计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(10): 2468−2480, 2025

BTFAT: 突破联邦对抗训练中鲁棒性与准确性权衡的框架

闫彩虹　芦效峰
（北京邮电大学网络空间安全学院　北京　100876）
（yancaihong@bupt.edu.cn）

BTFAT: A Framework Breaking the Trade-off Between Robustness and Accuracy in Federated Adversarial Training

Yan Caihong and Lu Xiaofeng
（*School of Cyberspace Security*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876）

Abstract　In the field of distributed artificial intelligence, federated learning models, like centrally trained models, are vulnerable to adversarial samples during reasoning. Currently, federated adversarial training has not been extensively explored and is facing two major challenges: 1) There is a trade-off between the model accuracy on clean samples and adversarial robustness, making it difficult to improve both simultaneously; 2) non-independent and identically distributed (Non-IID) data limit the performance improvement of federated adversarial training. To address these challenges, we propose the BTFAT framework, which breaks the trade-off between robustness and accuracy in federated adversarial training. The framework includes: 1) A decision-space tightening algorithm that performs initial intra-class localization using labels, while shrinking intra-class sample distances and increasing inter-class sample distances, thereby improving both robustness and accuracy; 2) A weight penalty optimization algorithm that uses the global model weights as the best unified target, penalizing local adversarial model training that deviates excessively, and assisting the decision-space algorithm to counter the impact of Non-IID data distributions. We theoretically analyze the key factors limiting the robustness and accuracy gains of adversarial training as well as the convergence of BTFAT. Simultaneously, we experimentally demonstrate that BTFAT comprehensively outperforms state-of-the-art baseline algorithms in terms of overall performance, convergence, time cost, and handling Non-IID data, providing a new perspective for research in federated adversarial training. Our code can be found in the website: https://anonymous.4open.science/r/BTFAT-11265.

Key words　distributed artificial intelligence; federated learning (FL); adversarial robustness; performance trade-off; non-independent and identically distributed

摘　要　在分布式人工智能领域，联邦学习模型与集中训练的模型一样，在推理时面对对抗样本是脆弱的. 目前，联邦对抗训练很大程度上仍未被探索，面临以下两大难题：1）干净样本的模型准确性和对抗鲁棒性之间存在权衡，难以共同提升两者效果；2）非独立同分布的数据限制了联邦对抗训练的性能提升. 针对以上难题，提出一种突破联邦对抗训练中鲁棒性与准确性权衡的框架 BTFAT，包括：1）决策空间收紧算法以标签进行初始类内定位，同时在决策空间中缩小类内样本的距离，增加类间样本的距离，实现鲁棒性和准确性的共同提升；2）权重惩罚优化算法以全局模型权重作为最佳统一目标，惩罚过度偏离的本地对抗模型训练，辅助决策空间算法抵御非独立同分布数据分布带来的影响. 在理论上分析出限制对抗训

收稿日期：2025-06-06；**修回日期**：2025-07-15
基金项目：国家自然科学基金项目（62136006, 62572065）
This work was supported by the National Natural Science Foundation of China (62136006, 62572065).
通信作者：芦效峰（Luxf@bupt.edu.cn）

练鲁棒性和准确性增益的关键以及 BTFAT 的收敛性，同时从实验上证明了 BTFAT 在整体性能、收敛性、时间开销以及面对非独立同分布数据时全面优于最先进基线算法的效果，为联邦对抗训练研究提供了一种新思路. 代码可见 https://anonymous.4open.science/r/BTFAT-11265.

关键词 分布式人工智能；联邦学习；对抗鲁棒性；性能权衡；非独立同分布

中图法分类号 TP311.13；TP309

DOI: 10.7544/issn1000-1239.202550442 **CSTR:** 32373.14.issn1000-1239.202550442

联邦学习（federated learning, FL）[1]是一种具有隐私保护的分布式人工智能范式，其中中央服务器收集参与者在其私有数据上客户端本地计算的模型更新，并聚合这些更新以训练全局模型. 目前，联邦学习吸引了来自不同领域的广泛关注，包括推荐系统[2-3]、医学诊断[4-5]、自然语言处理[6-7]、物联网[8]等领域.

研究表明，神经网络模型在推理过程中容易受到对抗样本的攻击[9]. 对抗样本[9-10]是模型的输入，目标是通过在测试用例上精心加入不可感知的扰动实现高置信度错误分类. 联邦学习训练得到的全局模型对于对抗样本具有脆弱性[11-12]：首先，联邦学习的分布式特性使得全局模型容易继承并放大各客户端间的数据分布偏差，从而降低模型对对抗样本的鲁棒性；其次，在非独立同分布（non-independent and identically distributed, Non-IID）数据下，全局模型可能在某些数据分布上过拟合，导致其在其他分布上的泛化能力不足，进而更容易被对抗样本误导；此外，联邦学习的聚合算法通常具有平滑效应，可能掩盖对抗样本引入的微小扰动，反而使全局模型对这类扰动更加敏感.

对抗样本攻击给联邦学习带来的严重影响不容忽视[13-14]. 攻击者可能通过恶意客户端上传精心构造的对抗样本，破坏全局模型的性能，例如导致医疗误诊、金融风控失效或自动驾驶系统误判. 此外，对抗样本攻击可能被用于增强隐私泄露，如结合成员推断攻击推测用户敏感数据，或植入隐蔽后门，使模型在特定场景下（如人脸识别、交通标志检测）产生预设错误. 因此，在联邦学习中防御对抗样本攻击也很重要.

针对对抗样本攻击的防御中对抗训练实现了最好的经验鲁棒性[9,15]. 目前，只有少量研究旨在处理联邦学习中的对抗样本攻击[16-18]，即联邦对抗训练（federated adversarial training, FAT）. 联邦对抗训练面临以下两大难题[19]：1）与普通联邦训练相比，联邦对抗训练下模型分类准确度较低，对抗训练损害了干净样本的性能，对抗鲁棒性和模型准确性之间存在权衡关系；2）在数据非独立同分布的联邦对抗训练场景中，模型性能显著下降. 针对以上难题，我们专注于突破联邦对抗训练鲁棒性和准确性之间的权衡，同时提升干净样本和对抗样本的分类性能，即使在数据非独立同分布下也能得到保证.

为实现该目标，我们从理论上分析出限制对抗训练鲁棒性和准确性增益的关键：样本到决策边界的距离. 同时，我们得出结论：样本到决策边界的距离越大，样本的鲁棒性和准确度越高. 在此基础上，我们创新性地提出决策空间收紧算法（针对难题 1）. 该算法以类别标签作为基准点进行类内样本定位，同时缩小类内样本在决策空间中的距离，扩大不同类样本在决策空间中的距离. 其次，为缓解非独立同分布数据对决策空间收紧算法的影响，提出权重惩罚优化算法（针对难题 2），以全局模型权重作为最佳统一目标，惩罚过度偏离的本地对抗模型训练.

具体地说，我们的主要贡献有 5 点：

1）提出一种即使在非独立同分布复杂数据下依旧可以突破模型准确性和对抗鲁棒性权衡的联邦对抗训练框架 BTFAT（Breaking the trade-off between robustness and accuracy in federated adversarial training）.

2）针对干净样本的模型准确性和对抗鲁棒性之间的权衡难题，提出决策空间收紧算法，突破了对抗鲁棒性和准确性间的权衡难题.

3）针对非独立同分布数据限制联邦对抗训练性能提升的难题，提出权重惩罚优化算法，辅助决策空间收紧算法即使面对非独立同分布数据也能实现全面性能提升.

4）从理论上证明了样本到决策边界的距离是限制对抗鲁棒性和准确性提升的关键、嵌入表示空间距离和决策空间距离之间的近似关系以及 BTFAT 框架在非凸下的收敛性.

5）实验结果表明，我们提出的方法 BTFAT 比最先进基线具备更佳的对抗鲁棒性和准确性，有效打破两者间的权衡，同时在非独立同分布下依旧性能稳定.

1 相关工作

对抗样本攻击可以分为白盒攻击和黑盒攻击2类. 白盒攻击可以访问威胁模型的全部细节，包括模型架构和参数. Goodfellow 等人[9]引入了快速梯度符号法(fast gradient sign method, FGSM)生成对抗性示例，通过单步一阶近似来执行梯度上升. Kannan 等人[20]迭代应用具有小步长的 FGSM，提出了迭代的 FGSM(iterative fast gradient sign method, I-FGSM). 基于以上研究，近年来提出了更强大的对抗攻击方法，包括动量迭代法(momentum iterative method, MIM)[21]、投影梯度下降攻击(projected gradient descent, PGD)DeepFool 攻击[22]. DeepFool 攻击通过迭代逼近分类器的决策边界来生成最小扰动的对抗样本. C&W 攻击(Carlini & Wagner attack, CW)[23]和自动攻击(autoattack, AA)[24]. MIM[21]在迭代过程中引入动量，每次迭代时，不仅考虑当前的梯度方向，还考虑之前梯度的累积效果，从而使得对抗样本的生成更加稳定和有效. PGD[22]通过多次迭代来逐步优化扰动，每次迭代都会根据模型的梯度信息调整扰动方向，并将扰动投影到一定的约束空间内，以确保生成的对抗样本在合法范围内. CW[23]通过优化算法生成对抗样本，这些样本通过添加几乎不可察觉的扰动来欺骗模型. AA[24]则是一种基于特征空间扰动的黑盒目标性对抗攻击方法. 该方法通过驱动源图像特征向目标图像特征靠近来生成对抗样本，从而增加对抗样本的迁移性. 然而，实验表明传统联邦学习场景在这些对抗样本攻击下具有脆弱性[25].

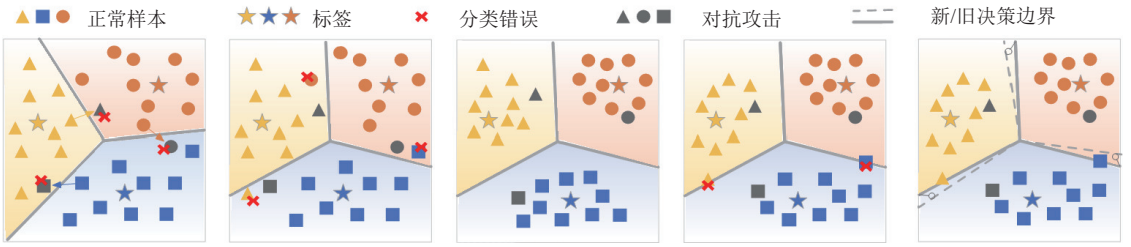
1)集中式对抗训练. 针对对抗样本攻击，集中式机器学习场景下提出了各种对抗训练方法. Madry 等人[22]将对抗训练视为 PGD 攻击下使用经验风险最小化的 min-max 公式. 对抗 Logit 配对(adversarial Logit pairing, ALP)算法[20]通过鼓励样本对的逻辑值相似以增强鲁棒的准确度. 为了量化准确度和鲁棒性之间的权衡，Zhang 等人[26]引入 TRADES 损失，以实现干净样本误差和鲁棒样本误差之间差距的严格上限. 基于边缘理论和软标记数据增强，Ding 等人[27]提出了最大间隔对抗训练(max-margin adversarial training, MMA)算法，Lee 等人[28]提出了对抗顶点混合(adversarial vertex mixup, AVmixup)算法. 然而，集中式对抗训练算法无法直接应用于联邦训练的分布式场景.

2)联邦对抗训练. 针对于联邦学习中的对抗样本攻击，鲁棒性提升方法包括对抗训练[29-30]、数据消毒[31]、知识蒸馏[32]、梯度正则化[33]和基于图注意力网络的防御[34]. 其中，对抗训练[23]是对抗样本攻击防御方法中最流行且最有效的方法之一. 然而，联邦对抗训练的研究较少. Zizzo 算法[16]率先研究了集中式下的对抗性训练协议在理想化联邦学习中的有效性及在集中式和分布式下的性能. Zhou 等人[17]则是将中央服务器的聚合误差分解为偏差和方差. 虽然文献[16-17]方法对于抵御对抗性攻击具有一定效果，但是都牺牲了干净样本下的准确度以获得鲁棒性，尤其是在非独立同分布数据下准确度严重下降. 为同时实现准确度和鲁棒性，基于决策边界的联邦对抗训练(decision boundary based federated adversarial training, DBFAT)算法[18]通过局部重新加权和全局正则化实现了基于决策边界的联邦对抗训练. 虽然通过加固决策边界有效地缓解了干净样本的误分类，但是这仅适用于被原有决策空间成功划分的样本，对于被原有决策空间错误分类的样本将更难以恢复. 同时，在非独立同分布数据场景下，全局正则化系数难以确定并且受初始全局模型的影响将会较大，可能得到全局次优模型. 因此，如何将对抗样本和干净样本收紧到正确的决策空间，同时保证在非独立同分布的数据场景中算法稳健性成为目前联邦学习领域未解决的共性挑战. 针对该难题，我们不仅从理论层面分析出突破联邦对抗训练鲁棒性和准确性矛盾的关键，同时提出了包括决策空间收紧算法和权重惩罚优化算法的 BTFAT 框架，并在实验层面取得了最佳效果，为今后的研究提供了新思路.

2 BTFAT 的提出

为缓解目前联邦对抗训练面临干净样本准确性和对抗鲁棒性难以权衡，同时非独立同分布数据影响联邦对抗训练性能提升两大关键难题，我们提出包含决策空间收紧算法和权重惩罚优化算法的联邦对抗训练框架 BTFAT. 决策空间收紧算法作为 BTFAT 的核心，打破了原有对抗训练中鲁棒性和准确性需要权衡的困境，同时保证了强鲁棒性和高性能. 权重惩罚优化算法则用于辅助缓解非独立同分布场景对于决策空间收紧算法的影响. 整体效果如图 1 所示.

1)决策空间收紧算法. 为实现各类样本决策空间的最佳收紧，我们首先在类内进行样本定位聚集，规划出类内样本的大致空间. 然后实现类内样本的收紧，内部缩小同类样本在决策空间中的距离，同时



(a) 简单联邦学习的决策边界 (b) 普通联邦对抗训练的决策边界 (c) 决策空间收紧算法在独立同分布数据下的决策边界 (d) 决策空间收紧算法在非独立同分布数据下的决策边界 (e) BTFAT 面对非独立同分布数据时的决策边界

Fig. 1 Five distinct decision boundaries

图 1 5 个不同的决策边界

放大与不同类别样本间的距离. 进一步为避免梯度消失, 将嵌入表示空间距离近似决策空间距离, 最终提出对抗嵌入收紧损失来打破对抗训练的权衡困境.

2)权重惩罚优化算法. 为辅助决策空间收紧算法在数据非独立同分布下获得全局最优对抗模型, 需要限制本地训练优化的路径. 我们将每轮的全局模型权重作为统一的最佳优化目标, 惩罚本地模型权重过度偏离全局模型的训练, 缓解数据异质导致的对抗效果不佳.

2.1 概念预定义

1)决策空间

在机器学习分类模型中, 决策空间是特征空间的划分, 表示模型如何基于输入特征决定输出类别. 具体来说, 模型将特征空间划分为多个区域, 每个区域对应一个类别标签. 例如, 以二分类问题为例, 线性模型(如逻辑回归)的决策空间通过式(1)定义.

$$f(x)=\boldsymbol{w}^{\mathrm{T}}x+b, \tag{1}$$

其中, $\boldsymbol{w}$是权重向量, b是偏置项. 模型通过决策函数$f(x)$对输入进行分类, 若$f(x)>0$则预测输入为类别 1, 若$f(x)<0$则预测输入为类别 0. 因此, 决策空间实际上是特征空间中由模型的分类规则$f(x)$所产生的不同类别区域.

2)决策空间收紧

在机器学习分类模型中, 通过调整模型的参数、结构或训练策略等, 使得决策空间中每个类别的划分变得更加明确、精准, 减少类别间的重叠区域, 并使得类别的决策边界更加紧凑和清晰.

2.2 BTFAT 底层支撑理论的分析

Tsipras 等人[19]认为在对抗性训练中学习到的特征不同于在标准训练中学习到的特征, 表明对抗训练鲁棒性和标准模型准确性之间可能存在权衡. 已有研究通过对抗样本协同、聚合泛化误差分解、对抗样本加权的策略来缓解联邦对抗训练中的权衡难题, 然而依旧效果不佳. 探索同时提升干净样本准确性和对抗鲁棒性的关键成为联邦对抗训练中的难点. 受“决策空间加固”理念[18]的启发, 我们通过定理 1 和定理 2 证明了增大各类别样本到决策边界的距离对于提高对抗训练鲁棒性和准确性具备共同增益效果, 这也成为决策空间收紧算法打破权衡的关键. 全文所用符号汇总如表 1 所示.

定理 1. 假设在K分类任务中, 函数$f(x;\theta)=[f_1(x;\theta),\cdots,f_K(x;\theta)]$在样本$x$处满足可微性, 那么$x$到决策边界的几何间距$\gamma(x)=\dfrac{f_y(x;\theta)-\max_{k\neq y}f_k(x;\theta)}{\left\|\nabla_x\left(f_y(x;\theta)-f_{k^*}(x;\theta)\right)\right\|}$和最小对抗扰动$\|\delta\|$大小相等. 因此, 几何间距$\gamma(x)$越远, 模型对抗鲁棒性越强.

证明. 在K分类任务中, 假设模型的输出为向量$\boldsymbol{f}(x;\theta)=(f_1(x;\theta),f_2(x;\theta),\cdots,f_K(x;\theta))$, 其中$f_k(x;\theta)$表示样本$x$被分为第$k$类的得分. 对于给定样本$x$及其真实标签$y$, 我们定义样本到决策边界的几何间距, 即样本$x$在输出空间中距离分类边界的“最近间隔”:

$$\gamma(x)=\frac{f_y(x;\theta)-\max_{k\neq y}f_k(x;\theta)}{\left\|\nabla_x\left(f_y(x;\theta)-f_{k^*}(x;\theta)\right)\right\|}, \tag{2}$$

其中$k^*=\arg\max_{k\neq y}f_k(x;\theta)$表示与真实类别$y$最接近的其他类别. 分子项表示正确类别与最接近错误类的“置信度差值”; 分母项为这个差值对输入变化的敏感程度; 整体表示在最敏感方向上, 要跨越多远才能使分类发生改变.

接下来构造对抗扰动δ, 其目标是最小化$\|\delta\|$以满足:

$$\arg\max_k\ f_k(x+\delta;\theta)\neq y. \tag{3}$$

即找到最小的扰动δ使得样本被分类器错误分类:

$$f_y(x+\delta;\theta)\leqslant f_{k^*}(x+\delta;\theta). \tag{4}$$

$f(x;\theta)$在样本x处满足可微性, 那么可以在x处对对抗扰动δ进行一阶泰勒展开, 得到:

Table 1 Summary List of Symbols

表 1 符号汇总表

符号	说明
K	分类数量
f	决策函数
γ	几何间距
x	训练样本
y	真实标签
δ	对抗扰动
R	泛化误差
$\hat{R}$	经验误差
N	所有样本数量
$\mathcal{F}$	假设空间
$\mathcal{R}_S$	Rademacher 复杂度
L	Lipschitz 常数
d_{decision}	决策空间距离
d_{embed}	嵌入空间距离
C	敏感度常数
e	嵌入映射函数
$\boldsymbol{J}$	雅可比矩阵
l_{agg}	类内样本定位聚集损失
m	类别区分掩码
l_{tig}	收紧阶段总损失
l_{adv}	对抗嵌入收紧损失
E	全部训练轮次
w	模型参数权重
α	权重惩罚系数
l_{w}	惩罚项损失
l_{total}	整体训练损失
F	本地目标函数

$$f_y(x+\delta;\theta)-f_{k^*}(x+\delta;\theta)\approx\left(f_y(x;\theta)-f_{k^*}(x;\theta)\right)+\nabla_x\left(f_y(x;\theta)-f_{k^*}(x;\theta)\right)^{\mathrm{T}}\delta. \tag{5}$$

为了使分类结果刚好发生翻转, 即令式(5)为 0.

$$f_y(x;\theta)-f_{k^*}(x;\theta)+\nabla_x\left(f_y(x;\theta)-f_{k^*}(x;\theta)\right)^{\mathrm{T}}\delta=0. \tag{6}$$

选取最陡下降方向作为扰动方向, 即令$\delta=-\alpha\times\nabla_x(f_y(x;\theta)-f_{k^*}(x;\theta))$, 其中$\alpha>0$. 将其带入边界条件解得:

$$\alpha=\frac{f_y(x;\theta)-f_{k^*}(x;\theta)}{\left\|\nabla_x\left(f_y(x;\theta)-f_{k^*}(x;\theta)\right)\right\|^2}. \tag{7}$$

因此, 最小扰动为:

$$\|\delta\|=\alpha\left\|\nabla_x\left(f_y(x;\theta)-f_{k^*}(x;\theta)\right)\right\|=\frac{f_y(x;\theta)-f_{k^*}(x;\theta)}{\left\|\nabla_x\left(f_y(x;\theta)-f_{k^*}(x;\theta)\right)\right\|}=\gamma(x). \tag{8}$$

由此可见, 几何间距$\gamma(x)$表示最小扰动的大小, 几何间距越大, 对抗鲁棒性越强. 证毕.

定理 2. 假设在K分类任务中, 函数$\boldsymbol{f}(x;\theta)=(f_1(x;\theta), f_2(x;\theta),\cdots,f_K(x;\theta))$中每个分类器$f_k(x;\theta)$是 L-Lipschitz 连续的, 那么泛化误差$R(f)$和经验误差$\hat{R}(f)$间满足$R(f)\leqslant\hat{R}(f)+O\left(\frac{L}{\gamma}\sqrt{\frac{\log N}{N}}\right)$. 因此, 几何间距$\gamma$越大, K分类泛化性能越好.

证明. 在K分类任务中, 设训练集为$\{(x_i,y_i)\}_{i=1}^N$, 基于统计学习理论, 泛化误差$R(f)$和经验误差$\hat{R}(f)$满足以下关系:

$$R(f)\leqslant\hat{R}(f)+2\mathcal{R}_{\mathrm{S}}(\mathcal{F})+O\left(\frac{1}{\sqrt{N}}\right), \tag{9}$$

其中$\mathcal{R}_S(\mathcal{F})$是假设空间$\mathcal{F}$的 Rademacher 复杂度, 用于衡量模型类对样本噪声的拟合能力, 是控制过拟合风险的关键因素. 泛化误差$R(f)$表示模型在真实数据分布下的平均分类错误率. 经验误差$\hat{R}(f)$表示模型在训练集上的平均分类误差.

假设每个分类器$f_k(x;\theta)$是 L-Lipschitz 连续的, 这意味着模型不会对输入的微小扰动产生过大的输出变化, 即满足:

$$|f_k(x;\theta)-f_k(x';\theta)|\leqslant L\|x-x'\|. \tag{10}$$

根据已有理论结果[35], 对于具有几何间距至少为γ的假设空间$\mathcal{F}$, Rademacher 复杂度满足:

$$\mathcal{R}_S(\mathcal{F})\leqslant\frac{L}{\gamma}\sqrt{\frac{\log N}{N}}. \tag{11}$$

这是由于几何间距γ越大, 模型对数据的区分越清晰, 从而减少对随机噪声的拟合能力, 等效于“缩小了假设空间的复杂度”.

将上述 Rademacher 复杂度界带入泛化误差不等式中, 即可简化得到:

$$R(f)\leqslant\hat{R}(f)+O\left(\frac{L}{\gamma}\sqrt{\frac{\log N}{N}}\right). \tag{12}$$

因此, 增大几何间距γ可以降低泛化误差上界, 进而提升泛化性能. 证毕.

2.3 决策空间收紧算法

在定理 1 和定理 2 的基础上, 决策空间收紧算法主要实现落实逐步增大各类样本到决策边界的距离的目标, 这往往需要可靠度量样本到决策边界的距离. 有些研究通过计算迭代攻击(例如 PGD 攻击)使得原始输入错误分类所需的最小步数来衡量原始输入到决策空间的距离. 然而, 这种衡量步数的方法无法用于对抗训练优化来实现决策空间的逐步收紧过程. 即使通过迭代攻击得到距离决策边界最远的样本来收紧也依旧效果不佳[18]. 这是由于分类任务的交

叉熵损失中 one-hot 编码的标签已经作为“最远的类内样本”进行了样本聚集，2 个“最远点”将产生冲突. 然而，交叉熵损失只能实现类内样本向决策空间中的标签聚集，而无法保证决策空间收紧，同时优化决策空间下的距离将出现梯度消失问题. 因此，为实现决策空间收紧的目标，我们首先对决策空间距离度量进行了替代，通过定理 3 证明了嵌入空间的距离可以近似衡量决策空间的距离.

定理 3. 假设分类任务中的嵌入映射函数 e 和决策函数 f 是平滑且可微的，对于任意 2 个样本 x_1 和 x_2，它们在嵌入空间的距离与在决策空间的距离之间存在线性关系，即 $d_{\text{decision}} \approx C \cdot d_{\text{embed}}(x_1, x_2)$.

证明. 模型通过嵌入函数 $e(x)$ 将原始输入样本 x 映射到高维表示空间，并通过决策函数 f 将嵌入表示映射到决策空间. 给定样本 x_1 和 x_2，在嵌入空间中的距离为：

$$d_{\text{embed}}(x_1, x_2) = \|e(x_1) - e(x_2)\|. \tag{13}$$

在决策空间中的距离为：

$$d_{\text{decision}}(x_1, x_2) = \|f(e(x_1)) - f(e(x_2))\|. \tag{14}$$

假设决策函数 f 是平滑且可微的，样本 x_1 和 x_2 足够接近时，可以通过一阶泰勒展开对决策函数进行局部线性化：

$$f(e(x_2)) \approx f(e(x_1)) + \boldsymbol{J}_f(e(x_1)) \cdot (e(x_2) - e(x_1)), \tag{15}$$

其中，$\boldsymbol{J}_f(e(x_1))$ 是 f 在 $e(x_1)$ 处的雅可比矩阵，表示输出对输入每维分量的偏导.

因此，得到决策空间的距离近似为：

$$d_{\text{decision}}(x_1, x_2) \approx \left\| \boldsymbol{J}_f(e(x_1)) \cdot (e(x_2) - e(x_1)) \right\|. \tag{16}$$

设 $C = J_f(e(x_1))$ 表示雅可比矩阵的谱范数，$J_f(e(x_1))$ 在 x_1 附近变化不大，则有：

$$d_{\text{decision}} \approx C \cdot d_{\text{embed}}(x_1, x_2), \tag{17}$$

其中，C 可视为决策函数 f 对嵌入距离变化的敏感度，在局部范围内近似为常数. 因此，嵌入空间的距离与决策空间的距离之间存在线性关系，且在模型平滑且局部线性时，二者变化趋势相似. 证毕.

决策空间收紧算法包括类内样本定位聚集和样本收紧 2 个过程. 前者主要是保证各类样本向决策空间中的标签聚集，以保证基础模型的准确性；后者则在确定聚集位置后全力实现空间收紧，进一步保证对抗鲁棒性.

1）类内样本定位聚集. 类内样本定位聚集过程主要是通过干净样本和对抗样本在对抗训练中的交叉熵损失函数 l_{ce} 优化来实现类内样本在决策空间的定位，向类内标签在决策空间的位置聚集可表示为：

$$l_{\text{agg}} = l_{\text{ce}}\left(f^{\text{loc}}\left(x^{\text{adv}}\right), y\right) + l_{\text{ce}}\left(f^{\text{loc}}(x), y\right), \tag{18}$$

其中，f^{loc} 是联邦对抗训练中参与方的本地模型，l_{ce} 是交叉熵损失，x^{adv} 是通过原始样本 x 生成的对抗样本，y 是标签.

2）样本收紧. 样本收紧过程包括类内样本的拉近和类间样本的疏远. 首先，为计算决策空间中样本距离的远近，我们通过样本嵌入表示的欧式距离 d_{ij} 来对样本 $x_i, x_j \in x^{\text{adv}} \cup x$ 进行距离度量：

$$d_{ij} = \|\boldsymbol{e}_i - \boldsymbol{e}_j\|_2 = \sqrt{\sum_{k=1}^{d} (e_{i,k} - e_{j,k})^2}, \tag{19}$$

其中，$\boldsymbol{e}_i$ 和 $\boldsymbol{e}_j$ 分别表示样本 x_i 和 x_j 的嵌入表示向量.

为区分同类样本对和非同类样本对，我们进一步创建了类别区分掩码 m_{ij}，如果类别相同，则 $m_{ij} = 1$，否则 $m_{ij} = 0$.

$$m_{ij} = \begin{cases} 1, & y_i = y_j, \\ 0, & y_i \neq y_j. \end{cases} \tag{20}$$

在以上距离度量和类别区分掩码基础上，我们将通过拉近类内样本距离和增大类间样本距离来实现空间收紧. 具体地说，对于标签相同的样本对，使它们之间的欧式距离尽可能地小：

$$\min \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} m_{ij} d_{ij}, \tag{21}$$

其中，N 表示所有样本数.

对于标签不同的样本对，则需要它们之间的距离尽可能地大. 我们加入了预设的 $margin$ 参数. $margin$ 参数主要用于控制不同类别样本之间的距离，确保训练过程中彼此有足够的间隔，从而提升模型的区分能力. 通常令 $margin = 1$ 以实现区分能力和训练稳定的权衡[36]. 如果不同类样本对之间的距离小于 $margin$，就需要对其进行惩罚，使得它们的距离在训练优化中增加. 类间疏远的优化目标为：

$$\min \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (1 - m_{ij}) \left(\max\left(0, margin - d_{ij}\right)\right)^2. \tag{22}$$

收紧阶段的总损失是同类样本对距离损失和非同类样本对距离损失的平均：

$$l_{\text{tig}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \Big[m_{ij} d_{ij} + (1 - m_{ij}) \left(\max\left(0, margin - d_{ij}\right)\right)^2 \Big]. \tag{23}$$

3）对抗嵌入收紧损失设计. 考虑到在对抗训练的不同阶段中类内样本定位聚集和样本收紧 2 个过程具有不同的重要性，需要对决策空间收紧算法的

2个阶段损失进行加权. 因此, 最终的对抗嵌入收紧损失l_{adv}是:

$$l_{\text{adv}}=\left(1-\frac{t}{E}\right)l_{\text{agg}}+\frac{t}{E}l_{\text{tig}}, \tag{24}$$

其中, t表示当前轮次, E表示全部的训练轮次.

联邦对抗训练过程中, 决策空间收紧算法在类内样本定位聚集和样本收紧2阶段中逐步完成目标决策空间定位和决策空间收缩调整, 保证干净样本和对抗样本大多数都约束在正确的类别决策空间中, 为打破联邦对抗训练中模型准确性和鲁棒性之间的权衡提出新思路.

2.4 权重惩罚优化算法

在联邦学习等分布式训练任务中, 不同客户端的本地模型可能会因数据分布的差异而产生较大的偏离, 导致对抗训练效果不佳. 为获得全局最优的对抗模型, 需要限制各参与方本地训练的优化. 我们将聚合后的全局对抗模型用来约束各参与方本地下一轮次的模型权重, 提出权重惩罚优化算法, 以辅助决策空间收紧算法应对数据非独立同分布场景对联邦对抗训练性能的影响. 权重惩罚优化算法主要包括权重差异度量和加权归一化2个步骤.

1)权重差异度量. 权重差异度量用于衡量本地对抗模型和全局对抗模型在参数空间中的偏离程度, 通过计算每个参数的l_2范数来实现. 对于本地模型参数p^{loc}和全局模型参数p^{glo}, 总的权重差异定义为:

$$\left\|p^{\text{loc}}-p^{\text{glo}}\right\|_2=\sum_{k=1}^{N}\left\|p_k^{\text{loc}}-p_k^{\text{glo}}\right\|_2=\sum_{k=1}^{N}\sqrt{\sum_{i=1}^{d}\left(p_{k,i}^{\text{loc}}-p_{k,i}^{\text{glo}}\right)^2}, \tag{25}$$

其中, k是模型的第k个参数, d表示参数的维度, N是模型参数的总数.

2)加权归一化. 加权归一化旨在消除模型规模对权重差异的影响, 并进一步通过权重惩罚系数α调节惩罚优化项的权重比例, 调节损失函数中惩罚项的强度, 以控制模型的复杂度和过拟合, 确保模型的拟合与正则化之间达到一个合理的平衡. 具体公式为:

$$l_{\text{w}}=\frac{\alpha}{N}\sum_{k=1}^{N}\left\|p_k^{\text{loc}}-p_k^{\text{glo}}\right\|_2. \tag{26}$$

将惩罚项融入对抗嵌入收紧损失l_{adv}, 得到本地对抗训练的最终损失函数:

$$l_{\text{total}}=l_{\text{adv}}+l_{\text{w}}. \tag{27}$$

权重惩罚优化算法辅助原始决策空间收紧算法, 引导模型训练过程更加平滑, 避免局部过拟合, 将有效缓解非独立同分布数据对于联邦对抗训练性能的影响.

虽然整体的损失函数形式上看似结合了交叉熵损失、三元组损失和正则化损失, 并利用了与对比学习类似的核心思想, 但是本质上并非简单拼接, 而是为解决具体任务问题而设计的具有针对性与创新性的组合. BTFAT创新性地通过类内样本定位聚集、类内样本拉近、类间样本疏远、非IID权重惩罚优化这一系列新逻辑策略, 并融合自适应权重调节机制, 突破了联邦学习对抗训练中鲁棒性和准确性权衡的特定关键难题, 构建出了联邦学习对抗训练新框架.

3 收敛性分析

在联邦对抗训练框架BTFAT中, 设全局优化目标为:

$$f\left(p^{\text{glo}}\right)=\frac{1}{K}\sum_{k=1}^{K}F_k\left(p^{\text{glo}}\right), \tag{28}$$

其中, $F_k\left(p^{\text{glo}}\right)$表示第$k$个设备上的局部目标函数, p^{glo}为全局模型参数, K是设备总数.

假设局部目标函数$F_k\left(p^{\text{glo}}\right)$是非凸的, 并满足L-Lipschiits平滑性, 梯度变化受控, 目标函数不会剧烈震荡, 即:

$$\left\|\nabla F_k\left(p_1^{\text{glo}}\right)-\nabla F_k\left(p_2^{\text{glo}}\right)\right\|\leqslant L\left\|p_1^{\text{glo}}-p_2^{\text{glo}}\right\|. \tag{29}$$

每个局部目标函数满足梯度下界假设, 即存在常数$L^-\gt 0$, 使得$\nabla^2F_k\left(p^{\text{glo}}\right)\geqslant L^-I$, 这保证目标函数在局部具有一定的“上凹性”, 使得优化过程稳定. 据此定义$\bar{\mu}:=\mu-L^-\gt 0$, 其中μ是对局部目标函数的强凸性控制参数.

在这样的条件下, BTFAT的收敛性分析如下:

在第t轮, 设选中设备集合为S_t, 设全局更新后模型参数为p_{t+1}^{glo}. BTFAT的局部更新及聚合策略可保证在合理学习率γ下, 存在梯度下降步长ρ使得:

$$\mathbb{E}_{S_t}\left[f\left(p_{t+1}^{\text{glo}}\right)\right]\leqslant f\left(p_t^{\text{glo}}\right)-\rho\left\|\nabla f\left(p_t^{\text{glo}}\right)\right\|^2. \tag{30}$$

式(30)是收敛性分析的核心, 表明每一轮全局更新能在期望意义下减少目标函数值, 且减少量与梯度模长的平方成正比. 其中, ρ是梯度下降步长, 具体由学习率γ、设备数量K、设备间数据不相似度参数B(用于量化统计异构性)共同决定.

为保证目标函数有效下降, 需满足以下约束: $\gamma B\lt 1$且$tB/\sqrt{K}\lt 1$. 这表示在数据异构性B给定的前提下, 需调控学习率和设备数量以抑制偏差积累.

将式(30)在T轮内累加并整理, 可得BTFAT的收敛速率为:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(p_t^{\mathrm{glo}})\|^2]\leqslant\frac{\Delta}{\rho T}\Rightarrow O\left(\frac{\Delta}{\rho\epsilon}\right). \tag{31}$$

BTFAT 经过多轮迭代后收敛至近似驻点集合：

$$S_s=\{w|\mathbb{E}\left[\|\nabla f\left(p^{\mathrm{glo}}\right)\|^2\right]\leqslant\epsilon\}, \tag{32}$$

其中，$\Delta=f(p_0^{\mathrm{glo}})-f^*$表示初始目标值与最优值之间的差，$\epsilon$表示收敛精度的阈值.

在凸目标函数的特殊情形下，即局部目标函数$F_k\left(p^{\mathrm{glo}}\right)$为凸函数，且局部优化问题被精确求解（即无近似解误差），BTFAT 的收敛速率可进一步提升，与分布式 SGD 保持一致级别：

$$O\left(\frac{L\Delta}{\epsilon}+\frac{L\Delta\sigma^2}{\epsilon^2}\right), \tag{33}$$

其中σ^2表示梯度估计的方差. 同时，当权重惩罚优化系数$\alpha>0$时，BTFAT 能够在更高数据异构性场景下保持稳定的收敛性能.

通过将设备不相似性、局部更新约束和参与率结合，BTFAT 的理论分析提供了保证非凸情形下收敛的充分条件，并验证了在异构数据分布下的稳定性.

4 实验评估

4.1 实验设置

1）数据集

MNIST[37] 是手写数字识别数据集，包括 60 000 张训练灰度图和 10 000 张测试灰度图，图像大小为 28×28，标签有 10 类；Fashion-MNIST[38] 是衣物分类数据集，具有和 MNIST 相同的图像尺寸、样本数、标签数；CIFAR-10[39] 是物体识别数据集，包括 50 000 张训练图像和 10 000 张测试图像，每个样本是 32×32 的 RGB 彩色图像，标签有 10 类.

2）攻击方法

FGSM[9] 通过对输入数据添加扰动，增加模型的损失值，从而使其分类错误；PGD[11] 是 FGSM 的扩展，采用多次迭代的方式生成对抗样本，并在每次迭代后将样本投影到可行域内以满足约束条件，其中 PGD-10 和 PGD-20 指分别进行 10 次和 20 次迭代的 PGD 攻击；CW[23] 通过优化目标函数生成对抗样本，其目标是最小化扰动的同时实现高效攻击.

以上攻击[9,11,23] 的实现方式包括白盒和黑盒 2 种. 白盒攻击下将基于联邦训练得到的全局模型展开攻击. 针对 MNIST 和 Fashion-MNIST 数据集，采用 3 层卷积层和 2 层全连接层训练. 针对 CIFAR-10 数据集，则采用 6 层卷积层、1 层池化层、3 层全连接层和 1 层 Dropout 层进行训练.

黑盒攻击下，针对 MNIST 和 Fashion-MNIST 数据集，采用 4 种经过预训练的模型：2 层卷积层、2 层全连接层和 1 层 Dropout 层；3 层卷积层、1 层全连接层和多处 Dropout 层；2 层卷积层、1 层池化层和 2 层全连接层；3 层全连接层和 1 层 Dropout 层. 针对 CIFAR-10 数据集，同样采用 4 种经过预训练的模型：基于残差网络的 ResNet18[40]；具有批归一化的 VGG11[41]；基于深度可分离卷积的 Xception[42]；轻量化设计的 MobileNetV2[43].

3）对比基线

FAT[16] 是将联邦学习客户端本地对干净样本和对抗样本一起执行训练；FedBVA[17] 将中央服务器的聚合的泛化误差分解为偏差和方差，通过在服务器更新期间利用共享数据集最大化偏差和方差来生成对抗性样本；DBFAT[18] 通过迭代 PGD 攻击需要的最少步骤来近似衡量对抗样本从原始输入到分类器决策边界的距离，接近决策边界的干净样本分配大权重，远离决策边界的干净样本分配小权重.

4）评估指标

①性能指标：在干净测试样本上分类正确的样本比例；

②鲁棒性指标：在以上白盒对抗样本攻击和黑盒对抗样本攻击下，正确分类的测试样本比例.

5）其他设置

客户端共 100 个，每轮随机选择 10% 的客户端进行联邦对抗训练，威胁和防御模型的最大扰动δ为 0.3，根据全部的测试样本生成了等量的对抗样本. 整体协同 100 轮次联邦训练，学习率为 0.01，批次大小为 64，BTFAT 的α默认值为 10. 数据集默认被均匀划分，非独立同分布场景下通过控制狄利克雷分布的β实现划分.

4.2 算法整体性能评估实验

在表 2 中展示了各种基线和本文方法在 3 种不同数据集、8 类攻击方法下的准确性和鲁棒性效果. 每种黑盒攻击下的鲁棒性性能都是在 4 种预训练模型下进行的平均. 由于 CW 攻击和 PGD-10/PGD-20 属于典型的强攻击范式，原始基线模型在未加防御的情况下表现出较差的鲁棒性，从而限制了可提升的防御空间. 因此，在部分攻击场景中，我们提出的 BTFAT 及其他联邦对抗训练方法的提升幅度相对有限. 另外，FAT 只是简单地将干净样本和生成的对抗样本进行联邦训练，干净样本上的准确性和对抗各种对抗样本攻击的能力较为薄弱. FedBVA 则改进了

Table 2 Overall Evaluation of Various Performance Under Different Datasets and Adversarial Attack Strategies

表 2 在不同数据集和对抗攻击策略下各种性能的全面评估

数据集	算法	准确性	白盒攻击下的鲁棒性				黑盒攻击下的鲁棒性			
		干净样本	FGSM	PGD-10	PGD-20	CW	FGSM	PGD-10	PGD-20	CW
MNIST	FAT	0.976 1	0.851 2	0.900 6	0.780 9	0.982 7	0.932 5	0.967 8	0.959 8	0.959 3
	FedBVA	0.976 5	0.860 0	0.903 9	0.783 5	0.982 7	0.935 6	0.968 9	0.961 5	0.960 8
	DBFAT	0.943	0.825 8	0.828 9	0.674 7	0.949 8	0.892 7	0.941 3	0.926 2	0.927 1
	BTFAT（本文）	**0.977 2**	**0.893 7**	**0.919 0**	**0.841 0**	**0.983 1**	**0.936 9**	**0.968 9**	**0.962 4**	**0.962 5**
Fashion-MNIST	FAT	0.808 2	0.616 9	0.372 4	0.174 1	0.826 0	0.766 5	0.783 8	0.779 2	0.780 3
	FedBVA	0.809 5	0.615 4	0.386 9	0.188 0	0.826 4	0.768 5	0.782 6	0.779 2	0.778 2
	DBFAT	0.613 3	0.608 1	0.449 9	0.247 0	0.625 2	0.681 1	0.655 5	0.660 8	0.660 9
	BTFAT（本文）	**0.811 4**	**0.666 4**	**0.450 3**	**0.254 6**	**0.840 9**	**0.781 7**	**0.794 8**	**0.793 3**	**0.792 9**
CIFAR-10	FAT	0.712 1	0.806 2	0.004 1	0.000 8	0.259 6	0.741 5	0.772 6	0.771 4	0.771 4
	FedBVA	0.729 3	0.819 0	0.004 5	0.000 3	0.268 5	0.741 0	0.772 5	0.772 7	0.772 2
	DBFAT	0.690 0	0.787 7	0.003 8	0.000 2	0.214 1	0.738 1	0.757 4	0.755 9	0.756 7
	BTFAT（本文）	**0.780 6**	**0.834 5**	**0.004 7**	**0.000 8**	**0.405 5**	**0.743 4**	**0.777 4**	**0.776 1**	**0.777 9**

注：黑体数值表示最优结果.

对抗样本生成的策略，利用共享数据集最大化偏差和方差来生成对抗性样本，提升准确性的同时增强了对抗鲁棒性. 平均来看，FedBVA 较 FAT 而言，准确性提升 0.63 个百分点，白盒攻击和黑盒攻击上的鲁棒性分别提升 0.53 个百分点、0.06 个百分点. 然而，DBFAT 由于利用 PGD 攻击步数衡量到决策边界距离方式的不准确性而导致整体效果甚至远不如 FAT，准确性降低 8.34 个百分点，白盒攻击和黑盒攻击上的鲁棒性分别降低 3 个百分点、5.27 个百分点. 不同的是，BTFAT 利用决策空间收紧算法有效提升了准确性和对抗鲁棒性. 最终，BTFAT 相比 FedBVA，准确性进一步平均提升 1.8 个百分点，白盒攻击和黑盒攻击上的鲁棒性分别提升 3.8 个百分点和 0.62 个百分点，证明了 BTFAT 的有效性.

4.3 不同客户端参与比例下的性能对比

在图 2 中展示了在 Fashion-MNIST 数据集下分别选择客户端参与比例为 10%、50%、100% 进行联邦对抗训练时的整体效果对比，最后一个子图是前 9 个子图中数据的平均值. 实验结果表明，随着客户端参与比例的增多，全局模型融合了更多的知识，不论是干净样本还是对抗样本上的准确性均有所提升. 然而，在默认设置下各参与方数据均匀分布，因此性

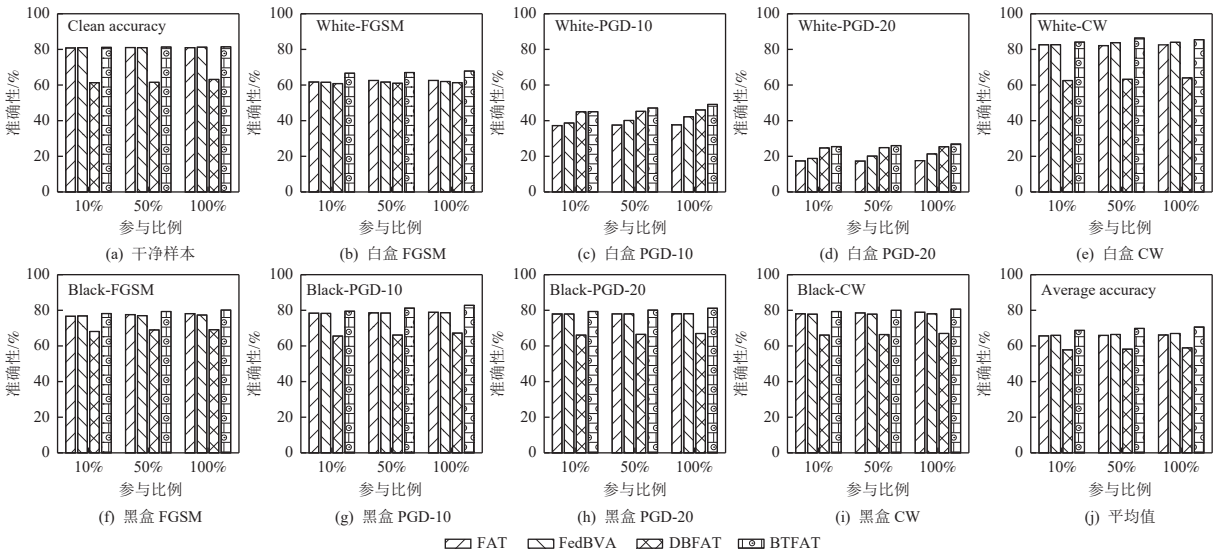


Fig. 2 Performance comparison of each method under different client participation ratios

图 2 不同客户端参与比例下各方法的性能对比

能提升不够明显. 另外, BTFAT 在干净样本准确性和对抗鲁棒性方面均表现最为优越. 在客户端全参与时, 其干净样本准确性达到 81.45%, 显著高于其他方法, 尤其比 DBFAT 高出超过 18 个百分点, 说明其在保持模型基本性能方面更为稳健. 不仅如此, 在白盒 PGD-10 攻击下, BTFAT 的准确性由 10% 客户端参与时的 45.03% 提升至 49.12%, 提升幅度达 4.1 个百分点, 高于 FAT 的 0.42 个百分点、DBFAT 的 1.03 个百分点和 FedBVA 的 3.5 个百分点. 在黑盒 PGD-10 攻击下, 其准确性提升至 82.78%, 比 FAT 高出近 4 个百分点. BTFAT 的优势源于其引入了类内样本聚集、类间样本疏远机制, 增强了模型对扰动结构的感知能力, 实现了在准确性与鲁棒性之间的双重优化.

4.4 不同非独立同分布程度下的性能对比

在表 3 中展示了在 Fashion-MNIST 数据集下利用狄利克雷分布的参数β实现不同非独立同分布数据分布时各方法的准确性和鲁棒性的效果对比. 参数β越大, 参与联邦训练的各客户端数据分布越均衡. 整体而言, 准确性和鲁棒性随着β的减小, 效果也逐渐变差. 数据的非独立同分布影响了对抗样本攻击防御的性能. 然而, 即使在不同的β下, BTFAT 依旧保持超过其他基线的准确性和鲁棒性. 平均而言, BTFAT 在干净样本上的准确性平均提升 6.91 个百分点, 白盒攻击上的鲁棒性提升 6.19 个百分点, 黑盒攻击上的鲁棒性提升 5.59 个百分点, 进一步证明了 BTFAT 的先进性.

4.5 联邦对抗训练的收敛性分析

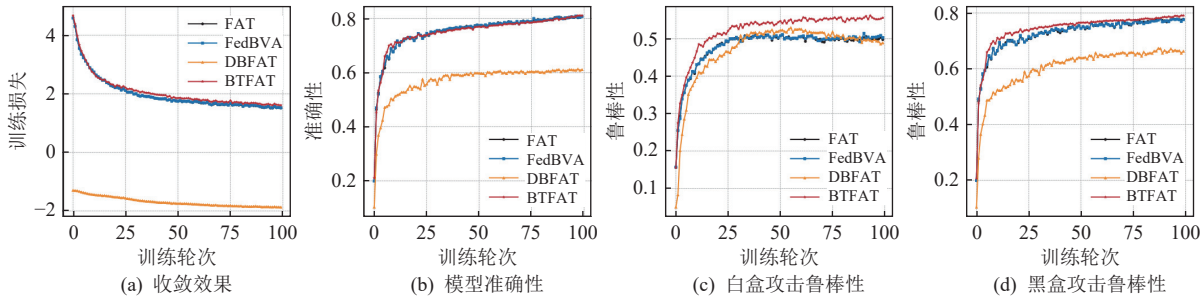
在图 3 中展示了在 Fashion-MNIST 数据集下各方法经过 5 次平均以后的训练结果. 在训练过程中, BTFAT 和 FedBVA、FAT 具有类似的收敛曲线. 它们在前 25 轮内快速收敛至低损失值, 同时损失曲线较为光滑, 可见算法的稳定性. DBFAT 则由于重加权的设计, 导致训练损失曲线值整体明显低于其余算法, 收敛速度也较慢. 从训练的收敛效果来看, BTFAT 不

Table 3 Performance Comparison on Fashion-MNIST Dataset Under Different Degrees of Non-IID Distribution

表 3 不同非独立同分布程度下 Fashion-MNIST 数据集的性能比较

β	算法	准确性	白盒攻击下的鲁棒性				黑盒攻击下的鲁棒性			
		干净样本	FGSM	PGD-10	PGD-20	CW	FGSM	PGD-10	PGD-20	CW
0.1	FAT	0.802 9	0.602 7	0.368 4	0.156 0	0.820 5	0.761 5	0.771 7	0.765 9	0.771 1
	FedBVA	0.804 6	0.608 4	0.372 9	0.172 7	0.823 1	0.763 0	0.771 8	0.767 2	0.771 3
	DBFAT	0.598 9	0.601 1	0.424 5	0.233 0	0.617 6	0.676 9	0.640 8	0.636 3	0.636 4
	BTFAT（本文）	**0.806 5**	**0.657 7**	**0.443 1**	**0.243 9**	**0.834 0**	**0.780 4**	**0.790 5**	**0.788 3**	**0.788 8**
0.5	FAT	0.803 1	0.606 4	0.371 2	0.160 0	0.821 8	0.763 7	0.780 4	0.776 1	0.775 7
	FedBVA	0.805 5	0.608 9	0.376 4	0.177 9	0.824 2	0.764 3	0.780 8	0.775 9	0.774 9
	DBFAT	0.610 8	0.602 9	0.426 9	0.237 8	0.622 3	0.671 4	0.647 8	0.643 3	0.652 6
	BTFAT（本文）	**0.807 3**	**0.660 3**	**0.446 9**	**0.253 8**	**0.837 7**	**0.780 9**	**0.792 2**	**0.791 0**	**0.789 9**
0.7	FAT	0.804 2	0.614 3	0.371 4	0.165 1	0.823 3	0.764 5	0.781 8	0.775 2	0.770 3
	FedBVA	0.806 5	0.613 4	0.380 9	0.182 3	0.824 3	0.765 1	0.781 2	0.775 6	0.770 4
	DBFAT	0.611 1	0.604 1	0.439 9	0.240 7	0.624 2	0.675 6	0.650 2	0.653 8	0.653 1
	BTFAT（本文）	**0.809 4**	**0.662 4**	**0.449 3**	**0.254 0**	**0.839 9**	**0.781 0**	**0.792 8**	**0.790 1**	**0.791 1**

注：黑体数值表示最优性能结果.



(a) 收敛效果 (b) 模型准确性 (c) 白盒攻击鲁棒性 (d) 黑盒攻击鲁棒性

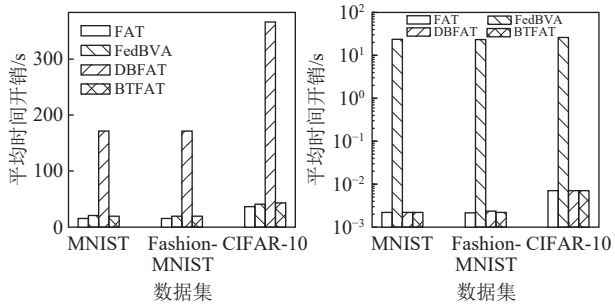
Fig. 3 Comparison of the convergence, performance, and robustness of different methods on Fashion-MNIST dataset

图 3 不同方法在 Fashion-MNIST 数据集上的收敛性、性能和鲁棒性的比较

仅提升了测试的准确性, 同时不论是在白盒攻击还是黑盒攻击下均优于其他基线. 可见, BTFAT 实现了稳定且可靠的收敛效果.

4.6 联邦对抗训练时间开销对比

在图 4 中展示了 3 种数据集下平均每一轮联邦对抗训练下各方法的本地训练时间和服务器聚合时间. 联邦对抗训练的算法设计主要是集中在客户端本地训练中, 因此该阶段时间开销上差异较大. DBFAT 在训练中需要计算所有样本到决策边界的 PGD 攻击步数, 导致巨大的时间开销. BTFAT 虽然需要计算嵌入表示间的距离, 但是依旧可以保持和 FAT、FedBVA 相近的较低开销. 由于 FedBVA 需要将中央服务器的聚合的泛化误差分解为偏差和方差, 而 BTFAT、FAT、DBFAT 只需要实现简单聚合即可, 因此 FedBVA 在服务器端具有远高于其他算法百倍的时间开销. 整体而言, BTFAT 具备出色的计算时间开销, 实现较为轻量化.



(a) 本地训练时间 (b) 服务器运行时间

Fig. 4 Average time cost per round for federated adversarial training of each method on three different datasets

图 4 在 3 个不同的数据集上比较每种方法每轮联邦对抗训练的平均时间开销

4.7 消融实验

在表 4 中展示了 3 种不同数据集下消除 BTFAT 中每一部分算法时的准确性、鲁棒性的变化. 决策空间收紧算法(Algo1)通过类内样本定位聚集初步规划出类内样本的大致空间, 然后根据类内样本收紧来缩小同类样本在决策空间中的距离并放大与不同类别样本间的距离, 实现了各类样本决策空间的最佳收紧. 权重惩罚优化算法(Algo2)将每轮的全局模型权重作为统一的最佳优化目标, 惩罚本地模型权重过度偏离全局模型的训练, 有效缓解了数据异质导致的联邦对抗训练效果不佳的难题. 平均来看, 单独加入决策空间收紧算法或是权重惩罚优化算法均比简单利用干净样本和对抗样本进行训练的效果更好. 在 Fashion-MNIST 数据集中, $\text{BTFAT}_{\text{NoAlgo2}}$ 比 $\text{BTFAT}_{\text{NoAll}}$ 准确性提升 0.19%, 鲁棒性提升 3.31%, 而 $\text{BTFAT}_{\text{NoAlgo1}}$ 比 $\text{BTFAT}_{\text{NoAll}}$ 的准确性还要低 0.41%, 因此决策空间收紧算法比权重惩罚优化算法起到更加关键的作用. $\text{BTFAT}_{\text{All}}$ 则在 3 种数据集和多重攻击下均保持超过 $\text{BTFAT}_{\text{NoAlgo1}}$、$\text{BTFAT}_{\text{NoAlgo2}}$、$\text{BTFAT}_{\text{NoAll}}$ 的性能, 缺失任何一个算法都无法达到 BTFAT 的最佳效果, 可见每个算法存在的必要性.

Table 4 Ablation Study Under Different Datasets and Adversarial Attack Strategies

表 4 不同数据集和对抗攻击策略下的消融实验

数据集	算法	准确性	白盒攻击下的鲁棒性	黑盒攻击下的鲁棒性
MNIST	$\text{BTFAT}_{\text{NoAll}}$	0.976 1	0.878 9	0.954 9
	$\text{BTFAT}_{\text{NoAlgo1}}$	0.976 6	0.888 0	0.956 9
	$\text{BTFAT}_{\text{NoAlgo2}}$	0.977 1	0.906 7	0.957 9
	$\text{BTFAT}_{\text{All}}$	**0.977 2**	**0.909 2**	**0.957 7**
Fashion-MNIST	$\text{BTFAT}_{\text{NoAll}}$	0.808 2	0.497 4	0.777 5
	$\text{BTFAT}_{\text{NoAlgo1}}$	0.804 1	0.500 2	0.777 0
	$\text{BTFAT}_{\text{NoAlgo2}}$	0.810 1	**0.553 6**	0.787 4
	$\text{BTFAT}_{\text{All}}$	**0.811 4**	0.553 1	**0.790 7**
CIFAR-10	$\text{BTFAT}_{\text{NoAll}}$	0.712 1	0.267 7	0.764 2
	$\text{BTFAT}_{\text{NoAlgo1}}$	0.771 2	0.300 2	0.764 1
	$\text{BTFAT}_{\text{NoAlgo2}}$	0.775 7	0.305 3	0.765 4
	$\text{BTFAT}_{\text{All}}$	**0.780 6**	**0.311 4**	**0.768 7**

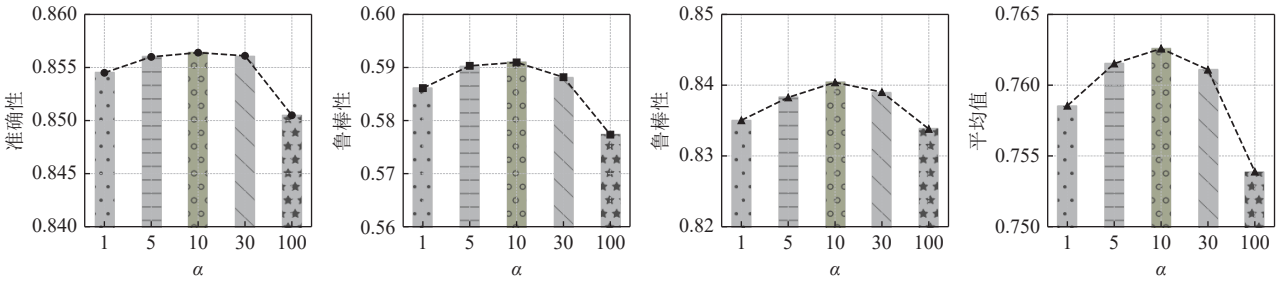
注: Algo1 为决策空间收紧算法, Algo2 为权重惩罚优化算法. “NoAll”表示 Algo1 和 Algo2 都不存在 BTFAT 中, “NoAlgo1”表示 Algo1 不存在 BTFAT 中, “NoAlgo2”表示 Algo2 不存在 BTFAT 中, “All”表示 Algo1 和 Algo2 都存在 BTFAT 中. 黑体数值表示最优实验结果.

4.8 超参数分析

在图 5 中展示了在 Fashion-MNIST 数据集下选择不同参数α时的准确性、平均鲁棒性和综合结果的变化. 综合结果计算为性能和鲁棒性的平均值. 其中图 5(d)中综合结果是前 3 项指标的平均. 随着参数α的增大, 权重惩罚优化算法将在对抗训练中占据更大的比重. 然而, 对抗非独立同分布的数据分布并不是联邦对抗训练的目标, 因此过大的α将会导致联邦对抗训练的准确性和鲁棒性下降; 过小则会导致训练过程容易受到数据分布不均衡的影响, 如表 3 中所示的导致效果不佳. 最终参数α选择 10 作为默认值, 以实现更好的平衡.

5 结 论

与集中训练的模型一样, 基于分布式人工智能架构的联邦学习模型在推理过程中容易遭受对抗样



(a) 模型准确性 (b) 白盒攻击鲁棒性 (c) 黑盒攻击鲁棒性 (d) 加权结果

Fig. 5 Performance, robustness, and integrated results under different parameters α on Fashion-MNIST dataset with multiple average computations

图 5 在 Fashion-MNIST 数据集上进行多次平均计算的性能、鲁棒性和不同参数α下的综合结果

本攻击，导致产生模型误判. 然而，联邦对抗训练作为最佳的解决策略主要面临鲁棒性与准确性权衡的限制，难以同时提升两者的性能，同时难以保证在数据非独立同分布时的效果. 针对该难题，我们从理论上分析出限制对抗训练鲁棒性和准确性增益的关键，提出了 BTFAT 框架. BTFAT 包括决策空间收紧算法和权重惩罚优化算法. 前者通过类内初始定位、类内决策空间距离缩小、类间决策空间距离增大，突破了联邦对抗训练鲁棒性和准确性间的权衡；后者作为辅助，惩罚过度偏离的本地对抗模型训练，保证了数据非独立同分布时的性能稳定. 我们不仅从理论上分析了 BTFAT 的收敛性，同时通过丰富的实验证明了方法的有效性、稳定性和高效性，为共同提升联邦对抗训练鲁棒性和准确性提供了一种新思路. 未来工作可聚焦于提升 BTFAT 框架在数据动态变化和复杂攻击场景下的稳定性与泛化能力，同时优化计算效率，并结合隐私保护技术增强框架的安全性.

作者贡献声明：闫彩虹负责方案整体设计并撰写论文；芦效峰提出指导意见并修改论文. 闫彩虹和芦效峰为共同一作.

参 考 文 献

[1] Le Junqing, Zhang Di, Lei Xinyu, et al. Privacy-preserving federated learning with malicious clients and honest-but-curious servers[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 4329−4344

[2] Wang Qinyong, Yin Hongzhi, Chen Tong, et al. Fast-adapting and privacy-preserving federated recommender system[J]. The VLDB Journal, 2022, 31(5): 877−896

[3] Yang Liu, Tan Ben, Zheng Vincent W, et al. Federated Recommendation Systems[M]//Federated Learning: Privacy and Incentive. Cham: Springer, 2020: 225−239

[4] Jiang Meirui, Wang Zirui, Dou Qi. HarmoFL: Harmonizing local and global drifts in federated learning on heterogeneous medical images[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022, 36(1): 1087−1095

[5] Chaddad A, Wu Yihang, Desrosiers C. Federated learning for healthcare applications[J]. IEEE Internet of Things Journal, 2023, 11(5): 7339−7358

[6] Ye Rui, Wang Wenhao, Chai Jingyi, et al. OpenfedLLM: Training large language models on decentralized private data via federated learning[C]//Proc of the 30th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2024: 6137−6147

[7] Cai Dongqi, Wu Yaozong, Wang Shangguang, et al. Efficient federated learning for modern NLP[C]//Proc of the 29th Annual Int Conf on Mobile Computing and Networking. New York: ACM, 2023: 1−16

[8] Le Junqing, Tan Zhouyong, Zhang Di, et al. Secure and efficient federated learning for continuous IoV data sharing[J]. Journal of Computer Research and Development, 2024, 61(9): 2199−2212 (in Chinese)
(乐俊青，谭州勇，张迪，等. 面向车联网数据持续共享的安全高效联邦学习[J]. 计算机研究与发展，2024，61(9)：2199−2212)

[9] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint, arXiv: 1412.6572, 2014

[10] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint, arXiv: 1312.6199, 2013

[11] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint, arXiv: 1706.06083, 2017

[12] Zhou Chunyi, Chen Dawei, Wang Shang, et al. Research and challenge of distributed deep learning privacy and security attack[J]. Journal of Computer Research and Development, 2021, 58(5): 927−943 (in Chinese)
(周纯毅，陈大卫，王尚，等. 分布式深度学习隐私与安全攻击研究进展与挑战[J]. 计算机研究与发展，2021，58(5)：927−943)

[13] Bhagoji A N, Chakraborty S, Mittal P, et al. Analyzing federated learning through an adversarial lens[C]//Proc of Int Conf on Machine Learning. Long Beach, California: PMLR, 2019: 634−643

[14] Huang Wenke, Ye Mang, Shi Zekun, et al. Federated learning for generalization, robustness, fairness: A survey and benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 9387−9406

[15] Papernot N, McDaniel P, Sinha A, et al. Sok: Security and privacy in

machine learning[C]//Proc of 2018 IEEE European Symp on Security and Privacy (EuroS&P). Piscataway, NJ: IEEE, 2018: 399−414

[16] Zizzo G, Rawat A, Sinn M, et al. Fat: Federated adversarial training[J]. arXiv preprint, arXiv: 2012.01791, 2020

[17] Zhou Yao, Wu Jun, Wang Haixun, et al. Adversarial robustness through bias variance decomposition: A new perspective for federated learning[C]//Proc of the 31st ACM Int Conf on Information & Knowledge Management. New York: ACM, 2022: 2753−2762

[18] Zhang Jie, Li Bo, Chen Chen, et al. Delving into the adversarial robustness of federated learning[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023, 37(9): 11245−11253

[19] Tsipras D, Santurkar S, Engstrom L, et al. Robustness may be at odds with accuracy[J]. arXiv preprint, arXiv: 1805.12152, 2018

[20] Kannan H, Kurakin A, Goodfellow I. Adversarial logit pairing[J]. arXiv preprint, arXiv: 1803.06373, 2018

[21] Dong Yinpeng, Liao Fangzhou, Pang Tianyu, et al. Boosting adversarial attacks with momentum[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 9185−9193

[22] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2574−2582

[23] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]/Proc of 2017 IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 39−57

[24] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[C]//Proc of Int Conf on Machine Learning. Long Beach, California: PMLR, 2020: 2206−2216

[25] Chen Cheng, Kailkhura B, Goldhahn R, et al. Certifiably-robust federated adversarial learning via randomized smoothing[C]//Proc of the 18th Int Conf on Mobile Ad Hoc and Smart Systems (MASS). Piscataway, NJ: IEEE, 2021: 173−179

[26] Zhang Hongyang, Yu Yaodong, Jiao Jiantao, et al. Theoretically principled trade-off between robustness and accuracy[C]//Proc of Int Conf on Machine Learning. Long Beach, California: PMLR, 2019: 7472−7482

[27] Ding G W G, Sharma Y, Liu K Y C, et al. MMA training: Direct input space margin maximization through adversarial training[J]. arXiv preprint, arXiv: 1812.02637, 2018

[28] Lee S, Lee H, Yoon S. Adversarial vertex mixup: Toward better adversarially robust generalization[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 272−281

[29] Shah D, Dube P, Chakraborty S, et al. Adversarial training in communication constrained federated learning[J]. arXiv preprint, arXiv: 2103.01319, 2021

[30] Hong Junyuan, Wang Haotao, Wang Zhangyang, et al. Federated robustness propagation: Sharing adversarial robustness in federated learning[J]. arXiv preprint, arXiv: 2106.10196, 2021

[31] Cretu G F, Stavrou A, Locasto M E, et al. Casting out demons: Sanitizing training data for anomaly sensors[C]//Proc of 2008 IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2008: 81−95

[32] Phuong M, Lampert C. Towards understanding knowledge distillation[C]//Proc of Int Conf on Machine Learning. Long Beach, California: PMLR, 2019: 5142−5151

[33] Ross A, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018, 36(1): 1660−1669

[34] Shen Shiwei, Jin Guoqing, Gao Ke, et al. APE-GAN: Adversarial perturbation elimination with GAN[J]. arXiv preprint, arXiv: 1707.05474, 2017

[35] Bartlett P L, Jordan M I, McAuliffe J D. Convexity, classification, and risk bounds[J]. Journal of the American Statistical Association, 2006, 101(473): 138−156

[36] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping[C]//Proc of 2006 IEEE Computer Society Conf on Computer Vision and Pattern Recognition (CVPR'06). Piscataway, NJ: IEEE, 2006, 2: 1735−1742

[37] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 2002, 86(11): 2278−2324

[38] Xiao Han, Rasul K, Vollgraf R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms[J]. arXiv preprint, arXiv: 1708.07747, 2017

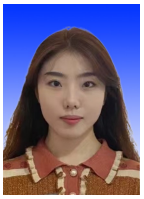
[39] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[R]. Toronto, Canada: University of Toronto, 2009

[40] He Kaiming, Zhang Xianyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[41] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint, arXiv: 1409.1556, 2014

[42] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1251−1258

[43] Sandler M, Howard A, Zhu Menglong, et al. MobileNetV2: Inverted residuals and linear bottlenecks[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4510−4520



Yan Caihong, born in 1997. PhD candidate. Her main research interest includes attack robustness and privacy security in federated learning.

闫彩虹, 1997 年生. 博士研究生. 主要研究方向为联邦学习的攻击鲁棒性与隐私安全.



Lu Xiaofeng, born in 1976. PhD, professor. Distinguished Member of CCF, Member of IEEE. His main research interests include artificial intelligence, cybersecurity, and privacy computing.

芦效峰, 1976 年生. 博士, 教授. CCF 杰出会员、IEEE 会员. 主要研究方向为人工智能、网络安全、隐私计算.