种基于粒子群优化的嵌入式纵向联邦特征选择算法。中国人工智能学会智能技术汇刊，2022b年。Zhao, L.、Hu, Q. 和 Wang, W. 使用多模态深度神经网络和稀疏组套索进行异构特征选择。*IEEE*多媒体汇刊，17(11)：1936–1948，2015年。Zhou, J.、Zhang, S.、Lu, Q.、Dai, W.、Chen, M.、Liu, X.、Pirt tikangas, S.、Shi, Y.、Zhang, W. 和 Herrera-Viedma, E. 联邦学习及其加速工业物联网应用的综述。*arXiv:2104.10501*，2021年。Zou, T.、Liu, Y.、Kang, Y.、Liu, W.、He, Y.、Yi, Z.、Yang, Q. 和 Zhang, Y.-Q. 防御纵向联邦学习中的批次级标签推断和替换攻击。*IEEE*大数据汇刊，2022年。

A. 纵向联邦学习算法

在算法2中，我们给出了使用神经网络进行标准纵向联邦学习训练的伪代码（Hu 等人，2019年；Ceballos 等人，2020年）。

算法2纵向联邦学习

1：初始化：为所有参与方 θ_m^0 和服务器模型 m $\theta_s^0$2：对于 $t \leftarrow 0$，……，$T_1 - 1$，执行 3：随机采样 $\mathcal{B} \subset [N]$4：并行地对 $m \leftarrow 1$，……，M执行 5：将 $\mathbf{h}_m(\theta_m^t; \mathbf{X}_m^{(\mathcal{B})})$ 发送给服务器 6：结束循环 7：$\Phi \leftarrow \{\theta_s^t, \mathbf{h}_1(\theta_1^t; \mathbf{X}_m^{(\mathcal{B})})$ ……，$\mathbf{h}_M(\theta_M^t; \mathbf{X}_m^{(\mathcal{B})})\}$8：$\theta_s^{t+1} \leftarrow U(\theta_s^t, \nabla_s R_{\mathcal{B}}(\Phi^t; \mathbf{y}^{\mathcal{B}^t}))$ 9：服务器将$\nabla_{\mathbf{h}_m}(\theta_m^t)$ $R_{\mathcal{B}}(\Phi^t; \mathbf{y}^{(\mathcal{B})})$发送给每个参与方 10：并行地对 $m \leftarrow 1$，……，M执行 11：$\nabla_m R_{\mathcal{B}}(\Phi^t) = \nabla_{\theta_m} \mathbf{h}_m(\theta_m^t)^\top \nabla_{\mathbf{h}_m}(\theta_m^t) R_{\mathcal{B}}(\Phi^t)$ 12：$\theta_m^{t+1} = U(\theta_m^t, \nabla_m R_{\mathcal{B}}(\Phi^t))$ 13：结束循环 14：结束循环

参与方首先就小批量样本 $\mathcal{B}$ 达成一致，然后将给定小批量的当前嵌入发送给服务器。我们令 $\mathbf{X}^{(\mathcal{B})}$ 和 $\mathbf{y}^{(\mathcal{B})}$ 分别表示小批量中的训练样本和标签。服务器使用关于 θ_s的小批量偏导数（记为 $\nabla_s R_{\mathcal{B}}(\cdot)$）以及某种优化器更新规则$U(\cdot)$（例如随机梯度下降、Adam 等）更新其模型。随后，服务器将关于参与方嵌入的偏导数发送给各参与方。然后，每个参与方 m 使用其小批量偏导数更新模型，该偏导数记为 $\nabla_m R_{\mathcal{B}}(\cdot)$。

B. 定理 4.3 的证明

在本节中，我们首先证明 4.7 和 4.6 在 $M = 1$ 情况下的定理，并将结果推广至 $M > 1$ 情况，最后证明定理 4.3。我们在表 4 中总结了本节所使用的符号。

B.1. M = $M = 1$ 时的补充符号说明

我们首先给出证明定理 4.7 和4.6所需的补充符号。我们定义最优参数集合中参与方和服务器模型参数的集合 $\mathcal{T}^*$:

$$\mathcal{C}^* = \{\boldsymbol{\theta}_m : \exists \boldsymbol{\theta}_s \text{ s.t. } [\boldsymbol{\theta}_s, \boldsymbol{\theta}_m] \in \mathcal{T}^*\}$$

and

$$\mathcal{S}^* = \{\boldsymbol{\theta}_s : \exists \boldsymbol{\theta}_m \text{ s.t. } [\boldsymbol{\theta}_s, \boldsymbol{\theta}_m] \in \mathcal{T}^*\}.$$

我们使用由Dinh 和 Ho (2020年)证明的以下引理:

引理 B.1。 设 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 分别为生成模型中的显著和非显著输入层权重 $\boldsymbol{\Theta}^\diamond$。设 $\phi(\boldsymbol{\Theta})$为参数 $\boldsymbol{\Theta}$，其中所有非显著输入层权重$\boldsymbol{V}$ 均设为零。在假设 *4.1*下，

- 存在 $c_0 > 0$，使得 $\|\boldsymbol{U}^k\| \geq c_0$ 对所有 $\boldsymbol{\Theta} \in \mathcal{T}^*$ 和 $k = 1, \ldots, d^s$ 均成立（其中 d^s 为显著特征的数量）。
- 若 $\boldsymbol{\Theta} \in \mathcal{T}^*$，则参数 $\phi(\boldsymbol{\Theta})$ 也属于 $\mathcal{T}^*$。

引理 B.2. 存在 $c_2, \nu > 0$，使得:

$$c_2 d(\boldsymbol{\Theta}, \mathcal{T}^*)^\nu \leq R(\boldsymbol{\Theta}) - R(\boldsymbol{\Theta}^\diamond)$$

对所有 $\boldsymbol{\Theta} \in \mathcal{T}$。

表4。 符号汇总。

符号	定义
N	训练样本数量。
M	参与方数量。
λ_s, λ_m	服务器和参与方的正则化系数。
$f(\cdot)$	纵向联邦学习模型的标签预测。
$\mathbf{h}(\cdot)$	参与方的本地嵌入函数。
$e(\cdot)$	两个嵌入之间的均方误差。
$R(\cdot)$	风险函数：所有可能样本的标签对应的均方误差。
$R_N(\cdot)$	经验风险函数：所有训练样本的标签对应的均方误差（MSE）。
$G(\cdot)$	组套索$L_{2,\ 1}$正则化项。
$H(\cdot)$	两个嵌入函数之间的预期均方差。
$H_N(\cdot)$	两个嵌入函数之间的经验均方差。
$d(\cdot)$	向量与向量集合之间的距离。
$\boldsymbol{\Theta}^\diamond$	假设中定义的生成模型参数4.1。
$\hat{\boldsymbol{\Theta}}$	通过最小化经验风险获得的预训练模型参数。
$\tilde{\boldsymbol{\Theta}}$	与生成模型具有相同风险且最接近预训练模型的模型。
$\bar{\boldsymbol{\Theta}}$	运行 LESS-VFL 后学习到的模型参数。
$\mathbf{U}_m$，$\mathbf{V}_m$	参与方 m的生成模型中显著特征和非显著特征的输入权重。
$\mathcal{T}^*$	与生成模型具有相同风险的模型集合。
$\mathcal{S}^*$	与生成服务器模型具有相同风险的服务器模型集合。
$\mathcal{C}^*$	与生成客户端模型具有相同风险的客户端模型集合。
$\mathcal{T}^*_\phi$	将非显著特征权重设为零的 $\mathcal{T}^*$ 子集。
$\mathcal{S}^*_\phi$	将非显著嵌入权重设为零的 $\mathcal{T}^*$ 中的服务器模型。
$\mathcal{C}^*_\phi$	将非显著特征权重设为零的参与方模型 $\mathcal{T}^*$ 。

引理 B.3。 对于任意 $\delta > 0$，存在 $c_1(\delta) > 0$，使得对于所有 $\boldsymbol{\Theta} \in \mathcal{T}$：

$$|R_N(\boldsymbol{\Theta}) - R(\boldsymbol{\Theta})| \leq c_1 \frac{\log N}{\sqrt{N}}$$

概率为 $1-\delta$。

我们还证明了以下引理：

引理 B.4。给 定一个由 $\hat{\boldsymbol{\Theta}}$ 定义的模型 (3)，对于任意 $\delta > 0$，存在 $C_\delta(\delta) > 0$ 和 $N \geq N_0(\delta)$ 使得 ：

$$d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) \leq C_\delta \frac{\log N}{\sqrt{N}} \tag{7}$$

概率为 $1-\delta$。

证明。令 $[\hat{\boldsymbol{\theta}}_m^\top, \hat{\boldsymbol{\theta}}_s^\top]^\top = \hat{\boldsymbol{\Theta}}$ 表示预训练步骤后的参与方和服务器模型参数。我们将 $\tilde{\boldsymbol{\Theta}} = \text{argmin}_{\boldsymbol{\Theta} \in \mathcal{T}^*} \|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}\|_2 = [\tilde{\boldsymbol{\theta}}_s^\top, \tilde{\boldsymbol{\theta}}_m^\top]^\top$ 定义为最接近预训练模型的最优模型。根据引理B.2和B.3，我们有如下结论：

$$\begin{aligned} c_2 d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*)^\mu &= c_2 \|\tilde{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}\|_2^\mu && (8) \\ &\leq R(\hat{\boldsymbol{\theta}}_s; \mathbf{h}(\hat{\boldsymbol{\theta}}_m)) - R(\tilde{\boldsymbol{\theta}}_s; \mathbf{h}(\tilde{\boldsymbol{\theta}}_m)) && (9) \\ &\leq 2c_1 \frac{\log N}{\sqrt{N}} + R_N(\hat{\boldsymbol{\theta}}_s; \mathbf{h}(\hat{\boldsymbol{\theta}}_m)) - R_N(\tilde{\boldsymbol{\theta}}_s; \mathbf{h}(\tilde{\boldsymbol{\theta}}_m)) && (10) \end{aligned}$$

^^~ ~注意， $R_N(\boldsymbol{\theta}_s; \mathbf{h}(\boldsymbol{\theta}_m)) \leq R_N(\boldsymbol{\theta}_s; \mathbf{h}(\boldsymbol{\theta}_m))$，因此：

$$c_2 d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*)^\mu \leq 2c_1 \frac{\log N}{\sqrt{N}} \tag{11}$$

$$d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) \leq \left(\frac{2c_1}{c_2} \frac{\log N}{\sqrt{N}} \right)^{1/\mu} \tag{12}$$

令 $\mu = 1$。由此完成引理 B.4 的证明。 □

B.2. 引理的证明4.5

证明。注意， $H(\boldsymbol{\theta}_m; \tilde{\boldsymbol{\theta}}_m; \mathcal{K}_m)$ 的最小化会导致对于所有显著嵌入分量 $k \in \mathcal{K}_m$ 和任意输入 x，$\mathbf{h}(\boldsymbol{\theta}_m; \mathbf{x})^k = \mathbf{h}(\tilde{\boldsymbol{\theta}}_m; \mathbf{x})^k$。根据 $\mathcal{T}^*$ 的定义和定义 2.1，这意味着 $R(\boldsymbol{\theta}_s^\circ; \mathbf{h}(\boldsymbol{\theta}_m^\circ))$ ~ ~ ~ ~ $= R(\boldsymbol{\theta}_s; \mathbf{h}(\boldsymbol{\theta}_m))$ $= R(\boldsymbol{\theta}_s; \mathbf{h}(\boldsymbol{\theta}_m))$ ）。因此， $[\boldsymbol{\theta}_s, \boldsymbol{\theta}_m] \in \mathcal{T}^*$。

<escape>~ ~</escape>根据（Dinh 和 Ho，2020年）中的引理 3.1，由于 <escape> $\boldsymbol{\Theta} \in \mathcal{T}^*$, $f(\boldsymbol{\Theta}; \mathbf{x}; y) = f(\boldsymbol{\Theta}^\circ; \mathbf{x}; y)$</escape> 对所有输入 x 均成立。这意味着，<escape> $f(\boldsymbol{\Theta}^\circ)$</escape> 的显著特征和非显著特征必须与 <escape> $f(\boldsymbol{\Theta})$</escape> 的相同。设 <escape> ~s</escape> 和 <escape> ~z</escape> 分别为 <escape> $f(\boldsymbol{\Theta})$</escape> 的显著特征和非显著特征。那么必然有 <escape> ~s $= \mathbf{s}$</escape> 和 <escape> ~z $= \mathbf{z}$</escape>。

令 $j \in$ ~$\mathbf{z}$ 为 $f(\tilde{\boldsymbol{\Theta}})$ 中的非显著特征，并令 $r \in \mathbb{R}$。令 $g^j(\mathbf{x}\ s)$ 为一个将 $\mathbf{x}^j$ 替换为值 s 的函数。我们知道，对于所有 $k \in \mathcal{K}_m$：

$$\mathbf{h}(\boldsymbol{\theta}_m; g^j(\mathbf{x}, r))^k = \mathbf{h}(\tilde{\boldsymbol{\theta}}_m; g^j(\mathbf{x}, r))^k = \mathbf{h}(\tilde{\boldsymbol{\theta}}_m; \mathbf{x})^k.$$

事实上，根据命题 4.4， $\mathcal{K}_m$ 中的所有嵌入分量仅依赖于 ~s。由于对于所有 $k \in \mathcal{K}_m$， $\mathbf{h}(\boldsymbol{\theta}_m; \cdot)^k$ 不受 ~z 中特征的影响~，这意味着 $e(\boldsymbol{\theta}_m; \boldsymbol{\theta}_m; \mathcal{K}_m)$ 的非显著特征集合包含 $f(\boldsymbol{\Theta})$ 的非显著特征集合~： $\mathbf{z}_h \supseteq$ ~z。

~类似地，令 k 为 f（$\boldsymbol{\Theta}$）的一个重要特征。根据命题4.4，对于所有 $k \in \mathcal{K}_m$：

$$\mathbf{h}(\boldsymbol{\theta}_m; g^k(\mathbf{x}, r))^k = \mathbf{h}(\tilde{\boldsymbol{\theta}}_m; g^k(\mathbf{x}, r))^k \neq \mathbf{h}(\tilde{\boldsymbol{\theta}}_m; \mathbf{x})^k$$

对于某个 $r \in \mathbb{R}$。这意味着：

$$\sum_{k \in \mathcal{K}_m} (\mathbf{h}(\boldsymbol{\theta}_m; g^k(\mathbf{x}, r))^k - \mathbf{h}(\tilde{\boldsymbol{\theta}}_m; \mathbf{x})^k)^2 \neq \sum_{k \in \mathcal{K}_m} (\mathbf{h}(\boldsymbol{\theta}_m; \mathbf{x})^k - \mathbf{h}(\tilde{\boldsymbol{\theta}}_m; \mathbf{x})^k)^2$$

~ ~我们可以说， $e(\boldsymbol{\theta}_m; \boldsymbol{\theta}_m; \mathcal{K}_s)$ 的显著特征包含 $f(\boldsymbol{\Theta})$ 的显著特征集合：$\mathbf{s}_h \supseteq$ ~$\mathbf{s}$。因此，$\mathbf{s}_h =$ ~$\mathbf{s}$ $= \mathbf{s}$ 且$\mathbf{z}_h =$ ~$\mathbf{z}$ $= \mathbf{z}$。 □

B.3. 定理证明4.6

接下来，我们证明，求解（4）的服务器会找到一个最优解，同时将不重要的嵌入分量~权重设为零。我们将 $\boldsymbol{\theta}_s$ 定义为求解（4）的服务器模型参数。首先，我们证明以下引理：

引理 *B.5.* 设 L 为 $f(\cdot)$ 的 *Lipschitz* 常数。给定一个预训练模型 ^ $\boldsymbol{\Theta} = [\hat{\boldsymbol{\theta}}_m^\top$，以及由 *(3)* 定义的 $\hat{\boldsymbol{\theta}}_s^\top]^\top$，令 $\tilde{\boldsymbol{\Theta}} = arg\ min_{\boldsymbol{\Theta} \in \mathcal{T}^*} \|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}\|_2 = [\tilde{\boldsymbol{\theta}}_s^\top, \tilde{\boldsymbol{\theta}}_m^\top]^\top$ 为距离预训练模型最近的最优模型。对于任意 $\delta > 0$，存在 $C_1(\delta)$ $C_2(\delta)$ $C_3(\delta)$ $C_4(\delta)$ $C_5 > 0$ 和 $N \geq N_0(\delta)$，使得：

$$d(\bar{\boldsymbol{\theta}}_s, \mathcal{S}^*) \leq d(\bar{\boldsymbol{\theta}}_s, \{\tilde{\boldsymbol{\theta}}_s\}) \leq \left(C_1 \frac{\log N}{\sqrt{N}} + C_2 \lambda_s^{\nu/(\nu-1)} + C_3 L d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) \right)^{1/\nu} \tag{13}$$

且非显著嵌入分量权重之和为

$$\sum_l \| \boldsymbol{V}_s^l \|_2 \leq C_4 \frac{\log N}{\lambda_s \sqrt{N}} + \frac{2L}{\lambda_s} d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) + C_5 d(\bar{\boldsymbol{\theta}}_s, \{\tilde{\boldsymbol{\theta}}_s\}). \tag{14}$$

概率为 $1 - \delta$。

~ ~ ~ ~证明。注意，$\{\tilde{\boldsymbol{\theta}}_s\}$ 是解析函数 $E(\boldsymbol{\theta}_s) = R(\boldsymbol{\theta}_s;\ \mathbf{h}(\boldsymbol{\theta}_m)) - R(\boldsymbol{\theta}_s;\ \mathbf{h}(\boldsymbol{\theta}_m))$的零水平集。我们可以应用 Łojasiewicz 不等式（Ji 等人，1992年）如下：

$$
\begin{aligned}
c_2 d(\bar{\boldsymbol{\theta}}_s, \mathcal{S}^*)^\nu &\le c_2 d(\bar{\boldsymbol{\theta}}_s, \{\tilde{\boldsymbol{\theta}}_s\})^\nu && (15)\\
&= c_2 \|\tilde{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|_2^\nu && (16)\\
&\le R(\bar{\boldsymbol{\theta}}_s; \mathbf{h}(\tilde{\boldsymbol{\theta}}_m)) - R(\tilde{\boldsymbol{\theta}}_s; \mathbf{h}(\tilde{\boldsymbol{\theta}}_m)) && (17)
\end{aligned}
$$

由于 $f(\cdot)$ 是解析的，我们知道风险函数是光滑的。令 L 为 $R(\cdot)$ 的 Lipschitz 常数。对于任意 $\boldsymbol{\theta}_s$，有：

$$
\begin{aligned}
\left| R(\boldsymbol{\theta}_s; \mathbf{h}(\tilde{\boldsymbol{\theta}}_m)) - R(\boldsymbol{\theta}_s; \mathbf{h}(\hat{\boldsymbol{\theta}}_m)) \right| &\le L \|[\boldsymbol{\theta}_s, \tilde{\boldsymbol{\theta}}_m] - [\boldsymbol{\theta}_s, \hat{\boldsymbol{\theta}}_m]\|_2 && (18)\\
&\le L \|\tilde{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}\|_2 && (19)\\
&= L d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*). && (20)
\end{aligned}
$$

应用（20）和引理B.3于（17），可得：

$$
\begin{aligned}
c_2 d(\bar{\boldsymbol{\theta}}_s, \mathcal{S}^*)^\nu &\le 2L d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) + R(\bar{\boldsymbol{\theta}}_s; \mathbf{h}(\hat{\boldsymbol{\theta}}_m)) - R(\tilde{\boldsymbol{\theta}}_s; \mathbf{h}(\hat{\boldsymbol{\theta}}_m)) && (21)\\
&\le 2c_1 \frac{\log N}{\sqrt{N}} + 2L d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) + R_N(\bar{\boldsymbol{\theta}}_s; \mathbf{h}(\hat{\boldsymbol{\theta}}_m)) - R_N(\tilde{\boldsymbol{\theta}}_s; \mathbf{h}(\hat{\boldsymbol{\theta}}_m)) && (22)
\end{aligned}
$$

¯根据（$\boldsymbol{\theta}_s$ 中4）的定义，可得：

$$
\begin{aligned}
R_N(\bar{\boldsymbol{\theta}}_s; \mathbf{h}(\hat{\boldsymbol{\theta}}_m)) + \lambda_s G(\bar{\boldsymbol{\theta}}_s) &\le R_N(\tilde{\boldsymbol{\theta}}_s; \mathbf{h}(\hat{\boldsymbol{\theta}}_m)) + \lambda_s G(\tilde{\boldsymbol{\theta}}_s) && (23)\\
R_N(\bar{\boldsymbol{\theta}}_s; \mathbf{h}(\hat{\boldsymbol{\theta}}_m)) - R_N(\tilde{\boldsymbol{\theta}}_s; \mathbf{h}(\hat{\boldsymbol{\theta}}_m)) &\le \lambda_s G(\tilde{\boldsymbol{\theta}}_s) - \lambda_s G(\bar{\boldsymbol{\theta}}_s) && (24)
\end{aligned}
$$

将（24）代入（22），并注意到正则化项 $G(\cdot)$ 是光滑的，可得：

$$
\begin{aligned}
c_2 \|\tilde{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|_2^\nu &\le 2c_1 \frac{\log N}{\sqrt{N}} + 2L d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) + \lambda_s (G(\tilde{\boldsymbol{\theta}}_s) - G(\bar{\boldsymbol{\theta}}_s)) && (25)\\
&\le 2c_1 \frac{\log N}{\sqrt{N}} + 2L d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) + \lambda_s C \|\tilde{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|_2 && (26)
\end{aligned}
$$

其中 C 是 $G(\cdot)$ 的 Lipschitz 常数。

根据 Young 不等式，我们有：

$$
\begin{aligned}
\lambda_s C \|\tilde{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|_2 &\le \frac{1}{\nu} \left(\frac{(c_2 \nu)^{1/\nu}}{2} \|\tilde{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|_2 \right)^\nu + \frac{\nu - 1}{\nu} \left(\frac{2C}{(2c_2)^{1/\nu}} \lambda_s \right)^{\nu/(\nu-1)} && (27)\\
&= \frac{c_2}{2} \|\tilde{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|_2 + \frac{2(\nu-1) C^{\nu/(\nu-1)}}{\nu (c_2 \nu)^{1/(\nu-1)}} \lambda_s^{\nu/(\nu-1)}. && (28)
\end{aligned}
$$

令 $C_0 = \frac{2}{\frac{(\nu-1}{\nu}} \frac{C^{\nu/(\nu-1)}}{1/(\nu-1)})(c_2\nu)$。将 (28) 代入 (26) 得：

$$
\begin{aligned}
c_2 \|\tilde{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|_2^\nu &\le 2c_1 \frac{\log N}{\sqrt{N}} + 2L d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) + C_0 \lambda_s^{\nu/(\nu-1)} + \frac{c_2}{2} \|\tilde{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|_2^\nu && (29)\\
\frac{c_2}{2} \|\tilde{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|_2^\nu &\le 2c_1 \frac{\log N}{\sqrt{N}} + 2L d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) + C_0 \lambda_s^{\nu/(\nu-1)} && (30)\\
\|\tilde{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|_2 &\le \left(\frac{4c_1}{c_2} \frac{\log N}{\sqrt{N}} + \frac{2C_0}{c_2} \lambda_s^{\nu/(\nu-1)} + \frac{4L}{c_2} d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) \right)^{1/\nu} && (31)
\end{aligned}
$$

请注意，$G(\cdot)$ 可重写为 $G(\cdot) = \sum_k \| \mathbf{U}_s^k \|_2 + \sum_l \|\mathbf{V}_s^l\|_2$。令 $K(\cdot) = \sum_k \| \mathbf{U}_s^k \|_2$ 表示正则化项 $G(\cdot)$ 中显著嵌入分量权重的总和。令 $\phi(\boldsymbol{\theta}_s)$ 表示参数 $\boldsymbol{\theta}_s$，其中所有不显著的嵌入分量权重 $\mathbf{V}_s$ 均设为零。请注意，$R(\boldsymbol{\theta}_s;\mathbf{h}(\boldsymbol{\theta}_m)) \leq R(\boldsymbol{\theta}_s;\mathbf{h}(\boldsymbol{\theta}_m))$，因为 $[\boldsymbol{\theta}_s, \boldsymbol{\theta}_m] \in \mathcal{T}^*$。根据 $_\boldsymbol{\theta}_s$ 的定义以及 $K(\cdot)$ 的光滑性，我们有如下结果：

$$\lambda_s \sum_l \| \mathbf{V}_s^l \|_2 \leq R_N(\phi(\tilde{\boldsymbol{\theta}}_s);\mathbf{h}(\hat{\boldsymbol{\theta}}_m)) - R_N(\bar{\boldsymbol{\theta}}_s;\mathbf{h}(\hat{\boldsymbol{\theta}}_m)) + \lambda_s(K(\phi(\tilde{\boldsymbol{\theta}}_s)) - K(\bar{\boldsymbol{\theta}}_s)) \tag{32}$$

$$\leq 2c_1 \frac{\log N}{\sqrt{N}} + 2Ld(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) + R(\tilde{\boldsymbol{\theta}}_s;\mathbf{h}(\tilde{\boldsymbol{\theta}}_m)) - R(\bar{\boldsymbol{\theta}}_s;\mathbf{h}(\tilde{\boldsymbol{\theta}}_m)) + \lambda_s(K(\tilde{\boldsymbol{\theta}}_s) - K(\bar{\boldsymbol{\theta}}_s)) \tag{33}$$

$$\leq 2c_1 \frac{\log N}{\sqrt{N}} + 2Ld(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) + \lambda_s C \|\tilde{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|_2 \tag{34}$$

$$\sum_l \| \mathbf{V}_s^l \|_2 \leq 2c_1 \frac{\log N}{\lambda_s \sqrt{N}} + \frac{2L}{\lambda_s} d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*) + C d(\bar{\boldsymbol{\theta}}_s, \{\tilde{\boldsymbol{\theta}}_s\}). \tag{35}$$

这就完成了引理的证明B.5。 □

¯为了完成定理 4.6 的证明，我们考察 $\boldsymbol{\theta}_s$ 与参数集合 $\mathcal{S}_\phi^*$之间的距离。注意，根据引理 B.1，$\phi(\tilde{\boldsymbol{\theta}}_s) \in \mathcal{S}_\phi^*$。令 $\mathbf{V}_{\boldsymbol{\theta}_s} = \sum_l \|\mathbf{V}_s^l\|_2$ 表示模型 $\boldsymbol{\theta}_s$ 中非显著嵌入分量权重的总和。

$$d(\bar{\boldsymbol{\theta}}_s, \mathcal{S}_\phi^*) \leq \|\bar{\boldsymbol{\theta}}_s - \phi(\tilde{\boldsymbol{\theta}}_s)\|_2 \tag{36}$$

$$\leq \|\bar{\boldsymbol{\theta}}_s - \tilde{\boldsymbol{\theta}}_s\|_2 + \|\phi(\tilde{\boldsymbol{\theta}}_s) - \tilde{\boldsymbol{\theta}}_s\|_2 \tag{37}$$

$$\leq \|\bar{\boldsymbol{\theta}}_s - \tilde{\boldsymbol{\theta}}_s\|_2 + \| \mathbf{V}_{\tilde{\boldsymbol{\theta}}_s} \|_2 \tag{38}$$

$$\leq \|\bar{\boldsymbol{\theta}}_s - \tilde{\boldsymbol{\theta}}_s\|_2 + \| \mathbf{V}_{\tilde{\boldsymbol{\theta}}_s} + \mathbf{V}_{\bar{\boldsymbol{\theta}}_s} - \mathbf{V}_{\bar{\boldsymbol{\theta}}_s} \|_2 \tag{39}$$

$$\leq \|\bar{\boldsymbol{\theta}}_s - \tilde{\boldsymbol{\theta}}_s\|_2 + \| \mathbf{V}_{\bar{\boldsymbol{\theta}}_s} \|_2 + \| \mathbf{V}_{\tilde{\boldsymbol{\theta}}_s} - \mathbf{V}_{\bar{\boldsymbol{\theta}}_s} \|_2 \tag{40}$$

$$\leq \|\bar{\boldsymbol{\theta}}_s - \tilde{\boldsymbol{\theta}}_s\|_2 + \| \mathbf{V}_{\bar{\boldsymbol{\theta}}_s} \|_2 + C\|\bar{\boldsymbol{\theta}}_s - \tilde{\boldsymbol{\theta}}_s\|_2 \tag{41}$$

定理的证明4.6 通过结合引理B.5、引理B.4和（41）得以完成。

B.4. 定理的证明4.7

接下来，我们证明，求解（5）的参与方能够找到最优解，并将所有不显著的输入层权重置为零。按照 Dinh 和 Ho (2020年) 给出的引理B.3的相同证明方法，我们可以证明以下引理：

引理 B.6. 对于任意 $\delta > 0$，存在 $c_1(\delta) > 0$，使得对于所有 $\boldsymbol{\theta}$、$\boldsymbol{\theta}'$ 以及集合 $\mathcal{K}$：

$$|H_N(\boldsymbol{\theta};\boldsymbol{\theta}';\mathcal{K}) - H(\boldsymbol{\theta};\boldsymbol{\theta}';\mathcal{K})| \leq c_1 \frac{\log N}{\sqrt{N}}$$

以 $1-\delta$ 的概率。

¯令 $\boldsymbol{\theta}_m$ 为求解（5）的参数。我们证明以下引理：

引理 *B.7.* 令 B 为 $H(\cdot)$ 的 *Lipschitz* 常数。令 $[\tilde{\boldsymbol{\theta}}_s^\top, \tilde{\boldsymbol{\theta}}_m^\top]^\top = arg\,min_{\boldsymbol{\Theta}\in\mathcal{T}^*} \|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}\|_2$，其中 $\hat{\boldsymbol{\Theta}}$ 是式（3）中定义的预˜ ˜训练模型。若式（5）中的 $\mathcal{K}_m$ 是 $f(\boldsymbol{\theta}_s;\mathbf{h}(\boldsymbol{\theta}_m))$ 的显著嵌入分量，则对于任意 $\delta > 0$，存在 $C_1(\delta)$ $C_2(\delta)$ $C_3(\delta)$ $C_4(\delta)$ $C_5 > 0$ 以及 $N \geq N_0(\delta)$，使得：

$$d(\bar{\boldsymbol{\theta}}_m, \mathcal{C}^*) \leq d(\bar{\boldsymbol{\theta}}_m, \{\tilde{\boldsymbol{\theta}}_m\}) \leq \left(C_1 \frac{\log N}{\sqrt{N}} + C_2 B d(\hat{\boldsymbol{\theta}}_m, \mathcal{C}^*) + C_3 (\lambda_m)^{\nu/(\nu-1)} \right)^{1/\nu} \tag{42}$$

非显著输入层权重的总和为

$$\sum_l \| \boldsymbol{V}_m^l \|_2 \leq C_4 \frac{\log N}{\lambda_m \sqrt{N}} + \frac{2B}{\lambda_m} d(\hat{\boldsymbol{\theta}}_m, \mathcal{C}^*) + C_5 d(\bar{\boldsymbol{\theta}}_m, \{\tilde{\boldsymbol{\theta}}_m\}) \tag{43}$$

以概率 $1-\delta$。

~ ¯ ~证明。注意，$\{\boldsymbol{\theta}_m\}$ 是 $H(\boldsymbol{\theta}_m;\boldsymbol{\theta}_m;\mathcal{K}_m)$ 的零水平集。由于 $H(\cdot)$ 是解析的，我们可以应用 Łojasiewicz 不等式，如下所示：

$$c_2 d(\bar{\boldsymbol{\theta}}_m,\mathcal{C}^*)^\nu \le c_2 d(\bar{\boldsymbol{\theta}}_m,\{\tilde{\boldsymbol{\theta}}_m\})^\nu \tag{44}$$
$$\le H(\bar{\boldsymbol{\theta}}_m;\tilde{\boldsymbol{\theta}}_m) \tag{45}$$
$$= H(\bar{\boldsymbol{\theta}}_m;\tilde{\boldsymbol{\theta}}_m) - H(\tilde{\boldsymbol{\theta}}_m;\tilde{\boldsymbol{\theta}}_m) \tag{46}$$

其中（46）由以下事实可得：$H(\boldsymbol{\theta}_m;\boldsymbol{\theta}_m)=0$ 对任意 $\boldsymbol{\theta}_m$。

由于 $H(\cdot)$ 是解析的，我们知道 $H(\cdot)$ 是光滑的。令 B 为 $H(\cdot)$ 的 Lipschitz 常数。对于任意 $\boldsymbol{\theta}_m$，我们有：

$$|H(\boldsymbol{\theta}_m;\hat{\boldsymbol{\theta}}_m) - H(\boldsymbol{\theta}_m;\tilde{\boldsymbol{\theta}}_m)| \le B\|\hat{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_m\|_2 \tag{47}$$
$$\le B d(\hat{\boldsymbol{\theta}}_m,\mathcal{C}^*) \tag{48}$$

将（48）和引理B.6应用于（46）：

$$c_2 d(\bar{\boldsymbol{\theta}}_m,\mathcal{C}^*)^\nu \le 2B d(\hat{\boldsymbol{\theta}}_m,\mathcal{C}^*) + H(\bar{\boldsymbol{\theta}}_m;\hat{\boldsymbol{\theta}}_m) - H(\tilde{\boldsymbol{\theta}}_m;\hat{\boldsymbol{\theta}}_m) \tag{49}$$
$$\le 2c_1\frac{\log N}{\sqrt{N}} + 2B d(\hat{\boldsymbol{\theta}}_m,\mathcal{C}^*) + H_N(\bar{\boldsymbol{\theta}}_m;\hat{\boldsymbol{\theta}}_m) - H_N(\tilde{\boldsymbol{\theta}}_m;\hat{\boldsymbol{\theta}}_m) \tag{50}$$

¯根据（中的 $\boldsymbol{\theta}_m$ 定义5）：

$$H_N(\bar{\boldsymbol{\theta}}_m;\hat{\boldsymbol{\theta}}_m) + \lambda_m G(\bar{\boldsymbol{\theta}}_m) \le H_N(\tilde{\boldsymbol{\theta}}_m;\hat{\boldsymbol{\theta}}_m) + \lambda_m G(\tilde{\boldsymbol{\theta}}_m) \tag{51}$$
$$H_N(\bar{\boldsymbol{\theta}}_m;\hat{\boldsymbol{\theta}}_m) - H_N(\tilde{\boldsymbol{\theta}}_m;\hat{\boldsymbol{\theta}}_m) \le \lambda_m(G(\tilde{\boldsymbol{\theta}}_m) - G(\bar{\boldsymbol{\theta}}_m)) \tag{52}$$

将（52）代入（50）：

$$c_2 d(\bar{\boldsymbol{\theta}}_m,\mathcal{C}^*)^\nu \le 2c_1\frac{\log N}{\sqrt{N}} + 2B d(\hat{\boldsymbol{\theta}}_m,\mathcal{C}^*) + \lambda_m(G(\tilde{\boldsymbol{\theta}}_m) - G(\bar{\boldsymbol{\theta}}_m)) \tag{53}$$
$$\le 2c_1\frac{\log N}{\sqrt{N}} + 2B d(\hat{\boldsymbol{\theta}}_m,\mathcal{C}^*) + \lambda_m C\|\tilde{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}_m\|_2 \tag{54}$$

其中 C 是 $G(\cdot)$的 Lipschitz 常数。

根据 Young 不等式：

$$\lambda_m C\|\tilde{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}_m\|_2 \le \frac{1}{\nu}\left(\frac{(c_2\nu)^{1/\nu}}{2}\|\tilde{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}_m\|_2\right)^\nu + \frac{\nu-1}{\nu}\left(\frac{2C}{(2c_2)^{1/\nu}}\lambda_s\right)^{\nu/(\nu-1)} \tag{55}$$
$$= \frac{c_2}{2}\|\tilde{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}_m\|_2 + \frac{2(\nu-1)C^{\nu/(\nu-1)}}{\nu(c_2\nu)^{1/(\nu-1)}}\lambda_m^{\nu/(\nu-1)}. \tag{56}$$

令 $C_0 = {}^{2}\!/_{\nu}{}^{(\nu-1\ C^{\nu/(\nu-1)})(c_2\nu)}{}_{1/(\nu-1)}$。将 (56) 应用于 (54)，可得：

$$c_2\|\tilde{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}_m\|_2^\nu \le 2c_1\frac{\log N}{\sqrt{N}} + 2B d(\hat{\boldsymbol{\theta}}_m,\mathcal{C}^*) + C_0(\lambda_m)^{\nu/(\nu-1)} + \frac{c_2}{2}\|\tilde{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}_m\|_2^\nu \tag{57}$$
$$\frac{c_2}{2}\|\tilde{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}_m\|_2^\nu \le 2c_1\frac{\log N}{\sqrt{N}} + 2B d(\hat{\boldsymbol{\theta}}_m,\mathcal{C}^*) + C_0(\lambda_m)^{\nu/(\nu-1)} \tag{58}$$
$$\frac{c_2}{2}\|\tilde{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}_m\|_2^\nu \le 2c_1\frac{\log N}{\sqrt{N}} + 2B d(\hat{\boldsymbol{\theta}}_m,\mathcal{C}^*) + C_0(\lambda_m)^{\nu/(\nu-1)} \tag{59}$$
$$d(\bar{\boldsymbol{\theta}}_m,\mathcal{C}^*) \le d(\bar{\boldsymbol{\theta}}_m,\{\tilde{\boldsymbol{\theta}}_m\}) \le \left(\frac{4c_1}{c_2}\frac{\log N}{\sqrt{N}} + \frac{4B}{c_2}d(\hat{\boldsymbol{\theta}}_m,\mathcal{C}^*) + \frac{2C_0}{c_2}(\lambda_m)^{\nu/(\nu-1)}\right)^{1/\nu} \tag{60}$$

注意，$G(\cdot)$ 可以重写为 $G(\cdot) = \sum_k \|\mathbf{U}_m^k\|_2 + \sum_l \|\mathbf{V}_m^l\|_2$。令 $K(\cdot) = \sum_k \|\mathbf{U}_m^k\|_2$ 表示正则化项 $G(\cdot)$ 中显著输入层权重之和。令 $\phi(\boldsymbol{\theta}_m)$ 表示参数 $\boldsymbol{\theta}_m$，其中所有非显著输入层权重 $\mathbf{V}_m$均设为零。注意，在我们假设 $\mathcal{K}_m$ 仅包含显著嵌入分量且根据命题 4.4 的条件下，$H(\phi(\boldsymbol{\theta}_m); \boldsymbol{\theta}_m; \mathcal{K}_m) = H(\boldsymbol{\theta}_m; \boldsymbol{\theta}_m; \mathcal{K}_m) = 0$，因为非显著特征不会影响显著嵌入分量。根据 $\boldsymbol{\theta}_m$的定义：

$$\lambda_m \sum_l \|\mathbf{V}_m^l\|_2 \leq H_N(\phi(\tilde{\boldsymbol{\theta}}_m); \hat{\boldsymbol{\theta}}_m; \mathcal{K}_m) - H_N(\bar{\boldsymbol{\theta}}_m; \hat{\boldsymbol{\theta}}_m; \mathcal{K}_m) + \lambda_m(K(\phi(\tilde{\boldsymbol{\theta}}_m)) - K(\bar{\boldsymbol{\theta}}_m)) \tag{61}$$

$$\leq 2c_1 \frac{\log N}{\sqrt{N}} + 2Bd(\hat{\boldsymbol{\theta}}_m, \mathcal{C}^*) + H(\phi(\tilde{\boldsymbol{\theta}}_m); \tilde{\boldsymbol{\theta}}_m; \mathcal{K}_m) - H(\bar{\boldsymbol{\theta}}_m; \tilde{\boldsymbol{\theta}}_m; \mathcal{K}_m) + \lambda_m(K(\tilde{\boldsymbol{\theta}}_m) - K(\bar{\boldsymbol{\theta}}_m)) \tag{62}$$

$$\leq 2c_1 \frac{\log N}{\sqrt{N}} + 2Bd(\hat{\boldsymbol{\theta}}_m, \mathcal{C}^*) + \lambda_m(K(\tilde{\boldsymbol{\theta}}_m) - K(\bar{\boldsymbol{\theta}}_m)) \tag{63}$$

$$\leq 2c_1 \frac{\log N}{\sqrt{N}} + 2Bd(\hat{\boldsymbol{\theta}}_m, \mathcal{C}^*) + \lambda_m C\|\tilde{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}}_m\|_2 \tag{64}$$

$$\sum_l \|\mathbf{V}_m^l\|_2 \leq 2c_1 \frac{\log N}{\lambda_m \sqrt{N}} + \frac{2B}{\lambda_m} d(\hat{\boldsymbol{\theta}}_m, \mathcal{C}^*) + Cd(\bar{\boldsymbol{\theta}}_m, \{\tilde{\boldsymbol{\theta}}_m\}) \tag{65}$$

至此，引理的证明完成B.5。 □

为完成定理4.7的证明，我们考察 $\boldsymbol{\theta}_m$ 到参数集合 $\mathcal{C}_\phi^*$的距离。注意，根据引理B.1，$\phi(\tilde{\boldsymbol{\theta}}_m) \in \mathcal{C}_\phi^*$。令$\mathbf{V}_{\boldsymbol{\theta}_m} = \sum_l \|\mathbf{V}_m^l\|_2$ 表示模型 $\boldsymbol{\theta}_m$中非显著特征权重的总和。

$$d(\bar{\boldsymbol{\theta}}_m, \mathcal{C}_\phi^*) \leq \|\bar{\boldsymbol{\theta}}_m - \phi(\boldsymbol{\theta}_s')\|_2 \tag{66}$$

$$\leq \|\bar{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_m\|_2 + \|\phi(\tilde{\boldsymbol{\theta}}_m) - \tilde{\boldsymbol{\theta}}_m\|_2 \tag{67}$$

$$\leq \|\bar{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_m\|_2 + \|\mathbf{V}_{\tilde{\boldsymbol{\theta}}_m}\|_2 \tag{68}$$

$$\leq \|\bar{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_m\|_2 + \|\mathbf{V}_{\tilde{\boldsymbol{\theta}}_m} + \mathbf{V}_{\bar{\boldsymbol{\theta}}_m} - \mathbf{V}_{\bar{\boldsymbol{\theta}}_m}\|_2 \tag{69}$$

$$\leq \|\bar{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_m\|_2 + \|\mathbf{V}_{\bar{\boldsymbol{\theta}}_m}\|_2 + \|\mathbf{V}_{\tilde{\boldsymbol{\theta}}_m} - \mathbf{V}_{\bar{\boldsymbol{\theta}}_m}\|_2 \tag{70}$$

$$\leq \|\bar{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_m\|_2 + \|\mathbf{V}_{\bar{\boldsymbol{\theta}}_m}\|_2 + C\|\bar{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_m\|_2. \tag{71}$$

请注意：

$$d(\hat{\boldsymbol{\theta}}_m, \mathcal{C}^*) = \|\tilde{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_m\|_2 \leq \|\tilde{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}\|_2 = d(\hat{\boldsymbol{\Theta}}, \mathcal{T}^*). \tag{72}$$

定理4.7 的证明通过结合引理 B.7、(71)、(72) 和引理 B.4 得以完成。

B.5. 扩展至 $M > 1$ 个参与方

命题 *B.8.* 考虑模型 $\boldsymbol{\Theta} = [\boldsymbol{\theta}_s^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_M^\top]^\top$。令 s 和 z 分别为 $f(\boldsymbol{\Theta})$ 的显著特征集和非显著特征集。令 $f(\boldsymbol{\theta}_s; h_1(\boldsymbol{\theta}_1^\diamond); \ldots; h_M(\boldsymbol{\theta}_M^\diamond))$ 的显著嵌入分量集为 s_s。令 $g^j(x_m, r)$ 将输入 x_m^j 替换为值 r。对于每个显著嵌入分量 $k \in s_s$，对所有 $j \in z$ 和 $m \in [M]$，以及任意 $r \in \mathbb{R}$，均有 $h_m(\boldsymbol{\theta}_m; x_m)^k = h_m(\boldsymbol{\theta}_m; g^j(x_m, r))^k$。

证明。 假设对于参与方 m，$\mathbf{h}_m(\boldsymbol{\theta}_m; \mathbf{x}_m)^k \neq \mathbf{h}_m(\boldsymbol{\theta}_m; g^j(\mathbf{x}_m, r))^k$ 对某个显著嵌入分量$k \in \mathbf{s}_s$、非显著特征 $j \in \mathbf{z}$以及 $r \in \mathbb{R}$ 成立。根据我们的假设，且由于分量 k 是显著的，

$$f(\boldsymbol{\theta}_s; \mathbf{h}_1(\boldsymbol{\theta}_1; \mathbf{x}_1); \ldots; \mathbf{h}_m(\boldsymbol{\theta}_m; \mathbf{x}_m); \ldots; \mathbf{h}_M(\boldsymbol{\theta}_M; \mathbf{x}_M)) \neq f(\boldsymbol{\theta}_s; \mathbf{h}_1(\boldsymbol{\theta}_1; \mathbf{x}_1); \ldots; \mathbf{h}_m(\boldsymbol{\theta}_m; g^j(\mathbf{x}_m; r)); \ldots; \mathbf{h}_M(\boldsymbol{\theta}_M; \mathbf{x}_M))$$

对于某个值 $r \in \mathbb{R}$。这与 j 是一个非显著特征这一事实相矛盾。 □

引理 *B.9.* 设 $\tilde{\boldsymbol{\Theta}} = [\tilde{\boldsymbol{\theta}}_s^\top, \tilde{\boldsymbol{\theta}}_1^\top, \ldots, \tilde{\boldsymbol{\theta}}_M^\top]^\top \in \mathcal{T}^*$。设 s 和 z 分别是 $f(\boldsymbol{\Theta}^\diamond)$的显著特征和非显著特征。设 $\mathcal{K}_m$ 是嵌入向量 $h_m(\boldsymbol{\theta}_m)$中，针对 $f(\boldsymbol{\theta}_s; h_1(\boldsymbol{\theta}_1); \ldots; h_M(\boldsymbol{\theta}_M))$ 的显著嵌入分量子集。设 $\boldsymbol{\theta}_m = arg\,min_{\boldsymbol{\theta}'_m} H(\boldsymbol{\theta}'_m; \tilde{\boldsymbol{\theta}}_m; \mathcal{K}_m)$，其中包括所有参与方 m。设 s_{h_m} 和 z_{h_m} 分别是每个参与方 m 在处针对 $e(\boldsymbol{\theta}_m; \boldsymbol{\theta}_m; \mathcal{K}_m)$且参数为 $\boldsymbol{\theta}_m$时的显著特征和非显著特征。那么：

$$[\tilde{\boldsymbol{\theta}}_s, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M] \in \mathcal{T}^*, \ \bigcup_{m=1}^{M} \boldsymbol{s}_{h_m} = \boldsymbol{s}, \ and \ \bigcup_{m=1}^{M} \boldsymbol{z}_{h_m} = \boldsymbol{z}\,.$$

证明。注意，对 $H(\boldsymbol{\theta}_m; \boldsymbol{\theta}_m; \mathcal{K}_m)$ 进行最小化会导致 $\mathbf{h}_m(\boldsymbol{\theta}_m; \mathbf{x}_m)_i = \mathbf{h}_m(\boldsymbol{\theta}_m; \mathbf{x}_m)_i$ 对所有显著嵌入分量 $i \in \mathcal{K}_m$ 以及任意输入 $\mathbf{x}_m$成立。根据 $\mathcal{T}^*$ 的定义和定义 2.1，这意味着：

$$R(\boldsymbol{\theta}_s^\diamond; \mathbf{h}_1(\boldsymbol{\theta}_1^\diamond); \ldots; \mathbf{h}_M(\boldsymbol{\theta}_M^\diamond)) = R(\tilde{\boldsymbol{\theta}}_s; \mathbf{h}_1(\tilde{\boldsymbol{\theta}}_1); \ldots; \mathbf{h}_M(\tilde{\boldsymbol{\theta}}_M)) = R(\tilde{\boldsymbol{\theta}}_s; \mathbf{h}_1(\boldsymbol{\theta}_1); \ldots; \mathbf{h}_M(\boldsymbol{\theta}_M)).$$

因此，$[\boldsymbol{\theta}_s, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M] \in \mathcal{T}^*$。

根据（Dinh 和 Ho，2020）中的引理 3.1，由于 $\boldsymbol{\Theta} \in \mathcal{T}^*$，$f(\boldsymbol{\Theta}; \mathrm{x}; y) = f(\boldsymbol{\Theta}^\diamond; \mathrm{x}; y)$ 对所有输入 x 均成立。这意味着，对于 $f(\boldsymbol{\Theta}^\diamond)$，显著特征和非显著特征必须与 $f(\boldsymbol{\Theta})$ 中的相同。令 s 和 z 分别表示 $f(\boldsymbol{\Theta})$ 中的显著特征和非显著特征。必然有 s = s 且 z = z。

令 $j \in \mathbf{z}_m$ 为参与方 m 在 $f(\tilde{\boldsymbol{\Theta}})$ 中的一个非显著特征，并令 $r \in \mathbb{R}$。令 $g^j(\cdot)$ 的定义与定义 2.1相同。我们知道，对于所有 $k \in \mathcal{K}_m$：

$$\mathbf{h}_m(\boldsymbol{\theta}_m; g^j(\mathbf{x}_m, r))^k = \mathbf{h}_m(\tilde{\boldsymbol{\theta}}_m; g^j(\mathbf{x}_m, r))^k = \mathbf{h}_m(\tilde{\boldsymbol{\theta}}_m; \mathbf{x}_m)^k$$

因为根据命题 4.4，$\mathcal{K}_m$ 中的所有嵌入分量仅取决于 s_m。由于 $\mathrm{h}_m(\boldsymbol{\theta}_m; \cdot)^k$ 不受 $\mathbf{z}_m$中的特征影响，这意味着，$e(\boldsymbol{\theta}_m; \boldsymbol{\theta}_m; \mathcal{K}_m)$的非显著特征集合包含参与方 m处 $f(\boldsymbol{\Theta})$的非显著特征集合：$\mathbf{z}_{h_m} \supseteq \mathbf{z}_m$。

类似地，令 $k \in \mathbf{s}_m$ 为参与方 m 在 $f(\boldsymbol{\Theta})$中的一个显著特征。根据命题 B.8，对于某个 $k \in \mathcal{K}_m$：

$$\mathbf{h}_m(\boldsymbol{\theta}_m; g^k(\mathbf{x}_m, r))^k = \mathbf{h}_m(\tilde{\boldsymbol{\theta}}_m; g^k(\mathbf{x}_m, r))^k \neq \mathbf{h}_m(\tilde{\boldsymbol{\theta}}_m; \mathbf{x})^k$$

对于某个 $r \in \mathbb{R}$。这意味着：

$$\sum_{k \in \mathcal{K}_m} (\mathbf{h}_m(\boldsymbol{\theta}_m; g^k(\mathbf{x}_m, r))^k - \mathbf{h}_m(\tilde{\boldsymbol{\theta}}_m; \mathbf{x}_m)^k)^2 \neq \sum_{k \in \mathcal{K}_m} (\mathbf{h}_m(\boldsymbol{\theta}_m; \mathbf{x}_m)^k - \mathbf{h}_m(\tilde{\boldsymbol{\theta}}_m; \mathbf{x}_m)^k)^2$$

因此，我们可以说，$e(\boldsymbol{\theta}_m; \boldsymbol{\theta}_m; \mathcal{K}_m)$的显著特征集合包含参与方 m处 $f(\boldsymbol{\Theta})$的显著特征集合：$\mathbf{s}_{h_m} \supseteq \mathbf{s}_m$。

由于对于每个参与方 m，$\mathrm{s}_{h_m} \cap \mathrm{z}_{h_m} = \emptyset$，$\mathrm{s}_{h_m} = \mathrm{s}_m$ 且 $\mathrm{z}_{h_m} = \mathrm{z}_m$。由于对于参与方 $m \neq j$，$\mathrm{s}_m \cap \mathrm{s}'_j = \emptyset$ 且 $\mathrm{z}_m \cap \mathrm{z}'_j = \emptyset$，$\bigcup_{m=1}^{M} \mathrm{s}_{h_m} = \mathrm{s}$ 且 $\bigcup_{m=1}^{M} \mathrm{z}_{h_m} = \mathrm{z}$。引理 B.9 的证明至此完成。 □

我们重新定义 $\mathcal{S}_\phi^*$ for $M > 1$:

$$\mathcal{S}_\phi^* = \{\boldsymbol{\theta}_s : \exists\, \boldsymbol{\theta}_m \ \forall m = 1, \ldots, M \text{ s.t. } [\boldsymbol{\theta}_s, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M] \in \mathcal{T}^* \text{ and } \mathbf{V}_s = \mathbf{0}\}.$$

我们将在以下定理中界定距离 $d(\boldsymbol{\theta}_s, \mathcal{S}_\phi^*)$。

定理 B.10. 令 L 为 $f(\cdot)$ 的 *Lipschitz* 常数。给定预训练模型 $\hat{\boldsymbol{\Theta}}$，其定义为 (3)，对于任意 $\delta > 0$，存在 C_N，$C_\delta(\delta) > 0$ 和 $N \geq N_0(\delta)$ 使得：

$$d(\boldsymbol{\theta}_s, \mathcal{S}_\phi^*) \leq LC_N \frac{\log N}{\lambda_s \sqrt{N}} + LC_\delta \left(\frac{\log N}{\sqrt{N}} + \lambda_s^{\nu/(\nu-1)} \right)^{1/\nu} \tag{73}$$

其概率为 $1 - \delta$。如果 $\lambda_s \sim N^{-1/4}$，则以 $1 - \delta$ 的概率存在 $C(\delta) > 0$，使得：

$$d(\boldsymbol{\theta}_s, \mathcal{S}_\phi^*) \leq LC \left(\frac{\log N}{N} \right)^{\frac{1}{4(\nu-1)}} \tag{74}$$

证明。 定理 B.10 的证明与附录 4.6 中定理 B.3 的证明相同，只需将 $R(\boldsymbol{\theta}_s;\ \mathbf{h}(\boldsymbol{\theta}_m))$替换为 $R(\boldsymbol{\theta}_s;\ \mathbf{h}_1(\boldsymbol{\theta}_1);\ .....\mathbf{h}_M(\boldsymbol{\theta}_M))$。 □

我们将参与方 m 在 $\mathcal{T}^*$ 中的参数集合定义为：其中局部非显著特征对应的权重均设为零。

$$\mathcal{C}_m^* = \{\boldsymbol{\theta}_m : \exists\, \boldsymbol{\theta}_s \text{ and } \boldsymbol{\theta}_j\ \forall j \neq m \text{ s.t. } [\boldsymbol{\theta}_s, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m, \ldots, \boldsymbol{\theta}_M] \in \mathcal{T}^* \text{ and } \mathbf{V}_m = \mathbf{0}\}.$$

我们在以下定理中界定距离 $d(\boldsymbol{\theta}_m, \mathcal{C}_m^*) \to 0$ 。

定理 *B.11.* 设 B 为 $H(\cdot)$的 *Lipschitz* 常数。令 $[\tilde{\boldsymbol{\theta}}_s^\top, \tilde{\boldsymbol{\theta}}_1^\top, \ldots, \tilde{\boldsymbol{\theta}}_M^\top]^\top = arg\,min_{\boldsymbol{\Theta}\in\mathcal{T}^*} \|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}\|_2$，其中 $\hat{\boldsymbol{\Theta}}$ 是式 *(3)* 中定义的预训练模型。若式 *(5)* 中的 $\mathcal{K}_m$ 是 $h_m(\boldsymbol{\theta}_m)$ 中 $f(\boldsymbol{\theta}_s; h_1(\boldsymbol{\theta}_1); \ldots; h_M(\boldsymbol{\theta}_M))$ 的显著嵌入分量子集，则对于每个参与方 m，对于任意 $\delta > 0$，均存在 $C_N, C_\delta(\delta) > 0$和 $N \geq N_0(\delta)$，使得：

$$d(\boldsymbol{\theta}_m, \mathcal{C}_m^*) \leq BC_N \frac{\log N}{\lambda_m \sqrt{N}} + C_\delta \left(B \frac{\log N}{\sqrt{N}} + (\lambda_m)^{\nu/(\nu-1)} \right)^{1/\nu} \tag{75}$$

概率为 $1-\delta$。若 $\lambda_m \sim N^{-1/4}$，则以概率 $1-\delta$ 存在 $C_m(\delta) > 0$ 使得：

$$d(\boldsymbol{\theta}_m, \mathcal{C}_m^*) \leq BC_m \left(\frac{\log N}{N} \right)^{\frac{1}{4(\nu-1)}} \tag{76}$$

定理B.11 可通过对每个参与方应用定理4.7 的证明 m，将 $\mathcal{C}_\phi^*$ 替换为 $\mathcal{C}_m^*$即可。

B.6. 定理4.3的证明

令常数 C 如定理B.10 中所定义，并令常数 C_m 按定理B.11 中的相同方式对所有参与方 m定义。令 B_m 为 $H(\cdot)$ 在参与方 m处的 Lipschitz 常数。由定理B.10 和B.11可知，以概率 $1-\delta$:

$$d(\bar{\boldsymbol{\Theta}}, \mathcal{T}_\phi^*) = \sqrt{d(\bar{\boldsymbol{\theta}}_s, \mathcal{S}_\phi^*)^2 + d(\bar{\boldsymbol{\theta}}_1, \mathcal{C}_1^\diamond)^2 + \ldots + d(\bar{\boldsymbol{\theta}}_M, \mathcal{C}_m^*)^2} \tag{77}$$

$$\leq \sqrt{L^2C^2 \left(\frac{\log N}{N} \right)^{\frac{1}{2(\nu-1)}} + \left(\frac{\log N}{N} \right)^{\frac{1}{2(\nu-1)}} \sum_{m=1}^{M} B_m^2 C_m^2} \tag{78}$$

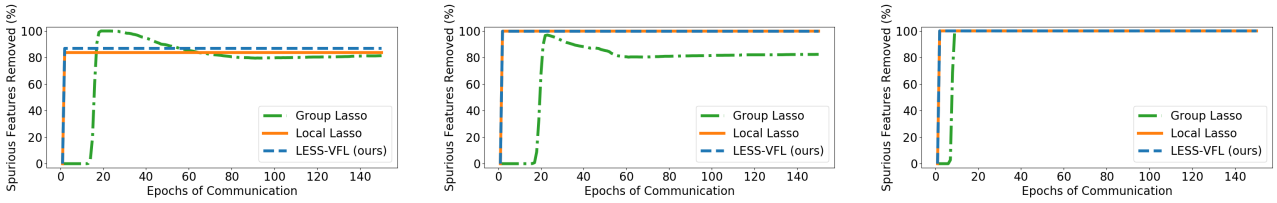
$$\leq \sqrt{\left(L^2C^2 + \sum_{m=1}^{M} B_m^2 C_m^2 \right)} \left(\frac{\log N}{N} \right)^{\frac{1}{4(\nu-1)}} \tag{79}$$

$$= O\left(\sqrt{M} \left(\frac{\log N}{N} \right)^{\frac{1}{4(\nu-1)}} \right). \tag{80}$$

C. 补充实验结果

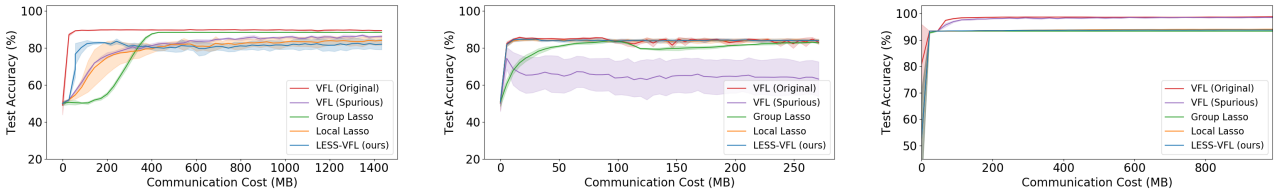
现在我们提供额外的实验结果。我们采用与第5节所述相同的实验设置，并给出此前未纳入的数据集（ MIMIC-III、Gina、Sylva）的结果。我们还提供网格搜索的完整结果，展示组套索、局部套索和 LESS-VFL 在不同正则化参数下移除的伪特征百分比和最终训练准确率。

在图5中，我们绘制了在 MIMIC-III、Gina 和 Sylva 数据集上训练的 150 个通信周期内移除的伪特征百分比。对于 MIMIC-III 和 Sylva，我们可以看到，所有方法在快速移除伪特征方面的表现都相似，不过组套索比其他方法落后几个通信轮次。对于 Gina，组套索需要额外约 20个通信周期才开始移除伪特征，并且在稳定下来之前有所波动，最终移除的比例低于其他方法。正如正文中的主要结论所强调的，通过允许特征选择在几乎无需前期通信的情况下进行，与组套索相比，可以用更少的通信轮次移除伪特征。



(a) MIMIC-III (b) Gina (c) Sylva

图 *5.* 在 150 个通信周期内移除的伪特征百分比。所示数值为 5 次运行的平均值。组套索逐步移除特征，而局部套索和 LESS-VFL 则在预训练后通过一轮通信移除特征。



(a) MIMIC-III (b) Gina (c) Sylva

图 *6.* 前 50 个通信周期内的测试准确率。实线表示 5 次运行的平均值，阴影区域表示标准差。

在图6中，我们绘制了所有基线方法的测试准确率与通信成本之间的关系。对于 MIMIC-III 和 Sylva，加入伪特征不会对纵向联邦学习的测试准确率造成明显的不利影响。在这种情况下，应用特征选择方法时，重要的是模型性能不能因此低于未移除任何伪特征时的表现。对于 MIMIC-III，所有方法的测试准确率相近。然而，对于 Sylva 数据集，在最初的 50个通信周期内，组套索无法达到其他方法相同的准确率。对于 Gina 数据集，所有特征选择方法的测试准确率都与不含伪特征的纵向联邦学习基线相近，不过组套索需要更多通信轮次才能收敛。

在表格5中，我们给出了在不同预训练轮数下，达到基线纵向联邦学习（VFL（原始））测试准确率的90% 并移除80%伪特征所需的通信成本。我们展示了预训练期间和特征选择后（Post-FS）产生的通信成本，以及总通信成本。所示数值为五次运行的平均值，$\pm$ 标准差。可以看到，在许多情况下，LESS-VFL 在特征选择后没有通信成本。这表明 LESS-VFL 在特征选择过程中移除了伪特征并达到了高准确率。可以看到，LESS-VFL 的通信成本始终不高于局部套索。此外，我们发现，在钓鱼数据集上，局部套索需要更多预训练轮数，才能以最低通信成本达到阈值。在 Activity 数据集中，当预训练轮数介于1 和 5之间时，LESS-VFL 的通信成本始终低于局部套索。局部套索的最低通信成本为26.56 MB，而 LESS-VFL 的最高通信成本为21.88MB。

在表 6、7 和 8中，我们给出了网格搜索的结果，用于确定每种方法的最佳正则化参数。我们给出了使用不同正则化参数时 group lasso、local lasso 和 LESS-VFL 的最终训练准确率及已移除的伪特征百分比：$(\lambda_m,\ \lambda_s)$。请注意，服务器正则化参数 λ_s 仅适用于 LESS-VFL。所示数值为五次运行的平均值，$\pm$ 标准差。

表 *5.* 达到基线 VFL（原始）测试准确率 90% 并移除 80% 伪特征所需的通信成本（不同预训练轮数）。所有数值均为 5 次运行的平均值。粗体数值表示该方法在该数据集上实现的最低通信成本。“预训练”指预训练期间的通信成本，“特征选择后”指特征选择完成后训练期间的通信成本，“总计”是前两者之和。

数据集	**预训练轮数**	**通信成本（MB）**					
		局部套索			LESS-VFL（我们的方法）		
		预训练	特征选择后	总计	预训练	特征选择后	总计
活动	1	3.59	30.15	33.74	3.59	15.97	19.56
	2	5.38	26.56	31.95	5.38	16.49	21.88
	3	7.18	19.39	**26.56**	7.18	10.90	18.08
	4	8.97	21.54	30.51	8.97	7.70	**16.68**
	5	10.77	18.85	29.62	10.77	10.51	21.28
网络钓鱼	1	3.24	4.86	8.10	3.24	0.75	**3.99**
	2	4.86	1.62	**6.48**	4.86	0.38	5.23
	3	6.48	1.62	8.10	6.48	0.00	6.48
	4	8.10	1.62	9.72	8.10	0.00	8.10
	5	9.72	1.62	11.34	9.72	0.00	9.72
MIMIC-III	1	7.17	0.00	**7.17**	7.17	0.00	**7.17**
	2	10.75	0.00	10.75	10.75	0.00	10.75
	3	14.34	0.00	14.34	14.34	0.00	14.34
	4	17.92	0.00	17.92	17.92	0.00	17.92
	5	21.51	0.00	21.51	21.51	0.00	21.51
Gina	1	1.35	0.54	**1.90**	1.35	0.13	**1.48**
	2	2.03	0.00	2.03	2.03	0.00	2.03
	3	2.71	0.00	2.71	2.71	0.00	2.71
	4	3.39	0.00	3.39	3.39	0.00	3.39
	5	4.06	0.00	4.06	4.06	0.00	4.06
Sylva	1	5.62	0.00	**5.62**	5.62	0.00	**5.62**
	2	8.43	0.00	8.43	8.43	0.00	8.43
	3	11.25	0.00	11.25	11.25	0.00	11.25
	4	14.06	0.00	14.06	14.06	0.00	14.06
	5	16.87	0.00	16.87	16.87	0.00	16.87

表 6. 活动和网络钓鱼数据集的训练准确率及伪特征移除百分比。

数据集	正则化项系数 (λ_m, λ_s)	组套索		局部套索		LESS-VFL（我们的方法）	
		最终准确率	伪特征移除	最终准确率	伪特征移除	最终准确率	伪特征移除
活动	(2。0，0。5)	18.22 ± 0.00	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	47.22 ± 2.49	100.00 ± 0.00
	(2.0, 0.25)	18.22 ± 0.00	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	69.38 ± 1.75	100.00 ± 0.00
	(2.0, 0.1)	18.22 ± 0.00	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	73.53 ± 0.71	100.00 ± 0.00
	(2.0, 0.05)	18.22 ± 0.00	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	74.38 ± 0.61	100.00 ± 0.00
	(1.0, 0.5)	18.22 ± 0.00	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	50.00 ± 0.29	100.00 ± 0.00
	(1。0，0。25)	18.22 ± 0.00	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	69.38 ± 1.75	100.00 ± 0.00
	(1。0，0。1)	18.22 ± 0.00	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	73.53 ± 0.71	100.00 ± 0.00
	(1。0，0。05)	18.22 ± 0.00	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	74.38 ± 0.61	100.00 ± 0.00
	(0.5, 0.5)	25.37 ± 8.79	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	45.95 ± 8.65	99.46 ± 0.54
	(0.5, 0.25)	25.37 ± 8.79	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	72.96 ± 2.34	100.00 ± 0.00
	(0。5，0。1)	25.37 ± 8.79	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	73.95 ± 1.02	100.00 ± 0.00
	(0.5, 0.05)	25.37 ± 8.79	100.00 ± 0.00	74.18 ± 0.77	100.00 ± 0.00	74.38 ± 0.61	100.00 ± 0.00
	(0.25, 0.5)	57.10 ± 1.74	100.00 ± 0.00	73.72 ± 4.78	100.00 ± 0.00	45.54 ± 7.09	91.90 ± 3.76
	(0.25, 0.25)	57.10 ± 1.74	100.00 ± 0.00	73.72 ± 4.78	100.00 ± 0.00	78.66 ± 3.40	100.00 ± 0.00
	(0.25, 0.1)	57.10 ± 1.74	100.00 ± 0.00	73.72 ± 4.78	100.00 ± 0.00	74.02 ± 7.81	100.00 ± 0.00
	(0.25, 0.05)	57.10 ± 1.74	100.00 ± 0.00	73.72 ± 4.78	100.00 ± 0.00	73.73 ± 5.01	100.00 ± 0.00
	(0.1, 0.5)	75.47 ± 1.93	88.93 ± 3.06	86.75 ± 2.04	100.00 ± 0.00	49.68 ± 4.03	59.05 ± 7.24
	(0.1, 0.25)	75.47 ± 1.93	88.93 ± 3.06	86.75 ± 2.04	100.00 ± 0.00	86.70 ± 3.13	87.93 ± 8.91
	(0。1，0。1)	75.47 ± 1.93	88.93 ± 3.06	86.75 ± 2.04	100.00 ± 0.00	88.31 ± 0.74	99.64 ± 0.71
	(0.1, 0.05)	75.47 ± 1.93	88.93 ± 3.06	86.75 ± 2.04	100.00 ± 0.00	87.14 ± 1.86	99.93 ± 0.14
	(0.05, 0.5)	89.98 ± 2.60	1.71 ± 0.61	90.17 ± 2.02	0.64 ± 0.35	43.91 ± 5.10	0.48 ± 0.34
	(0.05, 0.25)	89.98 ± 2.60	1.71 ± 0.61	90.17 ± 2.02	0.64 ± 0.35	89.13 ± 2.36	0.50 ± 0.36
	(0.05，0.1)	89.98 ± 2.60	1.71 ± 0.61	90.17 ± 2.02	0.64 ± 0.35	87.75 ± 2.18	0.57 ± 0.43
	(0.05, 0.05)	89.98 ± 2.60	1.71 ± 0.61	90.17 ± 2.02	0.64 ± 0.35	91.11 ± 1.58	0.64 ± 0.35
	(0.01, 0.5)	89.05 ± 2.24	0.00 ± 0.00	90.45 ± 1.12	0.00 ± 0.00	43.91 ± 5.10	0.00 ± 0.00
	(0.01, 0.25)	89.05 ± 2.24	0.00 ± 0.00	90.45 ± 1.12	0.00 ± 0.00	90.15 ± 1.86	0.00 ± 0.00
	(0.01, 0.1)	89.05 ± 2.24	0.00 ± 0.00	90.45 ± 1.12	0.00 ± 0.00	89.39 ± 1.61	0.00 ± 0.00
	（0.01, 0.05)	89.05 ± 2.24	0.00 ± 0.00	90.45 ± 1.12	0.00 ± 0.00	87.97 ± 2.30	0.00 ± 0.00
网络钓鱼	(2.0, 0.01)	55.63 ± 0.00	100.00 ± 0.00	90.27 ± 0.67	100.00 ± 0.00	53.38 ± 4.50	100.00 ± 0.00
	(2.0, 0.005)	55.63 ± 0.00	100.00 ± 0.00	90.27 ± 0.67	100.00 ± 0.00	53.38 ± 4.50	100.00 ± 0.00
	(1.0, 0.01)	55.63 ± 0.00	100.00 ± 0.00	90.27 ± 0.67	100.00 ± 0.00	53.38 ± 4.50	100.00 ± 0.00
	(1.0, 0.005)	55.63 ± 0.00	100.00 ± 0.00	90.27 ± 0.67	100.00 ± 0.00	53.38 ± 4.50	100.00 ± 0.00
	(0.5, 0.01)	55.63 ± 0.00	100.00 ± 0.00	90.27 ± 0.67	100.00 ± 0.00	53.38 ± 4.50	100.00 ± 0.00
	(0.5, 0.005)	55.63 ± 0.00	100.00 ± 0.00	90.27 ± 0.67	100.00 ± 0.00	53.38 ± 4.50	100.00 ± 0.00
	(0.25, 0.01)	89.26 ± 1.96	100.00 ± 0.00	90.27 ± 0.67	100.00 ± 0.00	53.38 ± 4.50	100.00 ± 0.00
	(0.25, 0.005)	89.26 ± 1.96	100.00 ± 0.00	90.27 ± 0.67	100.00 ± 0.00	51.13 ± 5.52	100.00 ± 0.00
	(0。1，0。01)	91.71 ± 0.20	84.00 ± 5.33	78.96 ± 15.00	92.00 ± 9.80	78.98 ± 15.01	90.67 ± 9.04
	(0.1, 0.005)	91.71 ± 0.20	84.00 ± 5.33	78.96 ± 15.00	92.00 ± 9.80	92.45 ± 0.00	93.33 ± 0.00
	(0.05, 0.01)	91.85 ± 0.43	0.00 ± 0.00	92.06 ± 0.11	0.00 ± 0.00	91.92 ± 0.11	0.00 ± 0.00
	(0.05, 0.005)	91.85 ± 0.43	0.00 ± 0.00	92.06 ± 0.11	0.00 ± 0.00	92.06 ± 0.12	0.00 ± 0.00
	(0.01, 0.01)	91.86 ± 0.21	0.00 ± 0.00	92.00 ± 0.11	0.00 ± 0.00	92.00 ± 0.12	0.00 ± 0.00
	(0.01, 0.005)	91.86 ± 0.21	0.00 ± 0.00	92.00 ± 0.11	0.00 ± 0.00	91.88 ± 0.26	0.00 ± 0.00

表 7。MIMIC-III 数据集中的训练准确率和伪特征移除百分比。‘–’表示未运行使用此正则化参数选项的实验。

数据集	正则化项系数 (λ_m, λ_s)	组套索		局部套索		LESS-VFL（我们的方法）	
		最终准确率	伪特征移除	最终准确率	伪特征移除	最终准确率	伪特征移除
MIMIC-III	(40.0, 0.5)	–	–	81.97 ± 2.64	100.00 ± 0.00	81.12 ± 2.08	99.89 ± 0.22
	(40.0, 0.25)	–	–	81.97 ± 2.64	100.00 ± 0.00	80.60 ± 1.59	100.00 ± 0.00
	(40.0, 0.1)	–	–	81.97 ± 2.64	100.00 ± 0.00	80.61 ± 2.62	100.00 ± 0.00
	(40.0, 0.05)	–	–	81.97 ± 2.64	100.00 ± 0.00	80.82 ± 1.44	99.21 ± 1.57
	(35。0, 0。5)	–	–	80.94 ± 3.45	98.60 ± 2.81	81.28 ± 1.89	96.07 ± 5.68
	(35。0, 0。25)	–	–	80.94 ± 3.45	98.60 ± 2.81	80.96 ± 1.92	98.60 ± 2.81
	(35.0, 0.1)	–	–	80.94 ± 3.45	98.60 ± 2.81	81.79 ± 1.03	98.43 ± 3.15
	(35。0, 0。05)	–	–	80.94 ± 3.45	98.60 ± 2.81	80.23 ± 2.09	98.71 ± 2.58
	(32.5, 0.1)	–	–	83.45 ± 2.47	87.53 ± 10.81	81.85 ± 2.66	89.94 ± 10.14
	(32.5, 0.05)	–	–	83.45 ± 2.47	87.53 ± 10.81	82.56 ± 1.37	86.85 ± 10.21
	(30.0, 0.5)	–	–	84.45 ± 2.08	76.24 ± 6.35	84.39 ± 2.84	66.15 ± 13.21
	(30.0, 0.25)	–	–	84.45 ± 2.08	76.24 ± 6.35	84.26 ± 2.26	76.12 ± 4.97
	(30.0, 0.1)	–	–	84.45 ± 2.08	76.24 ± 6.35	85.21 ± 1.62	74.38 ± 5.37
	(30。0, 0。05)	–	–	84.45 ± 2.08	76.24 ± 6.35	83.61 ± 2.54	78.30 ± 5.57
	(25。0, 0。5)	–	–	85.62 ± 1.24	55.22 ± 5.03	87.53 ± 1.19	53.43 ± 3.23
	(25.0, 0.25)	–	–	85.62 ± 1.24	55.22 ± 5.03	87.82 ± 0.65	53.54 ± 1.39
	(25。0, 0。1)	–	–	85.62 ± 1.24	55.22 ± 5.03	87.21 ± 1.45	56.07 ± 5.01
	(25.0, 0.05)	–	–	85.62 ± 1.24	55.22 ± 5.03	85.73 ± 1.13	55.11 ± 5.03
	(20。0, 0。5)	–	–	85.51 ± 1.34	44.61 ± 2.23	87.58 ± 0.53	46.40 ± 1.90
	(20.0, 0.25)	–	–	85.51 ± 1.34	44.61 ± 2.23	88.01 ± 0.40	44.94 ± 1.51
	(20.0, 0.1)	–	–	85.51 ± 1.34	44.61 ± 2.23	87.78 ± 0.66	43.71 ± 1.59
	(20.0, 0.05)	–	–	85.51 ± 1.34	44.61 ± 2.23	87.50 ± 0.54	45.51 ± 1.48
	(15.0, 0.5)	–	–	85.19 ± 0.86	32.25 ± 5.01	87.39 ± 0.79	29.72 ± 3.10
	(15.0, 0.25)	–	–	85.19 ± 0.86	32.25 ± 5.01	86.92 ± 0.76	31.40 ± 4.34
	(15.0, 0.1)	–	–	85.19 ± 0.86	32.25 ± 5.01	86.67 ± 0.75	31.63 ± 3.76
	(15.0, 0.05)	–	–	85.19 ± 0.86	32.25 ± 5.01	86.48 ± 1.32	30.73 ± 2.98
	(10。0, 0。5)	–	–	83.92 ± 1.70	8.99 ± 2.20	86.37 ± 2.26	9.89 ± 1.66
	(10.0, 0.25)	–	–	83.92 ± 1.70	8.99 ± 2.20	87.50 ± 0.83	8.26 ± 1.52
	(10.0, 0.1)	–	–	83.92 ± 1.70	8.99 ± 2.20	84.82 ± 2.11	9.72 ± 3.01
	(10.0, 0.05)	–	–	83.92 ± 1.70	8.99 ± 2.20	85.65 ± 0.86	8.88 ± 1.99
	(2。0, 0。5)	80.21 ± 1.25	100.00 ± 0.00	–	–	–	–
	(2.0, 0.25)	80.21 ± 1.25	100.00 ± 0.00	–	–	–	–
	(2.0, 0.1)	80.21 ± 1.25	100.00 ± 0.00	–	–	–	–
	(2.0, 0.05)	80.21 ± 1.25	100.00 ± 0.00	–	–	–	–
	(1.0, 0.5)	83.88 ± 2.24	98.93 ± 0.63	–	–	–	–
	(1。0, 0。25)	83.88 ± 2.24	98.93 ± 0.63	–	–	–	–
	(1.0, 0.1)	83.88 ± 2.24	98.93 ± 0.63	–	–	–	–
	(1。0, 0。05)	83.88 ± 2.24	98.93 ± 0.63	–	–	–	–
	(0.5, 0.5)	87.55 ± 0.74	93.37 ± 0.98	–	–	–	–
	(0.5, 0.25)	87.55 ± 0.74	93.37 ± 0.98	–	–	–	–
	(0。5, 0。1)	87.55 ± 0.74	93.37 ± 0.98	–	–	–	–
	(0.5, 0.05)	87.55 ± 0.74	93.37 ± 0.98	–	–	–	–
	(0.25, 0.5)	87.64 ± 0.92	81.35 ± 3.42	–	–	–	–
	(0.25, 0.25)	87.64 ± 0.92	81.35 ± 3.42	–	–	–	–
	(0.25, 0.1)	87.64 ± 0.92	81.35 ± 3.42	–	–	–	–
	(0.25, 0.05)	87.64 ± 0.92	81.35 ± 3.42	–	–	–	–
	(0.1, 0.5)	87.63 ± 0.47	60.84 ± 4.70	–	–	–	–
	(0.1, 0.25)	87.63 ± 0.47	60.84 ± 4.70	–	–	–	–
	(0。1, 0。1)	87.63 ± 0.47	60.84 ± 4.70	–	–	–	–
	(0.1, 0.05)	87.63 ± 0.47	60.84 ± 4.70	–	–	–	–
	(0.05, 0.5)	86.01 ± 1.41	43.99 ± 7.25	–	–	–	–
	(0.05, 0.25)	86.01 ± 1.41	43.99 ± 7.25	–	–	–	–
	(0.05, 0.1)	86.01 ± 1.41	43.99 ± 7.25	–	–	–	–
	(0.05, 0.05)	86.01 ± 1.41	43.99 ± 7.25	–	–	–	–
	(0.01, 0.5)	85.02 ± 0.84	0.00 ± 0.00	–	–	–	–
	(0.01, 0.25)	85.02 ± 0.84	0.00 ± 0.00	–	–	–	–
	(0.01, 0.1)	85.02 ± 0.84	0.00 ± 0.00	–	–	–	–
	(0.01, 0.05)	85.02 ± 0.84	0.00 ± 0.00	–	–	–	–

表 *8.* Gina 和 Sylva 数据集的训练准确率及伪特征移除百分比。‘–’表示未针对该正则化参数选择运行实验。

数据集	正则化项系数 (λ_m, λ_s)	组套索		局部套索		LESS-VFL（我们的方法）	
		最终准确率	伪特征移除	最终准确率	伪特征移除	最终准确率	伪特征移除
Gina	(2.0, 0.1)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	80.84 ± 0.00	100.00 ± 0.00
	(2.0, 0.05)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.00 ± 1.15	100.00 ± 0.00
	(2.0, 0.01)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.46 ± 1.25	100.00 ± 0.00
	(2.0, 0.005)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.37 ± 1.34	100.00 ± 0.00
	(1.0, 0.1)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	80.98 ± 0.00	100.00 ± 0.00
	(1。0，0。05)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.00 ± 1.15	100.00 ± 0.00
	(1.0, 0.01)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.46 ± 1.25	100.00 ± 0.00
	(1.0, 0.005)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.37 ± 1.34	100.00 ± 0.00
	(0。5，0。1)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	80.98 ± 0.00	100.00 ± 0.00
	(0.5, 0.05)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.00 ± 1.15	100.00 ± 0.00
	(0.5, 0.01)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.46 ± 1.25	100.00 ± 0.00
	(0.5, 0.005)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.37 ± 1.34	100.00 ± 0.00
	(0.25, 0.1)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	80.98 ± 0.00	99.59 ± 0.00
	(0.25, 0.05)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.00 ± 1.15	99.79 ± 0.21
	(0.25, 0.01)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.46 ± 1.25	100.00 ± 0.00
	(0.25, 0.005)	50.43 ± 0.00	100.00 ± 0.00	83.46 ± 1.10	100.00 ± 0.00	83.37 ± 1.34	100.00 ± 0.00
	(0。1，0。1)	81.16 ± 1.12	82.33 ± 0.47	83.46 ± 1.10	100.00 ± 0.00	81.99 ± 0.00	81.20 ± 0.00
	(0.1, 0.05)	81.16 ± 1.12	82.33 ± 0.47	83.46 ± 1.10	100.00 ± 0.00	83.57 ± 0.86	99.38 ± 0.62
	(0。1，0。01)	81.16 ± 1.12	82.33 ± 0.47	83.46 ± 1.10	100.00 ± 0.00	83.46 ± 1.25	100.00 ± 0.00
	(0.1, 0.005)	81.16 ± 1.12	82.33 ± 0.47	83.46 ± 1.10	100.00 ± 0.00	83.37 ± 1.34	100.00 ± 0.00
	(0.05，0.1)	82.71 ± 0.24	54.75 ± 3.12	83.29 ± 0.84	100.00 ± 0.00	81.99 ± 0.00	43.18 ± 0.00
	(0.05, 0.05)	82.71 ± 0.24	54.75 ± 3.12	83.29 ± 0.84	100.00 ± 0.00	83.93 ± 1.22	88.22 ± 5.17
	(0.05, 0.01)	82.71 ± 0.24	54.75 ± 3.12	83.29 ± 0.84	100.00 ± 0.00	83.54 ± 1.37	100.00 ± 0.00
	(0.05, 0.005)	82.71 ± 0.24	54.75 ± 3.12	83.29 ± 0.84	100.00 ± 0.00	83.34 ± 0.87	100.00 ± 0.00
	(0.025, 0.1)	–	–	84.03 ± 0.67	99.83 ± 0.15	81.99 ± 0.00	30.37 ± 0.00
	(0.025, 0.05)	–	–	84.03 ± 0.67	99.83 ± 0.15	83.57 ± 0.00	71.90 ± 0.00
	(0.025，0.01)	–	–	84.03 ± 0.67	99.83 ± 0.15	80.40 ± 0.62	99.71 ± 0.31
	(0.025, 0.005)	–	–	84.03 ± 0.67	99.83 ± 0.15	84.15 ± 0.92	99.75 ± 0.15
	(0.01, 0.1)	80.40 ± 0.75	0.00 ± 0.00	81.12 ± 1.01	0.00 ± 0.00	81.99 ± 0.00	0.00 ± 0.00
	(0.01, 0.05)	80.40 ± 0.75	0.00 ± 0.00	81.12 ± 1.01	0.00 ± 0.00	79.47 ± 0.94	0.00 ± 0.00
	(0.01，0.01)	80.40 ± 0.75	0.00 ± 0.00	81.12 ± 1.01	0.00 ± 0.00	81.38 ± 1.61	0.00 ± 0.00
	(0.01, 0.005)	80.40 ± 0.75	0.00 ± 0.00	81.12 ± 1.01	0.00 ± 0.00	80.03 ± 1.89	0.00 ± 0.00
Sylva	(2。0，0。01)	93.30 ± 0.00	100.00 ± 0.00	97.60 ± 0.28	100.00 ± 0.00	97.55 ± 0.18	100.00 ± 0.00
	(2.0, 0.005)	93.30 ± 0.00	100.00 ± 0.00	97.60 ± 0.28	100.00 ± 0.00	97.58 ± 0.23	100.00 ± 0.00
	(1.0, 0.01)	93.30 ± 0.00	100.00 ± 0.00	97.60 ± 0.28	100.00 ± 0.00	97.55 ± 0.18	100.00 ± 0.00
	(1.0, 0.005)	93.30 ± 0.00	100.00 ± 0.00	97.60 ± 0.28	100.00 ± 0.00	97.58 ± 0.23	100.00 ± 0.00
	(0.5, 0.01)	93.30 ± 0.00	100.00 ± 0.00	97.60 ± 0.28	100.00 ± 0.00	97.55 ± 0.18	100.00 ± 0.00
	(0.5, 0.005)	93.30 ± 0.00	100.00 ± 0.00	97.60 ± 0.28	100.00 ± 0.00	97.58 ± 0.23	100.00 ± 0.00
	(0.25, 0.01)	93.30 ± 0.00	100.00 ± 0.00	97.60 ± 0.28	100.00 ± 0.00	97.55 ± 0.18	100.00 ± 0.00
	(0.25, 0.005)	93.30 ± 0.00	100.00 ± 0.00	97.60 ± 0.28	100.00 ± 0.00	97.58 ± 0.23	100.00 ± 0.00
	(0.1, 0.01)	93.30 ± 0.00	100.00 ± 0.00	98.54 ± 0.06	100.00 ± 0.00	98.61 ± 0.06	100.00 ± 0.00
	(0.1, 0.005)	93.30 ± 0.00	100.00 ± 0.00	98.54 ± 0.06	100.00 ± 0.00	98.55 ± 0.04	100.00 ± 0.00
	(0.05, 0.01)	98.52 ± 0.08	23.89 ± 4.32	98.56 ± 0.07	11.11 ± 2.03	98.53 ± 0.09	10.00 ± 1.08
	(0.05, 0.005)	98.52 ± 0.08	23.89 ± 4.32	98.56 ± 0.07	11.11 ± 2.03	98.59 ± 0.11	10.56 ± 1.81
	(0。01，0。01)	98.46 ± 0.10	0.00 ± 0.00	98.37 ± 0.11	0.00 ± 0.00	98.49 ± 0.09	0.00 ± 0.00
	(0.01, 0.005)	98.46 ± 0.10	0.00 ± 0.00	98.37 ± 0.11	0.00 ± 0.00	98.53 ± 0.07	0.00 ± 0.00