# DPBR-Adapt：具有层级自适应差分隐私的联邦学习防御方案

胡荣磊，白晨阳，魏占祯，韩妍妍，段晓毅，张浩
（北京电子科技学院电子与通信工程系，北京 100070）

**摘　要：**针对联邦学习中隐私泄露与投毒攻击并存的双重威胁，现有防御方案往往将隐私保护与鲁棒性视为独立模块，导致噪声添加盲目、防御精度受限。基于此，提出一种隐私保护与鲁棒性深度耦合的防御方案DPBR-Adapt。首先，在隐私保护维度，引入层级变异系数与训练进度感知因子，实现了层级差异化的噪声分配策略；在鲁棒性维度，设计了基于欧氏距离与余弦相似度的双重过滤机制，确保在强噪声干扰下仍能精确识别恶意更新。其次，建立了一套基于鲁棒统计量的自适应闭环机制，利用拜占庭检测输出的良性梯度中位数动态校准差分隐私的裁剪阈值，实现了隐私敏感度随环境风险的实时反馈调节。实验结果表明，DPBR-Adapt实现了两者在防御过程中的互补增强；在多种拜占庭攻击场景下，模型准确率较现有先进方案有显著提升，实现了更优的隐私–效用平衡与系统鲁棒性。

**关键词：**联邦学习；层级差异化；动量机制；双重过滤机制
**中图分类号：**TP309.2
**文献标志码：**A
**DOI：**10.11959/j.issn.1000–436x.2026071

# DPBR-Adapt: a hierarchically adaptive differential privacy defence scheme for federated learning

Hu Ronglei, Bai Chenyang, Wei Zhanzhen, Han Yanyan, Duan Xiaoyi, Zhang Hao
Department of Electronic and Communication Engineering, Beijing Electronic Science and Technology Institute, Beijing 100070, China

**Abstract:** To address the dual threats of privacy leakage and poisoning attacks in federated learning, existing defense mechanisms often treated privacy protection and robustness as independent modules, resulting in indiscriminate noise injection and limited defense precision. DPBR-Adapt, a defense scheme characterized by the deep coupling of privacy protection and Byzantine was proposed. Firstly, in the dimension of privacy, a hierarchical noise allocation strategy was implemented by introducing the layer-wise coefficient of variation and a training progress perception factor. In terms of robustness, a dual-filtering mechanism based on Euclidean distance and cosine similarity was designed to ensure the accurate identification and exclusion of malicious updates even under strong noise interference. Furthermore, a closed-loop adaptive mechanism based on robust statistics was established. This mechanism utilized the median of benign gradients, output by the Byzantine detection module, to dynamically calibrate the clipping threshold of differential privacy. Consequently, the privacy sensitivity was adjusted in real-time through a feedback loop based on environmental risk. Experimental results demonstrate that DPBR-Adapt achieves mutual reinforcement between defense processes. Under various Byzantine attack scenarios, the proposed scheme achieves a significant improvement in model accuracy compared to state-of-the-art methods, attaining a superior balance between privacy utility and systemic robustness.

**Keywords:** federated learning, hierarchical differentiation, momentum mechanism, dual-filtering mechanism

收稿日期：2025–12–04；修回日期：2026–03–12
通信作者：胡荣磊，hurl77225@163.com
基金项目：国家自然科学基金资助项目（No.62476013）；中央高校基本科研业务费专项资金资助项目（No.3282024058, No.3282024052）
**Foundation Items**：The National Natural Science Foundation of China (No.62476013), The Fundamental Research Funds for the Central Universities (No.3282024058, No.3282024052)

## 0　引言

联邦学习（federated learning, FL）[1]作为一种新兴的分布式机器学习范式，允许客户端在不共享原始数据的前提下协作训练全局模型，为打破医疗[2]、通信[3]、互联网[4]等领域的数据孤岛提供了有效途径。然而，随着FL应用的深入，其开放式的分布式架构也暴露了严峻的安全隐患：隐私泄露与拜占庭攻击（Byzantine attack，BA）的双重威胁已成为制约其落地应用的核心瓶颈。

一方面，尽管数据不出本地，但梯度的交互过程仍包含丰富的数据特征，攻击者可通过推断攻击反推用户敏感信息。另一方面，由于服务器无法验证本地数据的真实性，恶意参与者可发起拜占庭攻击[5-6]，即使是微量的恶意更新也足以破坏全局模型的收敛甚至发生控制模型的行为。

面对上述双重挑战，现有的防御研究往往陷入了简单叠加的技术困境[7]。具体而言，针对拜占庭攻击的鲁棒聚合规则[8-9]主要依赖基于均值或其变体等非鲁棒统计量来过滤异常，但这要求诚实梯度的分布相对集中；针对隐私保护的差分隐私(differential privacy, DP）技术[10]则通过引入噪声来掩盖特征。当两者被机械地结合时，随即产生了严重的内生冲突：DP引入的隐私噪声被人为地增大了梯度的离散度，使基于非鲁棒统计量的聚合规则失效，极易将良性噪声误判为恶意攻击而剔除，导致模型难以收敛[11-12]。

尽管近期已有部分工作尝试探索两者的结合，但仍未从根本上解决这一矛盾。Lyu[13]提出的DP-SIGNSGD（differentially private SignSGD）虽然结合了DP与符号聚合，但缺乏对梯度幅度的动态感知，在强攻击下精度波动剧烈；Zhu等[14]尝试了双重防护范式，但由于隐私预算与鲁棒阈值相互独立，无法在复杂攻击下动态平衡两者关系；Gu等[15]提出的DP-BREM（differentially-private and By-zantine-robust federated learning with client momentum）虽然引入了动量机制，但其静态的裁剪策略难以应对动态变化的拜占庭攻击，且严格的记录级DP导致了较大的效用损失。这些方案的共同局限在于将隐私保护与鲁棒性视为两个独立的模块，缺乏机制上的深层耦合与反馈。

此外，现有的防御方案[16]大多忽略了深度神经网络参数的“层级异构性”。具体而言，现有方法往往对模型所有层采用同质化的噪声添加策略。然而，研究表明，不同网络层对梯度的敏感度及模型收敛的贡献截然不同[17]。若不加区分地注入噪声，极易导致关键层信息淹没或非关键层保护不足。同时，在遭受拜占庭攻击时，恶意梯度会严重扭曲层级重要性的评估指标，导致自适应算法失效。因此，如何在保障鲁棒性的前提下，利用合理的层级特征来实现精准的噪声分配，是打破隐私–效用僵局的关键。

针对上述问题，本文提出了DPBR-Adapt，一种具有层级自适应特性的联邦学习防御方案。区别于现有策略，本文并未将隐私与鲁棒性视为独立的两个环节，而是构建了一种基于鲁棒统计量反馈的层级化自适应闭环机制，利用中位数等鲁棒统计量对异常值的不敏感特性，不仅能够有效抵御拜占庭攻击，还能为差分隐私提供稳定的敏感度估计。通过利用双重过滤机制输出的鲁棒统计量来校准层级重要性评估，再结合层级重要性与训练进度的动态感知，实现了噪声在不同网络层间的精细化动态分配，使隐私保护与鲁棒防御互补增强。本文的主要贡献归纳如下。

1) 提出了层级差异化的自适应噪声添加机制。针对神经网络不同层级对噪声敏感度的异构性，利用变异系数（coefficient of variation，CV）量化神经网络不同层级的敏感度差异，构建了尺度不变的层级重要性评估模型。结合层级功能重要性、训练进度因子与层位置权重，构建了三维动态噪声分配模型，同时引入动量机制使全局模型与局部更新更加平稳，平滑噪声影响，有效实现了对数据的隐私保护。

2) 提出了基于几何约束与方向一致性的双重过滤鲁棒聚合算法。本文模型结合欧几里得距离测量模型的几何偏差，同时利用余弦相似度测量模型方向的一致性，精确构建了梯度的“可信集合”，抵御拜占庭攻击。此外，引入信任评分机制对聚合过程进行加权平滑，进一步降低了残留异常值对全局模型的影响。

3) 提出了基于鲁棒统计量反馈的隐私–安全闭环机制。针对现有方案割裂隐私与鲁棒性的弊端，利用双重过滤后的鲁棒中位数动态校准差分隐私的裁剪阈值。这种机制赋予了隐私保护感知环境风险的能力：当攻击增强时，系统自动收紧敏感度边

界，从而在不增加隐私预算消耗的前提下，有效抑制恶意梯度导致的噪声膨胀，实现了隐私保护与鲁棒防御的内在统一与互补增强。

4) 在各种数据集上广泛测试了DPBR-Adapt模型的性能。与基线模型相比，本文模型在图像分类任务中模型精度提高了6.3%，收敛速率提升了20%，面对拜占庭攻击时，模型精度提高了27.14%。

## 1　相关知识

### 1.1　联邦学习

联邦学习通用模型如图1所示。在传统联邦学习过程中，首先由服务器端向各客户端传递初始全局模型，用户可以根据全局模型进行本地数据的训练，提取特征。

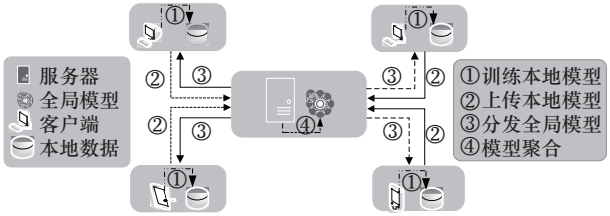



图1　联邦学习通用模型

客户端完成本地更新后，上传梯度参数，服务器根据各客户端上传的信息进行聚合，更新模型参数信息，形成新的全局模型，循环以上步骤，直至满足收敛性要求。

研究表明，即使客户端只上传梯度参数，也存在隐私泄露的可能性，这些参数本质上是本地数据训练的映射，包含本地数据的部分特征，这些数据特征在某些情况下是可以提取出来的，从而对用户隐私安全构成威胁。因此传统的联邦学习存在巨大的安全隐患。

### 1.2　差分隐私

差分隐私在机器学习中，通过添加噪声来掩盖一部分真实信息，即攻击者在参数分析的基础上对原始数据的了解是有限度的，并随着噪声的增强，了解得越少；但随着噪声的增强，在训练过程中所能获取的真实信息也越少。因此，在差分隐私的使用过程中，联邦学习的隐私–效用平衡是非常重要的研究内容。

**定义1**　设有一个随机算法A，其输入是一个数据集$D$，输出是一个随机变量$\mathrm{A}(D)$。如果两个数据集$D$和$D'$是邻近的（即它们仅在一个数据点上不同），则算法A满足$(\varepsilon,\delta)$–差分隐私，当且仅当式(1)成立，即

$$\Pr\left[\mathrm{A}(D) \in S\right] \leqslant \mathrm{e}^{\varepsilon}\Pr\left[\mathrm{A}(D') \in S\right] + \delta \quad (1)$$

其中，$\varepsilon$和$\delta$是隐私参数，$\varepsilon$和$\delta$越小，代表添加的噪声越大，隐私机制越严格。

### 1.3　联合防御

恶意客户端通过上传精心构造的恶意梯度向量，旨在将全局梯度的更新方向从目标损失函数的负梯度方向偏移，从而达到破坏全局模型的效果。即便系统中仅存在极少数的拜占庭节点，若缺乏有效的防御手段，传统的算术平均聚合也会被恶意偏差完全主导，导致训练任务彻底失败。因此，研究具有拜占庭鲁棒性的聚合规则显得至关重要。然而，单纯的鲁棒聚合在面对差分隐私环境时会面临新的挑战：DP引入的随机噪声会模糊正常梯度与拜占庭梯度之间的界限，使传统的基于几何距离或统计分布的防御方案准确率大幅下降。因此，设计一种既能抵御拜占庭攻击、又能适应差分隐私环境下的噪声扰动的内在统一防御机制，是本文DPBR-Adapt方案的核心动力。

## 2　具有层级自适应差分隐私的联邦学习

本节将详细介绍DPBR-Adapt联邦学习框架并概述整个系统工作流程。

### 2.1　系统模型

DPBR-Adapt由服务器、若干客户端组成，其中部分客户端被选为拜占庭节点，作为攻击者。DPBR-Adapt联邦学习模型如图2所示。

该联邦学习模型通过一个多阶段的客户端–服务器交互流程来运作，旨在增强模型的鲁棒性和隐私保护。整个流程可以概括为4个主要阶段。

1) 客户端本地训练与更新。服务器将当前的全局模型分发给选定的客户端，每个客户端在本地数据集上独立训练模型。这个本地训练过程确保了原始数据不会离开设备，从而保护了用户隐私。训练完成后，客户端将更新后的模型参数（如权重或梯度）传回服务器。

2) 服务器更新处理与过滤。服务器接收到客户端更新后，并不会直接聚合，而是启动一个精细的预处理流程；服务器评估模型不同层的重要性，以应对非独立同分布数据导致的客户端模型差异，

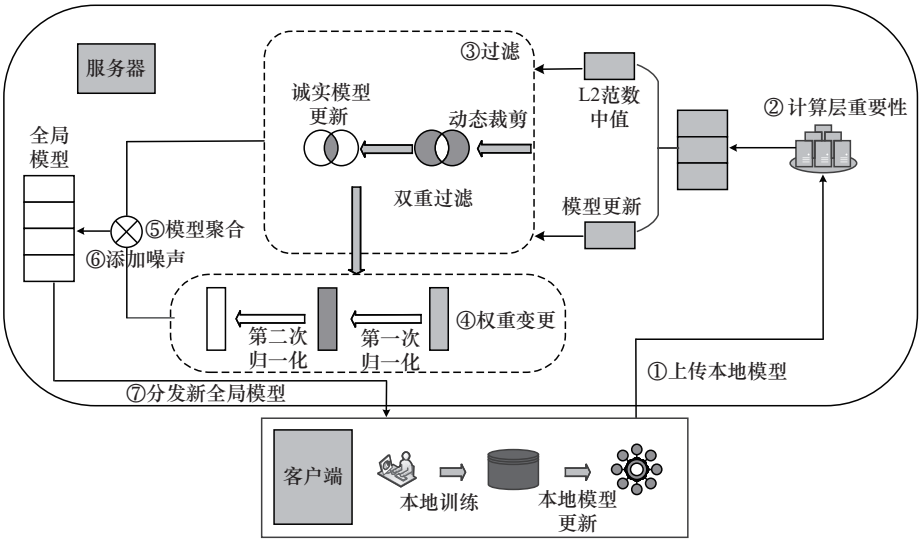



图2　DPBR-Adapt联邦学习模型

从而更有效地解决“客户端漂移”问题。服务器应用多层过滤机制来识别和剔除异常或恶意更新。这通常包括基于L2范数的离群值剔除，该方法通过比较更新的范数中位数来识别具有异常大范数的更新。然后框架对更新执行裁剪操作，限制其范数上限。这一步不仅能防御模型中毒攻击，也是实现差分隐私的关键先决条件，因为它能限制数据敏感度，从而校准后续添加的噪声。

3) 模型聚合与归一化。经过过滤和裁剪的“诚实模型更新”进入聚合阶段。该框架采用一种先进的归一化方法，这个过程基于客户端训练损失的归一化和反转来动态加权客户端贡献，从而有效应对数据异质性问题。

服务器将经过处理和加权的更新进行聚合，生成一个融合了所有良性客户端知识的新全局模型。为了提供正式的隐私保证，服务器向最终的聚合模型中添加经过校准的噪声。这是中央差分隐私的一种实现方式，通过扰动聚合结果来防止从模型更新中推断出敏感的用户信息。

4) 全局模型分发。服务器将经过隐私保护处理的新全局模型分发给客户端，作为下一轮训练的起点。

### 2.2　威胁模型

假定本文中有一个半诚实的服务器，即服务器会严格遵守既定的联邦学习协议，忠实地执行本文的双重过滤机制、鲁棒统计量计算及自适应噪声注入算法。然而，服务器可能具有分析客户端上传的梯度的动机，试图推断用户的私有属性。

隐私威胁：即使数据不离开本地，客户端上传的梯度向量仍可能通过推理攻击或成员重构攻击导致隐私泄露。此外，服务器无法访问客户端的本地训练数据，但可能有动机推断客户端的私人信息。因此，隐私威胁主要来自好奇的服务器以及在训练结束后能够访问全局模型的任何第三方。

拜占庭威胁：拜占庭攻击的目标是破坏全局模型的收敛性。只考虑恶意客户端作为拜占庭攻击的对手，同时假设系统中的恶意客户端占比不超过总参与者的半数。因为服务器的主要目标是训练一个鲁棒模型，没有实现拜占庭攻击的动机。这些恶意客户端可以进行任何行为，并且完全控制其本地训练数据和提交给服务器的数据，但无法直接访问或篡改诚实客户端的本地数据。

### 2.3　自适应差分隐私算法

本文模型在服务器端经过过滤选取出可信集合后，按照动态权重和层级重要性实现了层级差异化的隐私保护。差分隐私在联邦学习中已有广泛应用，传统方法通常采取固定的隐私预算和噪声的注入策略，但会导致模型性能下降。近年来，自适应差分隐私方法开始受到关注，但现有方法往往只关注单一维度的自适应调整。本文模型在层级重要性、训练进度和噪声分布等多个维度上实现自适应调整，提供更全面的隐私保护。自适应差分隐私算法通过多维度动态调整机制，在保护数据隐私的同时优化模型性能。本文模型结合了层级重要性分析、训练进度感知和自适应噪声注入，实现了隐私保护与模型性能的平衡。

由于 DPBR-Adapt 的服务器端被假设为可信的，因此通过服务器端产生高斯噪声可以保证差分隐私的安全性。高斯机制是实现差分隐私的主要技术之一，对于近似确定性实值函数$f$：$D\to R$，高斯机制定义为

$$M(D)=f(D)+\mathcal{N}\left(0,s_f^2\sigma^2\right) \tag{2}$$

其中，$\mathcal{N}\left(0,s_f^2\sigma^2\right)$表示从高斯分布中提取的噪声；$s_f$是函数$f$的敏感度，定义为任意相邻数据集$D$和$D'$之间的最大绝对距离，即

$$s_f=\max_{D,D'}\left|f(D)-f(D')\right| \tag{3}$$

根据研究，当$\delta\geqslant 20\mathrm{e}^{-\frac{(\delta\varepsilon)^2}{2}}$且$\varepsilon<1$时，机制$M$满足$(\varepsilon,\delta)$−差分隐私[18]。

本文对传统高斯机制进行拓展，提出层自适应高斯机制（layer-adaptive Gaussian mechanism, LAGM），该机制根据神经网络各层的特性和重要性，动态调整每层的噪声参数。形式上，层自适应高斯机制定义为

$$M_{\mathrm{LAGM}}(D)=f(D)+\mathcal{N}\left(0,s_{f,l}^2\sigma_l^2\alpha_t\right) \tag{4}$$

其中，$s_{f,l}$是第$l$层的敏感度，$\delta_l$是第$l$的基础噪声标准差，$\alpha_t$是在训练轮次$t$时的时间适应因子。

在 DPBR-Adapt 实际应用层自适应高斯机制时，LAGM 首先引入了变异系数（CV）作为评估层重要性的统计指标，对每一层进行计算，表达式为

$$\mathrm{CV}(i)=\frac{\sigma\left(N_i\right)}{\mu\left(N_i\right)} \tag{5}$$

其中，$\mathrm{CV}(i)$表示的是层变异系数，$\sigma\left(N_i\right)$是层范数的标准差，$\mu\left(N_i\right)$是层范数的均值。这一指标通过分析各层参数范数的离散程度，量化不同层对模型性能的贡献，将各层按照贡献分为4个层级；并提出时间适应因子，该因子用于分配各层的隐私保护资源，各层在不同的训练阶段需要不同的时间适应因子用于调配资源。具体如算法1所示。

**算法1** 自适应差分隐私层级划分

**输入** 初始模型$w_0$，客户端集合$K$，隐私预算$\epsilon,\delta$，裁剪阈值$\left[C_{\min},C_{\max}\right]$，总迭代轮数$T$

**输出** 最终全局模型$w_T$

1)初始化：服务端随机初始化 $w_0$

2) for $t$ = 1 to $T$ do

3) 收集客户端更新集合 $\mathcal{U}_t=\{\theta_1^t,\cdots,\theta_n^t\}$

4) 双重筛选：$d_{ij}=\|\theta_i^t-\theta_j^t\|_2$，$\cos\left(\theta_{ij}\right)$

5) 计算范数中位数：$M=\mathrm{median}\left(\{\|\theta_i^t\|_2|\theta_i^t\in\mathcal{U}_t\}\right)$

6) 信任评分 $\mu_i$：

$$\mu_i=\mathrm{e}^{\beta R_i}\left(\sum_{j=1}^{n}\mathrm{e}^{\beta R_j}\right)^{-1}$$

7) for每一参数层 $l$ = 1 to $L$ do

8) 计算范数分布$\mathcal{N}_l$与层重要性得分 $I_l^t$

9) end for

10) 计算重要性分位数$\{Q_1,Q_2,Q_3\}$

11) 执行算法2自适应差分隐私噪声添加

12) end for

13) return $w_T$

同时引入了训练进度因子$p_t$，通过时间维度的动态调整，更好地配合时间适应因子发挥作用。在训练初期，当训练进度因子较小时，为了提高模型收敛效率，需要对重要层添加较少的噪声，以便提取图形特征，上传较为有用的梯度参数。随着训练进度的加深，训练进度因子逐渐增大，越来越多的有用参数被聚合到全局模型中，因此需要更多的隐私保护资源去保护，此时自适应因子应分配更多的资源到重要层，以此实现更大程度的安全保护。这一机制的加入也解决了传统差分隐私在训练初期加入过度噪声难以收敛的问题。

传统的差分隐私方案通常预设固定的裁剪阈值，这在投毒攻击场景下难以兼顾隐私性与可用性。为此，本文提出了基于鲁棒统计反馈的动态敏感度机制。层自适应差分隐私的数学表现形式为

$$\sigma_{i,t}=\frac{\Delta f_t\sqrt{2\ln\left(\dfrac{1.25}{\sigma}\right)}}{\epsilon_t\sqrt{nBS_{\mathrm{type},i}S_{\mathrm{imp},i}S_{\mathrm{global},t}}} \tag{6}$$

其中，$\Delta f_t=\mathrm{clip}\left(M_{t-1},C_{\min},C_{\max}\right)$为动态敏感度函数，$M_{t-1}$是由双重过滤机制在上一轮迭代中产出的良性梯度范数中位数，$S_{\mathrm{type},i}$是层类型，$S_{\mathrm{imp},i}$是层的重要性，$S_{\mathrm{global},t}$是全局时间进度。全局敏感度通常受限于裁剪阈值，本文模型将双重过滤机制输出的鲁棒中位数范数$M$实时反馈给差分隐私模块。由于$M$是剔除异常偏差后的良性梯度特征表征，将其

作为动态裁剪阈值能够确保敏感度随当前全局模型的实际收敛状态和环境攻击风险实时调节，从而在防御投毒攻击的同时，避免注入过量噪声。

LAGM针对神经网络不同层的功能特性和隐私敏感度，通过层差异化裁剪策略精确控制敏感度$\Delta f$，对不同重要性的层采用不同的裁剪阈值。这种策略在保证差分隐私有界敏感度的同时，最大限度地保留了重要层的有用信息，显著提高了模型的学习效率。

LAGM的计算复杂度为$O(L \times N + L \times D)$，其中，$L$为网络层数，$N$为客户端数量，$D$为最大层参数维度。相比传统高斯机制的$O(L \times D)$，LAGM增加了与客户端数量相关的计算开销，但这在实际应用中是可接受的，且不影响客户端的本地计算负担。具体如算法2所示。

**算法2**　自适应差分隐私噪声添加

**输入**　更新集合$\mathcal{U}_t$，动量因子$m_t$，层系数$\alpha_l$，层得分$\{I_l\}$，当前隐私预算$\epsilon_t$，反馈参数$M_{t-1}$,$\mu_i$

**输出**　聚合后的全局模型$w_t$

1) 动态敏感度校准

2) 根据层重要性$I_l$分段调整裁剪阈值：$\Delta f_t \leftarrow \mathrm{clip}\left(M_{t-1}, C_{\min}, C_{\max}\right)$

3)自适应梯度裁剪$\theta_{i,l}^{*} \leftarrow \theta_{i,l}^{t} \cdot \min\left(1, \dfrac{\left(M_{t-1} I_l\right)}{\|\theta_{i,l}^{t}\|_2}\right)$

4) 差异化加噪与聚合：$\sigma_{\text{final}} = \dfrac{S_t \sqrt{2 \ln\left(\dfrac{1.25}{\sigma}\right)}}{\epsilon_t n}$

$$[\alpha_l I_l];\ \Delta \overline{w}_l = \sum_{i=1}^{l} \left(\mu_i \theta_{i,l}^{*}\right) + \mathcal{N}\left(0, \sigma_{\text{final}}^2 I\right)$$

5) 模型更新：$w_t \leftarrow f\left(w_{t-1}, \Delta \overline{w}_l, m_t\right)$

6) return $w_t$

算法2满足$(\varepsilon, \delta)$－差分隐私。对于每个客户端，其贡献的敏感度通过自适应裁剪机制控制，噪声注入基于高斯机制实现。根据差分隐私的组成定理，整个算法满足$(\varepsilon, \delta)$－差分隐私，其中$\varepsilon$和$\delta$由算法参数控制。在联邦学习训练过程中引入差分隐私，能够很好地保证客户端数据的安全。

### 2.4　层级隐私预算分配方案

在模型添加噪声前，DPBR-Adapt需要对过滤后形成的可信集合进行综合评估，噪声强度根据拜占庭检测结果动态调整模型更新的隐私预算。隐私预算是差分隐私的核心参数，在联邦学习中需要经过训练轮数进行分配，确保累计泄露不超过总预算数。目前根据应用和目标，主流策略可分为3类：均匀分配、自适应分配、基于概率的分配。本文模型中，DPBR-Adapt提出根据层级重要性进行隐私预算分配，首先引入了变异系数作为评估层重要性的统计指标，通过统计所有客户端第$l$层参数L2范数的变异系数，量化该层参数的异构性，如式(7)所示。异构性越高，表明该层对客户端数据分布差异更敏感，需分配更多隐私预算以保留个性化特征。引入变异系数的主要依据在于其尺度不变性[19]。神经网络中不同层的参数量级差异较大，直接使用范数或方差会导致量级大的层主导重要性评估[20]。变异系数通过将标准差除以均值，消除了参数绝对量级的影响，能够更准确地衡量各层参数分布的相对离散程度，从而有效量化层级异构性。

$$I_l = \frac{\mathrm{std}\left(\left\{\|W_l^i\|_2\right\}_{i=1}^{N}\right)}{\mathrm{mean}\left(\left\{\|W_l^i\|_2\right\}_{i=1}^{N}\right)} \tag{7}$$

其中，$\|W_l^i\|_2$指第$i$个客户端在第$l$层模型参数的范数。$I_l$越大，表明该层对不同客户端的数据分布越敏感，包含的个性化信息和有用特征越多。从特征表征角度看，高波动通常意味着该层捕获了更多来自不同客户端的、具有高度个性化的特征信息[21]。因此，CV指标能比范数更精准地识别出受异构数据驱动最大的“敏感层”，从而在隐私保护中赋予其更高的优先级[22]。

然后，依据功能重要性和参数敏感度进行定级，因为层级机制的特殊性，模型需要根据实际使用场景、隐私保护需求对各层级噪声进行调整，层级重要性定级如表1所示。

**表1**　层级重要性定级

| 层类型 | 层重要性$\alpha_{\text{type}}(l)$ |
|---|---|
| 卷积层 | 0.9 |
| 全连接层 | 0.7 |
| 偏置层 | 0.3 |
| 批归一化层 | 0.1 |

设置训练进度因子，通过训练进度因子动态调整隐私预算分配策略，如式(8)所示。

$$p_t = \frac{t}{T}\gamma + (1-\gamma),\ \gamma = 0.8 \tag{8}$$

其中，$\gamma$用于控制调整幅度。

在层级中层位置的权重也相当重要，它为神经网络中不同位置的层赋予重要性系数。首层因为要直接处理原始输入，泄露重要隐私信息的风险极高；末层输出模型更新可能会影响到最终的模型精度。因此首末两层需要更多的隐私预算去保证隐私安全，如式(9)所示。

$$\omega_{\text{pos}}(l) = \begin{cases} 1.2, & \text{首层和末层} \\ 1.0, & \text{其他层} \end{cases} \tag{9}$$

层级隐私预算分配是根据层重要性和敏感度为主要因素进行分配，随着训练轮数的增大，训练因子也随着发生变化，从而动态调整各层级的隐私预算，以保证数据的隐私安全。

初始分配依据最初给定的变异系数、层级重要性和层位置权重分配初轮的隐私预算，如式(10)所示。

$$\varepsilon_l^{\text{init}} = \varepsilon_{\text{total}}\left(I_l \alpha_{\text{type}}(l)\omega_{\text{pos}}(l)\right)\cdot \left(\sum_{k=1}^{n} I_k \alpha_{\text{type}}(k)\omega_{\text{pos}}(k)\right)^{-1} \tag{10}$$

随着时间的推移、训练轮数的增多，训练进度因子开始变大，逐步向重要层添加更多的隐私预算，如式(11)所示。

$$\varepsilon_l^{\text{final}} = \varepsilon_l^{\text{init}} \cdot \left(1 + (1-p_t)\cdot 0.5\right) \tag{11}$$

基于最终预算计算噪声强度，如式(12)所示。

$$\sigma_l = \frac{C_l}{\varepsilon_l^{\text{final}}}\sqrt{2\ln\left(\frac{1.25}{\delta}\right)} \tag{12}$$

其中，$C_l$为裁剪阈值，$\delta$为失败概率上界（$10^{-5}$）。

层级自适应机制的核心在于优化隐私预算的利用效率。根据差分隐私组合定理，模型的总隐私消耗是各层消耗的组合。理论上，不同层对模型梯度的贡献度不同。通过层级自适应，将有限的隐私预算$\epsilon$倾斜分配给高敏感度层（即变异系数大的层），在保证整体满足$(\epsilon,\delta)$ – DP 的前提下，最大化梯度的有效信息保留率，从而在不牺牲隐私安全界限的情况下提升模型效用。

### 2.5 拜占庭防御

在联邦学习中，拜占庭攻击实际上是依靠在客户端嵌入恶意节点得以达成目标的，那些恶意节点会蓄意传播错误或者虚假的信息，以此来扰乱整个系统的正常运转。这种攻击的主要目的在于破坏系统的稳定性、一致性以及安全性，妨碍系统对数据进行有效的整合与处理，最终有可能致使系统性能出现下降，甚至完全失效。

在模型聚合与归一化前，DPBR-Adapt 构建了隐私–鲁棒深度耦合的过滤策略。该策略通过动态裁剪技术将模型更新约束在合理范围，并联合运用欧几里得距离与余弦相似度从几何偏差与方向一致性两个维度精确识别恶意更新。更重要的是，该策略在消除拜占庭攻击影响的同时，提取良性梯度统计特征范数中位数，并将其作为关键反馈参数实时传递至差分隐私模块。 这一设计使差分隐私的裁剪阈值能够根据防御层的实时检测结果进行动态校准，从而在确保系统鲁棒性的基础上，实现隐私保护强度的自适应调节。

首先假设存在$n$个客户端参与模型的更新，记为$\{\theta_1,\theta_2,\theta_3,\cdots,\theta_n\}$，将第$j$维的参数$\theta_i$记为$(\theta_i)_j$，并将中位数模型作为初始参考模型$\theta_{\text{MED}}$，如式(13)所示。

$$\left(\theta_{\text{MED}}\right)_j = \text{MED}\left(\{\left(\theta_i\right)_j\}_{i=1}^{N}\right) \tag{13}$$

中位数对异常值具备很好的鲁棒性，因此可以用来代表大多数用户更新的特征，用于拜占庭防御也有很好的效果，所有客户端模型更新$\theta_i$的L2 范数的中位数。

$$M = \text{MED}\left(\{\|\theta_i\|_2\}_{i=1}^{N}\right) \tag{14}$$

$$\|\theta_i\|_2 = \sqrt{\sum_{j=1}^{K}\left(\theta_i\right)_j^2} \tag{15}$$

为了保证每次更新都保持在一定幅度内，本模型引入了动态裁剪技术，如式(16)所示。

$$\widetilde{\theta}_i = \theta_i \cdot \min\left(1, \frac{M}{\|\theta_i\|_2}\right) \tag{16}$$

然后经过两次过滤（基于欧几里得距离过滤和基于余弦相似度过滤），来识别和消除偏离的模型更新。

计算裁剪后的模型更新$\widetilde{\theta}_i$与参考模型$\theta_{\text{MED}}$的欧氏距离。

$$d_i = \|\widetilde{\theta}_i - \theta_{\text{MED}}\|_2 \tag{17}$$

其中，$d_i$衡量几何偏差，值越大表明更新越可能是恶意的。

$$D_1 = \left\{ i | d_i \text{是前 } n - f - 1 \text{小的值} \right\} \tag{18}$$

基于欧几里得距离选择最接近的模型更新。

然后在余弦相似度的角度进行测量，保证方向的一致性。

$$a_i^{\text{dist}} = 1 - \frac{\widetilde{\theta}_i \cdot \theta_{\text{MED}}}{|\widetilde{\theta}_i|_2 \cdot |\theta_{\text{MED}}|_2} \tag{19}$$

选择余弦距离 $a_i^{\text{dist}}$ 最小的 $n$–$f$–1 个更新筛选。

$$D_2 = \left\{ i | a_i^{\text{dist}} \text{是前 } n - f - 1 \text{小的值} \right\} \tag{20}$$

最终选取 $D_1$ 与 $D_2$ 的交集来使两种过滤原则全部满足。

虽然这两种过滤机制能够显著降低恶意更新的干扰，但仍不能完全排除所有的恶意模型更新。为了进一步减轻其影响，本文模型提出了信任评分机制，通过为每个模型更新分配信任评分并将其用作聚合权值来提高模型更新的鲁棒性。

首先将 $n$ 个客户端参与模型的更新对应的 L2 范数记为 $\{\|\theta_1\|,\|\theta_2\|,\|\theta_3\|,\cdots,\|\theta_n\|\}$，通过计算每次更新的范数与所有模型更新的范数中位数的绝对差来对本地更新模型的偏差进行评估，如式(21)所示。

$$r_i = \left| M - \|\theta_i\| \right| \tag{21}$$

较大的偏差表示模型的更新与参考模型的偏差越大，一般意味着存在异常行为。为了量化这种偏差，基于偏差的反比关系为每次模型定义了初始信任分数，如式(22)所示。

$$R_i = \frac{1}{r_i} \tag{22}$$

为了消除各种模型更新之间尺度差异的影响，对初始信任分数应用第一次归一化，标准化信任分数，使所有更新之间保证公平。

$$R_i' = \frac{R_i}{\sqrt{R_1^2 + R_2^2 + \cdots + R_n^2}} \tag{23}$$

这一步归一化对于维持计算稳定性和公平性至关重要，特别是在由异构客户端数据分布引起的模型更新多样化的场景中。

为了进一步增强良性和异常模型更新之间的区别，本模型加入了第二次归一化，如式(24)所示。

$$\mu_i = \mathrm{e}^{\beta R_i} \left( \sum_{j=1}^{n} \mathrm{e}^{\beta R_j} \right)^{-1} \tag{24}$$

其中，$\beta$ 是控制分数分化的正平滑参数。该函数用于放大良性更新的信任分数，同时抑制潜在恶意更新的信任分数。通过调整 $\beta$，算法可以对良性更新的放大和异常更新的抑制进行微调。较大的 $\beta$ 值增加了分数之间的分离，使良性更新在其影响力上更具优势。这两次归一化的结合确保了信任分数的一致性且对恶意攻击者具备鲁棒性。

在聚合过程中，将每个模型更新乘以其相应的信任分数，然后对所有加权更新进行求和，最后将该和添加到全局模型中，如式(25)所示。

$$w_i = w_0 + \sum_{i=1}^{n} \mu_i \theta_i \tag{25}$$

其中，$w_0$ 为聚合前的全局模型，$\mu_i$ 为第 $i$ 次模型更新的信任分数，$\theta_i$ 为第 $i$ 次模型更新。

至此，双重过滤机制不仅实现了对恶意梯度的精准剔除，还产生了两个关键的反馈参数：良性中位数 $M$ 与信任评分 $\mu_i$。

在客户端完成本地训练后，将本地更新上传至服务器。服务器在接收到所有客户端更新后，首先，执行双重过滤机制以剔除恶意更新并构建可信集合，同时产生两个关键反馈项：良性梯度范数中位数与用户信任评分。随后，该方案利用良性中位数动态校准差分隐私的裁剪阈值，并结合层级重要程度与训练进度感知，实现噪声的精准自适应添加。这一过程确保了隐私敏感度随环境风险实时调节，实现了隐私保护与鲁棒性的深度耦合。最后，服务器依据信任评分执行加权聚合策略，更新全局模型并进入下一轮迭代。

## 3　理论分析

本文将探讨 3 个核心问题：收敛性分析、隐私分析、鲁棒性理论分析。这些理论分析为 DPBR-Adap 方案在实际应用中的高效性、安全性和可靠性提供了有力的理论支持，基本数学符号定义如表 2 所示。

### 3.1　收敛性分析

在联邦学习中，加入差分隐私的主要挑战在于差分隐私引入的噪声会导致模型的精度降低和阻碍模型的收敛。因此，为了保证模型的收敛，就必须彻底检查在特定场景下加入噪声后是否会导致模型发散。为了便于研究，本文引入两个关键性假设。

**表2 基本数学符号**

| 数学符号 | 定义 |
|---|---|
| $N$ | 客户端总数 |
| $m$ | 恶意客户端数量 |
| $w_t$ | 第t轮全局模型参数 |
| $T$ | 总训练轮次 |
| $t$ | 当前训练轮次 |
| $\tilde{g}_t^{(i)}$ | 第$i$个客户端在第$t$轮的梯度 |
| $C_l$ | 裁剪阈值 |
| $G$ | 梯度的全局上限 |
| $\mu$ | 强凸参数 |
| $L$ | 函数的光滑常数 |
| $\eta$ | 学习率 |
| $\alpha$ | 隐私增强因子 |
| $\gamma_1$ | 第$l$层的噪声缩放因子 |
| $p_t$ | 训练进度因子 |
| $s_t$ | 全局噪声缩放因子 |

**假设1** 对于全局损失函数$f$：$R^d \to R$，存在常数$\mu > 0$，使对任意参数$w_1, w_2 \in R^d$式(26)成立，且$f$满足$\mu$−强凸性条件。

$$f\left(w_2\right) \geqslant f\left(w_1\right) + \nabla f\left(w_1\right)^{\mathrm{T}}\left(w_2 - w_1\right) + \frac{\mu}{2}\|w_2 - w_1\|^2 \tag{26}$$

**假设2** 有界性。局部小批量随机梯度是无偏的，且随机梯度是有界的，即

$$\|\nabla f_i(w)\| \leqslant G \tag{27}$$

此外，对于诚实的客户端，梯度方差也是有界的，即

$$\mathrm{E}\left[\|\nabla f_i(w) - \nabla f(w)\|^2\right] \leqslant \sigma^2 \tag{28}$$

通过引入这两个关键假设，建立了必要的理论框架来分析所考虑的目标函数的性质和行为。梯度和强凸性的结合，为严格检查算法性能和收敛保证奠定了基础。同时为了保证收敛的严谨性，本文还需要以下几个引理。

**引理1** 自适应裁剪误差界。假设第$l$层的梯度范数有界，即满足$\|g_t^{(i),l}\| \leqslant G$。对于阈值为$C_l$的自适应裁剪操作，存在一个与层级特性相关的常数$\beta_l$，使裁剪带来的期望平方误差满足

$$\mathrm{E}\left[\|\widetilde{g_t^{(i),l}} - g_t^{(i),l}\|^2\right] \leqslant \beta_l \cdot \frac{G^2}{C_l^2} \tag{29}$$

其中，$\beta_l$是与层重要性相关的常数。

**证明** 由于$\widetilde{g_t^{(i),l}} = g_t^{(i),l}\min\left(1, \frac{C_l}{\|g_t^{(i),l}\|}\right)$，因此存在两种情形：当$\|g_t^{(i),l}\| \leqslant C_l$时，不需要经过裁剪，误差为0；当$\|g_t^{(i),l}\| > C_l$时，梯度的模被压缩至$C_l$，即$\widetilde{g_t^{(i),l}} = g_t^{(i),l}\frac{C_l}{\|g_t^{(i),l}\|}$。误差的范数为

$$\|\widetilde{g_t^{(i),l}} - g_t^{(i),l}\| = \|g_t^{(i),l}\left(\frac{C_l}{\|g_t^{(i),l}\|} - 1\right)\| = \|g_t^{(i),l}\| - C_l \tag{30}$$

对式(30)两边平方，得到该情形下的误差$\phi$为

$$\phi = \left(\|g_t^{(i),l}\| - C_l\right)^2 \tag{31}$$

根据假设2中的梯度有界性$\|g_t^{(i),l}\| \leqslant G$，对误差$\phi$进行放缩，即

$$\phi \leqslant \left(G - C_l\right)^2 \tag{32}$$

令$p_l = P\left(\|g_t^{(i),l}\| > C_l\right)$表示第$l$层梯度触发裁剪的概率，期望误差即两种情形的加权和。

$$\mathrm{E}[\phi] \leqslant p_l\left(G - C_l\right)^2 \tag{33}$$

对式(33)不等号右边进行代数变换

$$p_l\left(G - C_l\right)^2 = \left[p_l\frac{C_l^2\left(G - C_l\right)^2}{G^2}\right]\frac{G^2}{C_l^2} \tag{34}$$

此时，将$\beta_l$进行显式定义，即

$$\beta_l \triangleq p_l\frac{C_l^2\left(G - C_l\right)^2}{G^2} \tag{35}$$

其中，$\beta_l$均由层级$l$的统计特性及算法参数决定，是层重要性相关的常数。证毕。

调整裁剪阈值$C_l$使其与层的变异系数相关，通过为重要层分配更高的裁剪阈值，可以最大程度地保留其有用的信息，从而降低裁剪误差。

**引理2** 双重筛选保证。在双重筛选机制下，筛选后的梯度满足

$$\mathrm{E}\left[\|\overline{g_t} - \nabla f\left(w_t\right)\|^2\right] \leqslant \left(\sigma^2 + \sum_{l=1}^{L}\frac{\beta_l G^2}{C_l^2}\right)(N - m)^{-1} \tag{36}$$

**证明** 设$S$为筛选后的诚实客户端集合。由于双重筛选机制基于欧几里得距离和余弦相似度进行过滤，能够有效识别并排除恶意更新，因此可以保证$|S| \geqslant N - m$，其中，$N$为客户端总数，$m$为拜占庭客户端数量。则聚合后的梯度满足

$$\mathrm{E}\left[\|\overline{g_t}-\nabla f\left(w_t\right)\|^2\right]=\mathrm{E}\left[\|\frac{1}{|S|}\sum_{i\in S}\tilde{g}_t^{(i)}-\nabla f\left(w_t\right)\|^2\right]\leqslant \frac{1}{|S|}\mathrm{E}\left[\sum_{i\in S}\|\tilde{g}_t^{(i)}-\nabla f\left(w_t\right)\|^2\right] \tag{37}$$

利用引理 1 和有界性假设，并考虑双重筛选机制对恶意梯度的排除作用，可以得到上述结果。证毕。

**引理 3**　噪声方差界。自适应噪声的总方差满足

$$\mathrm{E}\left[\|n_t\|^2\right]=\sum_{l=1}^{L}d_l\left(\sigma_l^{(t)}\right)^2\leqslant\Sigma_t^2 \tag{38}$$

其中

$$\Sigma_t^2=\frac{2\ln\left(\frac{1.25}{\sigma}\right)}{\epsilon^2\alpha^2}s_t^2\sum_{l=1}^{L}d_l\gamma_l^2 \tag{39}$$

其中，$\gamma_l$是层级噪声缩放因子。

**证明**　自适应噪声$\eta_t$的标准差$\sigma_t$由多个因子共同决定，包括层敏感度、训练进度和层级重要性

$$\sigma_{i,t}=\frac{\Delta f_t\sqrt{2\ln\left(\frac{1.25}{\sigma}\right)}}{\epsilon_t\sqrt{nBS_{\mathrm{type},i}S_{\mathrm{imp},i}S_{\mathrm{global},t}}} \tag{40}$$

自适应噪声的方差为$\sigma_{i,t}^2=\frac{\left(\Delta f_t\right)^2\cdot 2\ln\left(\frac{1.25}{\sigma}\right)}{\epsilon_t^2 nBS_{\mathrm{type},i}S_{\mathrm{imp},i}S_{\mathrm{global},t}}\leqslant\Sigma_t^2$，通过层级差异化的噪声分配策略，在保持隐私预算的同时，动态调整噪声大小，以免噪声过度干扰导致模型收敛困难。通过证明噪声方差总和存在一个上界，可以确保整个训练过程中注入的噪声不会无限增大，从而保证了模型的收敛稳定性和可靠性。证毕。

**引理 4**　梯度偏差界。对于双重筛选后的梯度估计，存在常数$\kappa$使

$$\mathrm{E}\left[\|\overline{g_t}-\nabla F\left(w_t\right)\|^2\right]\leqslant\kappa\left(\frac{\sigma^2}{N-m}+\beta_{\max}\frac{G^2}{C_{\min}^2}\right) \tag{41}$$

**证明**　设$S$为筛选后的诚实客户端集合，由于双重筛选的鲁棒性保证，有$|S|\geqslant N-m$。利用范数不等式分解，表达式为

$$\mathrm{E}\left[\|\overline{g_t}-\nabla F\left(w_t\right)\|^2\right]\leqslant \underbrace{2\,\mathrm{E}\left[\|\frac{1}{|S|}\sum_{i\in\mathcal{S}}\left(g_{t,i}-\nabla F\left(w_t\right)\right)\|^2\right]}_{\text{项 }A\text{(方差)}}+ \underbrace{2\,\mathrm{E}\left[\|\frac{1}{|S|}\sum_{i\in\mathcal{S}}\left(\widetilde{g_{t,i}}-g_{t,i}\right)\|^2\right]}_{\text{项 }B\text{(裁剪偏差)}} \tag{42}$$

基于本地梯度的无偏性及方差$\sigma^2$的定义，由于客户端之间相互独立，有

$$\text{项 }A=\frac{1}{|S|^2}\sum_{i\in\mathcal{S}}\mathrm{E}\left[\|g_{t,i}-\nabla F\left(w_t\right)\|^2\right]\leqslant\frac{\sigma^2}{N-m} \tag{43}$$

根据引理 1 相关结论，有

$$\text{项 }B\leqslant\frac{1}{|S|}\sum_{i\in\mathcal{S}}\mathrm{E}\left[|\widetilde{g_{t,i}}-g_{t,i}|^2\right]\leqslant\beta_{\max}\frac{G^2}{C_{\min}^2} \tag{44}$$

将项$A$与项$B$的放缩结果合并，并令结论中的常数$\kappa=2$。证毕。

**引理 5**　自适应裁剪的收敛性改进。相比于固定的裁剪阈值$C$，自适应裁剪策略使重要层的收敛误差减少。

$$\sum_{l\in\mathcal{L}_{\mathrm{hg}}}\mathrm{E}\left[\|\widetilde{g_t^{(i),l}}-g_t^{(i),l}\|^2\right]\leqslant 0.25\mathrm{E}\left[\|g_{\mathrm{ci}}^{(i),l}-g_t^{(i),l}\|^2\right] \tag{45}$$

其中，$\mathcal{L}_{\mathrm{hg}}$是高重要性层的集合，$g_{\mathrm{ci}}^{(i),l}$是使用固定阈值$C$的裁剪结果。

**证明**　设$e_{\mathrm{ada}}$和$e_{\mathrm{fixed}}$分别为自适应裁剪和固定裁剪的误差，根据引理 1，自适应裁剪通过为重要层（其范数均值和标准差都较大）设置更大的裁剪阈值，使$C_{\mathrm{ada}}>C_{\mathrm{fixed}}$。在裁剪操作中，当$\|g\|$较大时，误差主要取决于$\|g\|^2$的大小，自适应裁剪通过更大的阈值保留了更多大范数梯度信息，其误差$\|g\|^2\left(1-\frac{C}{\|g\|}\right)^2$显著小于固定裁剪的误差。特别是当梯度范数远大于裁剪阈值时，自适应裁剪的误差比固定裁剪的误差要小得多，因为固定裁剪会强制将所有超出阈值的梯度范数削减到同一个值，从而丢失了相对幅度信息。因此，在预期意义下，自适应裁剪对高重要性层的误差减少是显著的。证毕。

**引理 6**　双重筛选的方差减少。双重筛选较单一筛选具有更好的方差控制。

$$\mathrm{Var}\left[\overline{g_t^{\mathrm{double}}}\right]\leqslant\rho\mathrm{Var}\left[\overline{g_t^{\mathrm{single}}}\right] \tag{46}$$

其中，$\rho\leqslant 0.8$是方差减少因子。

**证明**　双重筛选$D=D_1\cap D_2$，单一筛选依赖于一个标准（如欧氏距离），双重筛选结合了两个互补的标准：欧氏距离（衡量几何偏差）和余弦相似度（衡量方向一致性）。因此，一个恶意更新如

果通过了单一筛选，可能因为其方向与诚实客户端平均方向不一致而在双重筛选中被排除。设$\mathcal{E}_1$,$\mathcal{E}_2$分别为筛选器误选事件，则

$$\Pr\left[\text{误选}\right]=\Pr\left[\mathcal{E}_1\cup\mathcal{E}_2\right]\leqslant \Pr\left[\mathcal{E}_1\right]+\Pr\left[\mathcal{E}_2\right]-\Pr\left[\mathcal{E}_1\cap\mathcal{E}_2\right] \tag{47}$$

由于筛选器的互补性，$\Pr\left[\mathcal{E}_1\cap\mathcal{E}_2\right]$显著小于单独的概率，从而减小方差。证毕。

以上引理系统性地将梯度裁剪、拜占庭容错、差分隐私与数据异质性四大扰动因素纳入统一的上界框架，为主定理的线性收敛保证与隐私–效用–鲁棒性三角权衡提供严格的数学基础。

**定理 1** 模型的收敛性。在假设 1 和假设 2 成立的前提下，本文算法在存在拜占庭攻击和差分隐私噪声的情况下，其全局模型的收敛速度满足

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left|\left|\nabla f\left(w_{t-1}\right)\right|\right|^2\leqslant\frac{G^2}{\left(N-m\right)^2}+\frac{\sigma^2}{N-m}+\frac{C_t}{T}+\frac{R}{T} \tag{48}$$

**证明** 该证明遵循梯度下降算法的收敛性分析，通过结合前述引理来量化每轮迭代的误差。设$w_t$为第$t$轮训练后的全局模型参数。从损失函数$f\left(w_t\right)$的强凸性开始，得到

$$\mathrm{E}\left[f\left(w_t\right)\right]-f\left(w^*\right)\leqslant\left(1-\frac{\mu\eta}{2}\right)\mathrm{E}\left[f\left(w_{t-1}\right)\right]-f\left(w^*\right)+\frac{\eta}{2}\mathrm{E}\left|\left|\widetilde{g_{t-1}}\right|\right|^2 \tag{49}$$

其中，$\widetilde{g_{t-1}}$是经过双重筛选和自适应加噪的聚合梯度，为了进一步分析算法的收敛性质，需对聚合梯度二阶矩$\mathrm{E}\left[\|\widetilde{g_{t-1}}\|^2\right]$进行放缩。根据引理 3 和引理4，聚合梯度由真实梯度、拜占庭偏差及隐私噪声组成。利用杨氏不等式及独立性假设，可将其界定为

$$\mathrm{E}\left[\|\widetilde{g_{t-1}}\|^2\right]\leqslant\left(1+\varphi\right)\mathrm{E}\left[\|\nabla f\left(w_{t-1}\right)\|^2\right]+\left(\frac{G^2}{\left(N-m\right)^2}+\frac{\sigma^2}{N-m}\right) \tag{50}$$

其中，$\varphi$为与算法参数相关的常数。通过选取满足条件的学习率$\eta$（如$\eta\leqslant\dfrac{1}{L\left(1+\rho\right)}$），利用$L$–平滑性产生的负梯度项抵消掉误差项中的梯度项，从而将单步递归式简化为

$$\mathrm{E}\left[\|g_{t-1}\|^2\right]\leqslant\mathrm{E}\left[\|\nabla f\left(w_{t-1}\right)\|\right]+\frac{G^2}{\left(N-m\right)^2}+\frac{\sigma^2}{N-m} \tag{51}$$

将式(51)代入式(49)可得

$$\mathrm{E}\left[f\left(w_t\right)\right]-f\left(w^*\right)\leqslant\left(1-\frac{\mu\eta}{2}\right)\left(\mathrm{E}\left[f\left(w_{t-1}\right)\right]-f\left(w^*\right)\right)+\frac{\eta}{2}\left(\frac{G^2}{\left(N-m\right)^2}+\frac{\sigma^2}{N-m}\right) \tag{52}$$

为了求得$T$次迭代后的平均收敛性，对上述单步递归不等式进行 $t=1$ 到 $T$ 的累加。通过移项，整理得到关于函数残差的差分形式

$$\frac{\mu\eta}{2}\sum_{t=1}^{T}\left(\mathrm{E}\left[f\left(w_{t-1}\right)\right]-f\left(w^*\right)\right)\leqslant\left(\mathrm{E}\left[f\left(w_0\right)\right]-f\left(w^*\right)\right)-\left(\mathrm{E}\left[f\left(w_T\right)\right]-f\left(w^*\right)\right)+\frac{T\eta}{2}\left(\frac{G^2}{\left(N-m\right)^2}+\frac{\sigma^2}{N-m}\right) \tag{53}$$

利用$L$ –平滑性及Polyak-Lojasiewicz不等式性质，在强凸下，梯度范数与函数最优值差距满足$2\mu\left(f\left(w\right)-f\left(w^*\right)\right)\leqslant\|\nabla f\left(w\right)\|^2$。将此关系代入式(52)左侧，并将不等式右侧的$-\left(\mathrm{E}\left[f\left(w_T\right)\right]-f\left(w^*\right)\right)$项舍去，可得

$$\sum_{t=1}^{T}\frac{\eta}{4}\mathrm{E}\left[\|\nabla f\left(w_{t-1}\right)\|^2\right]\leqslant\left(\mathrm{E}\left[f\left(w_0\right)\right]-f\left(w^*\right)\right)+\frac{T\eta}{2}\left(\frac{G^2}{\left(N-m\right)^2}+\frac{\sigma^2}{N-m}\right) \tag{54}$$

将式(54)两边同时除以$T\dfrac{\eta}{4}$，并对系数进行合并处理，定义$R=\dfrac{4\left(\mathrm{E}\left[f\left(w_0\right)\right]-f\left(w^*\right)\right)}{\eta}$，并将推导过程中的常数余项记为$C_t$。证毕。

基于前面引理所建立的理论基础，进一步推导并证明了算法的最终收敛性质。这些定理明确量化了算法在不同条件下的性能，为收敛性与效用误差提供了理论上的最优性保证。

**推论 1** 线性收敛率。噪声项满足一定条件，即

$$\frac{\eta^2L}{2}\left(\left(\sigma^2+\sum_{l=1}^{L}\frac{\beta_lG^2}{C_l^2}\right)\left(N-m\right)^{-1}+\Sigma_t^2\right)\leqslant$$

$\frac{\eta\mu}{8}\mathrm{E}\left[f\left(w_t\right)-f\left(w^*\right)\right]$时，算法将以线性速率收敛，其收敛因子为$\left(1-\frac{\eta\mu}{8}\right)$，满足

$$\mathrm{E}\left[f\left(w_{t+1}\right)-f\left(w^*\right)\right]\leqslant\left(1-\frac{\eta\mu}{8}\right)\mathrm{E}\left[f\left(w_t\right)-f\left(w^*\right)\right] \tag{55}$$

**定理 2**　全局收敛性。在假设 1 和假设 2 的前提下，对于适当选择的学习率$\eta$，算法最终会收敛到最优解附近的一个邻域，其误差上界的表达式为

$$\lim_{T\to\infty}\mathrm{E}\leqslant\frac{4\eta L}{\mu}\left(\left(\sigma^2+\sum_{l=1}^{L}\frac{\beta_l G^2}{C_l^2}\right)(N-m)^{-1}+\Sigma_t^2\right) \tag{56}$$

其中，$\overline{\Sigma^2}=\lim\limits_{t\to\infty}\Sigma_t^2$是稳态噪声方差。

**证明**　在达到稳态时，假设模型达到平衡，即$\mathrm{E}\left[f\left(w_t+1\right)\right]=\mathrm{E}\left[f\left(w_t\right)\right]$，由定理 1 可得

$$0\geqslant-\frac{\eta\mu}{4}\mathrm{E}\left[f\left(w_\infty\right)-f\left(w^*\right)\right]+\frac{\eta^2 L}{2}\left(\left(\sigma^2+\sum_{l=1}^{L}\frac{\beta_l G^2}{C_l^2}\right)(N-m)^{-1}+\Sigma_t^2\right) \tag{57}$$

重新整理即可证明。

定理 2 建立了算法能够达到的最终收敛精度界限。本文算法通过自适应噪声注入和平滑动量机制，确保噪声不会压倒梯度更新，且双重筛选能够有效抵抗拜占庭攻击，使聚合误差被限制在合理范围内。最终，当总训练轮数$T$足够大时，模型能够线性收敛到全局最优解的某个邻域，其收敛速率取决于引理 2 和引理 3 所证明的各项误差之和。定理 2 表明，本文模型在严格的隐私保护和拜占庭鲁棒性要求下，能够保证模型的稳定收敛，并且收敛速度损失可控，从而实现了隐私–效用–鲁棒性三者的平衡。

### 3.2　隐私分析

本节将深入分析本文模型中的自适应机制如何从理论上提升收敛性能。以下定理和推论证明了本文的动态策略在收敛速度和精度上相较于静态方法具有显著优势。

本文模型采用的训练进度自适应策略使噪声方差$\Sigma_t^2$随时间递减，从而保证算法在训练后期具有更优的收敛精度。

$$\Sigma_t^2=\Sigma_0^2\cdot\left(0.1+0.7\cdot p_t\right)^2\leqslant\Sigma_0^2\cdot(0.8)^2 \tag{58}$$

其中，$p_t=\min\left(1,\frac{t}{0.7T}\right)$是训练进度因子。

**推论 2**　层级差异化。通过层级差异化噪声分配，本文模型能够有针对性地为重要层分配噪声预算，从而显著降低其收敛误差。

$$\sum_{l\in\text{important}}d_l\left(\sigma_l^{(t)}\right)^2\leqslant 0.5\sum_{l\in\text{important}}d_l\left(\sigma_{l\,\text{uniform}}^{(t)}\right)^2 \tag{59}$$

其中，$\sum\limits_{l\in\text{important}}d_l$是重要层的集合，$\sigma_{l\,\text{uniform}}^{(t)}$是均匀噪声分配的标准差。

**推论 3**　训练进度自适应的收敛加速。训练进度因子$p_t$的引入不仅改善了 FL 模型的精度，也显著提高了算法后期的收敛速度。推论 3 对加速进行量化，即

$$\lim_{t\to\infty}\frac{\mathrm{E}\left[f\left(w_{t+1}\right)-f\left(w^*\right)\right]}{\mathrm{E}\left[f\left(w_t\right)-f\left(w^*\right)\right]}\leqslant 1-\frac{\eta\mu}{8}\left(\frac{s_\infty}{s_0}\right)^2 \tag{60}$$

其中，$s_\infty=0.8$，$s_0=0.1$，因此收敛率改进了 64 倍。

本文模型通过对隐私预算的精细化动态管理，确保了在严格的隐私保护下，模型依然能够实现稳定、快速且高精度的收敛，从而在隐私与效用之间取得了卓越的平衡。

### 3.3　鲁棒性理论分析

除了全局收敛分析和隐私分析，同样需要对拜占庭攻击下的鲁棒性进行严谨的理论证明。以下定理和推论量化了本文模型所能容忍的恶意攻击上限，并证明了其对模型收敛的有限影响。

**定理 3**　拜占庭容错界。在最多$m<\frac{N}{3}$个恶意客户端存在的情况下，本文模型提出的加权几何聚合机制保证聚合后的梯度误差具有一个明确的上界。

$$\mathrm{E}\left[\|\overline{g_t}-\nabla f\left(w_t\right)\|^2\right]\leqslant\left(1+\frac{2m}{N-2m}\right)\frac{\sigma^2}{N-m} \tag{61}$$

**证明**　利用几何中位数性质和恶意客户端的有界影响进行分析，设$\mathcal{H}$为诚实客户端集合，$\mathcal{M}$为恶意客户端集合。基于几何中位数进行筛选，对于任意梯度集合$\{g_1,\cdots,g_N\}$，几何中位数$g_{\text{med}}$满足

$$g_{\text{med}}=\arg\min_g\sum_{i=1}^{N}\|g-g_i\|_2 \tag{62}$$

设筛选后的客户端集合为$S$，其中包含$h$个诚实客户端和$m'$个恶意客户端。对于诚实客户端$i\in\mathcal{H}$，其梯度满足

$$\mathrm{E}\left[\|g_i-\nabla f\left(w_t\right)\|^2\right]\leqslant\sigma^2 \tag{63}$$

筛选后的平均梯度为

$$\overline{g_t} = \frac{1}{|S|}\sum_{i\in S} g_i = \frac{1}{|S|}\left(\sum_{i\in S\cap\mathcal{H}} g_i + \sum_{i\in S\cap\mathcal{M}} g_i\right) \tag{64}$$

其中，诚实客户端比例$\rho = \frac{h}{|S|} \geqslant \frac{N-2m}{N-m}$。筛选后的梯度表示为

$$\overline{g_t} = \frac{h}{|S|}\overline{g_{\mathcal{H}}} + \frac{m'}{|S|}\overline{g_{\mathcal{M}}} \tag{65}$$

利用三角不等式和有界性假设，诚实客户端方差为

$$\mathrm{E}\left[\|\overline{g_{\mathcal{H}}} - \nabla f(w_t)\|^2\right] \leqslant \frac{\sigma^2}{h} \tag{66}$$

恶意客户端方差为

$$\|\overline{g_{\mathcal{M}}} - \nabla f(w_t)\| \leqslant 2G \tag{67}$$

结合权重关系，$\frac{h}{|S|} \geqslant \frac{N-2m}{N-m}$和$\frac{m'}{|S|} \leqslant \frac{m}{N-m}$在$G^2 \approx \sigma^2$假设下有

$$\mathrm{E}\left[\|\overline{g_t} - \nabla f(w_t)\|^2\right] \leqslant \left(1 + \frac{2m}{N-2m}\right)\frac{\sigma^2}{N-m} \tag{68}$$

证毕。

本文模型为聚合后的梯度误差设定了明确的上界，并量化了在拜占庭攻击者数量不超过客户端总数三分之一的条件下，模型所能保持的容错能力。证明过程利用了几何中位数的固有鲁棒性，并通过对诚实客户端与恶意客户端的贡献进行区分，严谨地推导了聚合误差的边界。这一分析不仅验证了本文模型在对抗环境中的稳定性，也为在实际联邦学习系统中部署鲁棒性防御策略提供了理论支持。

## 4 实验仿真

本节根据DPBR-Adapt框架进行对比实验，对比的差分隐私框架分别使用DP-FedSGD（differentially private FedSGD）、DP-SIGNSGD和DP-BREM。

### 4.1 实验环境

整体实验在一块GeForce RTX 4090 GPU、一块Intel Core CPU上进行训练、评估，用Pytorch实现联邦学习框架，所有实验使用Python实现。

### 4.2 数据集

实验使用的数据集有MNIST、Fashion-MNIST、CIFAR-10、CIFAR-100。其中MNIST和Fashion-MNIST数据集是通过一个具备3个卷积层和2个完全连接层的卷积神经网络（convolutional neural network，CNN）进行处理，CIFAR-10和CIFAR-100数据集是通过ResNet-18进行处理。

### 4.3 基线

为了全面评估DPBR-Adapt的性能，本节选取了3种具有不同技术特征的基准模型进行对比实验，不同模型的DP和防御机制介绍如表3所示。

DP-FedSGD：差分隐私应用于联邦学习的经典基准方案。其核心机制是在每轮聚合前，对客户端梯度进行统一的范数裁剪并添加固定强度的高斯噪声。

DP-SIGNSGD：该方案仅传输梯度的符号信息，通过对符号位进行概率翻转来实现差分隐私保护。在防御层面，它利用多数投票机制聚合符号，能够天然抵御数值极端的异常攻击。

DP-BREM：该方案利用动量项平滑梯度的随机波动，并基于动量历史信息进行动态噪声注入和异常检测。

### 4.4 参数设置

DPBR-Adapt模型参数包括：设置100个客户端，其中20%为拜占庭攻击者所控制，服务器与客户端进行200次交互通信，学习率默认为0.1，隐私预算为1、2、3、4；数据分布采用Non-IID设置，通过狄利克雷分布划分数据以模拟异构性；每轮通信中，服务器随机选取比率的客户端参与聚合；客户端本地训练轮数$E=5$，批次大小

表3 不同模型的DP和防御机制介绍

| 模型 | 方法 | DP机制 | 防御机制 |
|---|---|---|---|
| DP-FedSGD | 标准差分隐私联邦梯度下降 | 固定高斯噪声 | 梯度裁剪 |
| DP-SIGNSGD | 基于梯度符号的压缩传输 | 位翻转机制 | 利用多数投票机制，天然抵御数值异常攻击。 |
| DP-BREM | 基于动量的鲁棒聚合 | 基于动量的动态噪声 | 利用动量历史信息辅助异常检测 |
| DPBR-Adapt | 层级自适应结合训练进度感知 | 多维自适应噪声 | 双重过滤：梯度范数动态筛选和余弦方向一致性筛选 |

$B = 64$。实验结果以5轮实验结果平均值±标准差的形式呈现。

### 4.5 实验评估

本节通过多轮对比实验，对DPBR-Adapt的安全防御能力、效用–隐私平衡能力、隐私保护能力进行全面评估，并测试了4种先进的攻击方法，即Min-Max[23]、ByzMean[24]、ALIE[25]和LF。

不同模型在拜占庭攻击下的最佳准确率比较如表4所示。实验使用DPBR-Adapt、DP-FedSGD、DP-SIGNSGD、DP-BREM这4个模型，在MNIST、Fashion-MNIST、CIFAR-10数据集中面对不同类型的拜占庭攻击进行准确率对比。本次实验采用相同的隐私预算（$\varepsilon$=2），这样可以更好地观察噪声对模型性能的影响。由表4可得如下实验结果。

1)在没有攻击（No-Attack）的情况下，DPBR-Adapt模型在各个数据集的训练中达到了与DP-FedSGD近似甚至更高的精度，这表明双重过滤机制和自适应权重调整机制在拜占庭鲁棒设计中并未影响到模型的精度，甚至对模型进行了一定的调优，特别是面对复杂场景，DPBR-Adapt表现出优越的性能和适应性，远优于其他模型。

2)随着拜占庭攻击者的加入，在不同的场景下，DPBR-Adapt模型精度下降较小，表现更好，表明在不同攻击下，它仍然提供了很好的鲁棒性，优于其他模型。其他模型都出现了不同程度的下降，表明大量聚合的噪声使其模型鲁棒性受到影响，更容易遭受到拜占庭攻击。

3) 在面对拜占庭攻击时，DPBR-Adapt与其他模型相比，模型准确率受到拜占庭攻击的影响更小，这表明在$\varepsilon$=2时，DPBR-Adapt具备很好的抵抗效果。通过攻击实例证明，相对于其他模型而言，本模型的隐私保护效果更佳。

4) 尽管ALIE和LF攻击利用真实数据分布的伪装性对DPBR-Adapt的双重过滤机制构成了挑战，导致模型在CIFAR-10等复杂数据集上的稳定性出现小幅下降；但实验数据证明，相较于基准算法，DPBR-Adapt对这类语义级攻击仍具备较强的抵抗效能，有效抑制了恶意信息的扩散，确保了全局模型的收敛性能与分类精度，体现了优异的鲁棒性。

5)在相同的隐私预算下，DPBR-Adapt模型随着训练的进度对不同层级采用了不同强度的隐私保护，有效地保证了模型的精度。在各类场景下，其他模型精度显著下降，甚至部分模型几乎丧失鲁棒性，而DPBR-Adapt模型仍能保持较高的准确率。

不同隐私预算下DPBR-Adapt的训练准确率如表5所示。为了充分展示噪声对整个系统模型的影响，这一实验将不再添加拜占庭攻击者。对于4个数据集，DPBR-Adapt采用4种不同的隐私预算（$\varepsilon$=1、2、3、4）。

**表4　不同模型在拜占庭攻击下的最佳准确率比较**

| 数据集 | 模型 | No-Attack | Min-Max | ByzMean | ALIE | LF |
|---|---|---|---|---|---|---|
| MNIST | DP-FedSGD | 93.12%±0.42% | 66.32%±1.94% | 27.56%±1.87% | 83.35%±0.63% | 74.91%±0.72% |
| | DP-SIGNSGD | 92.19%±0.37% | 56.86%±0.74% | 65.38%±0.82% | 80.47%±0.81% | 81.76%±0.56% |
| | DP-BREM | **93.98%**±0.26% | 70.41%±0.47% | 82.17%±0.32% | **89.72%**±0.27% | 87.03%±0.32% |
| | DPBR-Adapt | 90.43%±0.39% | **89.15%**±0.41% | **87.15%**±0.39% | 86.31%±0.48% | **87.36%**±0.43% |
| CIFAR10 | DP-FedSGD | 43.77%±2.34% | 21.41%±4.01% | 20.31%±3.72% | 34.01%±3.12% | 20.18%±4.09% |
| | DP-SIGNSGD | 40.11%±2.27% | 34.42%±3.07% | 31.30%±3.01% | 40.16%±3.01% | 37.22%±3.41% |
| | DP-BREM | 57.34%±1.39% | 52.58%±1.72% | 53.89%±1.44% | 42.73%±1.78% | 40.51%±1.42% |
| | DPBR-Adapt | **61.41%**±1.07% | **55.04%**±1.40% | **60.73%**±1.72% | **51.63%**±1.79% | **42.37%**±2.49% |
| FMNIST | DP-FedSGD | 70.13%±1.23% | 45.62%±2.76% | 50.22%±1.62% | 58.32%±1.63% | 41.74%±1.67% |
| | DP-SIGNSGD | **85.14%**±0.77% | 40.15%±1.75% | 56.34%±1.76% | 62.14%±1.64% | 52.17%±1.79% |
| | DP-BREM | 69.41%±0.86% | 49.13%±1.72% | 60.29%±1.37% | 60.17%±0.92% | 59.83%±1.32% |
| | DPBR-Adapt | 80.74%±0.53% | **75.22%**±1.02% | **79.82%**±1.28% | **72.16%**±0.57% | **64.29%**±0.72% |

**表5　不同隐私预算下DPBR-Adapt的训练准确率**

| 隐私预算 | MNIST | FMNIST | CIFAR10 | CIFAR100 |
|---|---|---|---|---|
| $\varepsilon$=4 | 92.89% | 81.55% | 64.51% | 58.67% |
| $\varepsilon$=3 | 91.11% | 81.07% | 63.99% | 57.46% |
| $\varepsilon$=2 | 90.43% | 80.74% | 61.41% | 55.91% |
| $\varepsilon$=1 | 89.68% | 80.12% | 59.26% | 54.28% |

当隐私预算为1、2、3、4时，DPBR-Adapt在MNIST数据集的准确率分别达到了89.68%、90.43%、91.11%、92.89%；在Fashion-MNIST数据集的准确率分别达到80.12%、80.74%、81.07%、81.55%；在较为复杂的CIFAR-10数据集的准确率分别达到59.26%、61.41%、63.99%、64.51%；在更为复杂的CIFAR-100数据集的准确率分别达到54.28%、55.91%、57.46%、58.67%。

随着隐私预算的增大，噪声的强度在逐步减小，对模型的性能影响也在减弱，这也符合差分隐私的基本性质，即预算越大、强度越小。在本次对比实验中，随着隐私预算的减少，差分隐私保护的强度也越大。由于采用了层级化噪声自适应，可以将噪声更平衡地分给各层级，特别是关键层（如卷积层），将从空间和时间上获得更严格的保护。在空间上，卷积层获得更高权重的噪声倾斜，这样有利于在训练过程中避免重要信息泄露，但同时也会影响整体计算效率。因此，采用在卷积层首层权重加大、中间层权重减小的策略，最大程度地减少对通信和计算效率的影响。在时间上，随着训练因子的不断增大，从初期的对关键层噪声倾斜，到训练后期的对关键层减少噪声、增大预算，避免影响模型的正常收敛和准确度。通过变异系数量化敏感度、动态训练因子调控、层级重要性分级三重机制，最终实现自适应噪声隐私保护。

$\varepsilon$=2时各模型在不同数据集上训练轮数与准确率的变化如图3所示。综合分析图3的结果可以发现，本文模型可以在$\varepsilon$=2的情况下完全收敛。值得注意的是，本文模型在保持相同隐私保护水平的同时，在准确性方面基本优于DP-FedSGD、DP-SIGNSGD、DP-BREM这3个基线模型，保证了模型的收敛性和有效性。此外，通过对训练轮数与准确率的变化散点图进行分析，能够更清晰地得到4种不同安全分布式学习模型的差异，以及本文模型在计算和训练效率上的提升。通过多轮实验可以看出，在MNIST数据集上，本文模型在达到80%准确率上的效率较其他模型提升了47%；在Fashion-MNIST数据集上，本文模型在达到65%准确率上的效率较其他模型提升了57%；在CIFAR10数据集上，本文模型在达到39%准确率上的效率较其他模型提升了25%。

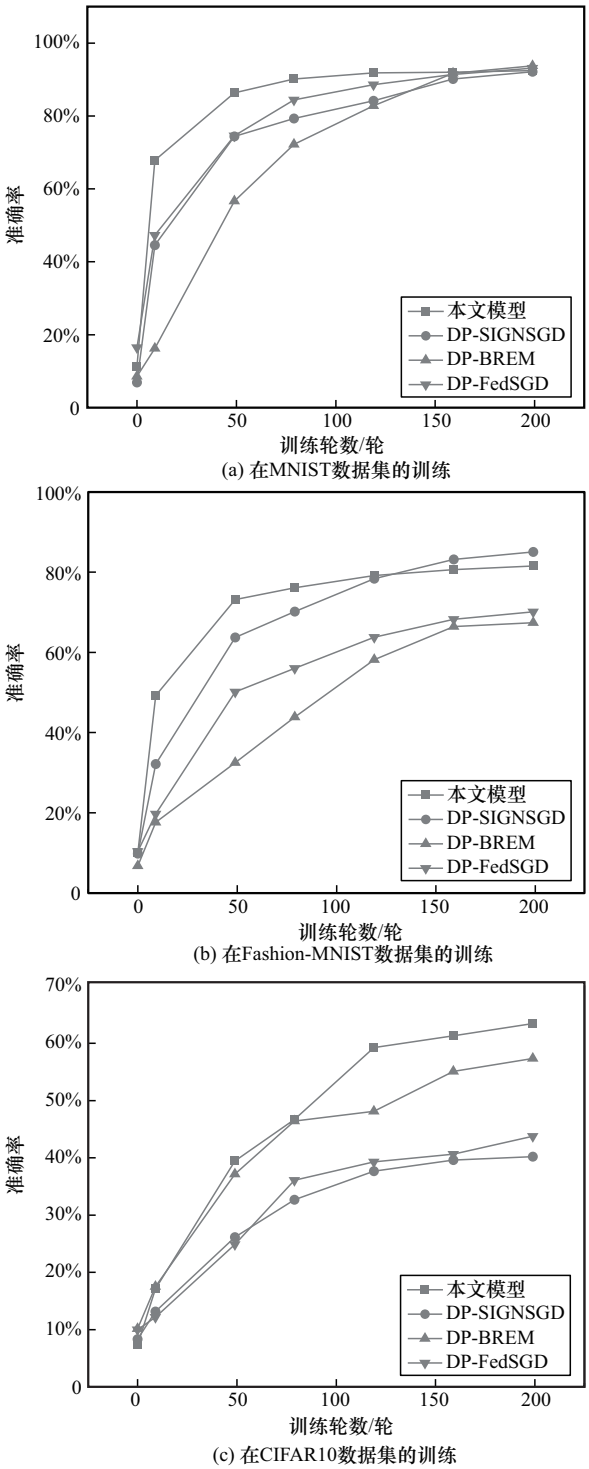



图3　$\varepsilon$=2时各模型在不同数据集上训练轮数与准确率的变化

4 个模型在不同数据集上每轮训练时间如表 6 所示。

**表 6　4 个模型在不同数据集上每轮训练时间**

| 模型 | MNIST/s | FMNIST/s | CIFAR10/s |
|---|---|---|---|
| DP-FedSGD | 3.81 | 4.77 | 56.14 |
| DP-SIGNSGD | 3.37 | 4.16 | 57.39 |
| DP-BREM | 2.84 | 3.07 | 40.16 |
| 本文模型 | 2.79 | 2.46 | 32.17 |
| 本文改进 | 0.05~1.02 | 0.61~2.31 | 7.99~25.22 |

每个模型使用相同的隐私预算（$\varepsilon$=2）进行训练，对 DPBR-Adapt 模型的计算效率进行评估。在相同的实验条件下，DPBR-Adapt 模型在 MNIST、Fashion-MNIST、CIFAR10 数据集上的计算效率均优于其他模型。相比于其他模型，DPBR-Adapt 模型在 MNIST 数据集上每轮训练时间最多缩短了 1.02 s，在 Fashion-MNIST 数据集上每轮训练时间最多缩短了 2.31 s，在 CIFAR10 数据集上每轮训练时间最多缩短了 25.22 s。

4 个模型在不同数据集的隐私预算如表 7 所示。

**表 7　4 个模型在不同数据集的隐私预算**

| 模型 | MNIST | FMNIST | CIFAR10 |
|---|---|---|---|
| DP-FedSGD | 3.28 | 5.86 | 5.72 |
| DP-SIGNSGD | 2.23 | 4.11 | 5.48 |
| DP-BREM | 2.11 | 4.07 | 4.23 |
| 本文模型 | 2.00 | 4.00 | 4.00 |
| 本文改进 | 0.11~1.28 | 0.07~1.86 | 0.23~1.72 |

由表 7 可知，DPBR-Adapt 模型在 MNIST、Fashion-MNIST、CIFAR10 数据集上保持了较高准确率的同时，实现了最低的隐私预算值：在 MNIST 数据集上最多降低了 1.28 的隐私预算，在 Fashion-MNIST 数据集上最多降低了 1.86 的隐私预算，在 CIFAR10 数据集上最多降低了 1.72 的隐私预算。隐私预算的下降恰恰证明了 DPBR-Adapt 模型在隐私保护和准确率之间实现了更高效的隐私效用平衡。

实验结果表明，DPBR-Adapt 模型在不同的隐私预算下达到了较高的准确率，在复杂的场景和严格的隐私保护下，DPBR-Adapt 模型在精度、计算效率、隐私–效用平衡方面优于其他的联邦学习模型。当面对拜占庭攻击时，DPBR-Adapt 在所有参与的数据集中均实现了高准确率，优于其他模型。这也证明了将有效的隐私保护机制与安全聚合方法相结合，是联邦学习隐私保护的一个重要方向。

### 4.6　消融实验

虽然实验结果证实了 DPBR-Adapt 的有效性，但检查其各组成部分的贡献是必要的。本节以 CIFAR-10 数据集为重点，评估各组成部分在不同攻击场景下对测试准确性的影响，并探索各超参数的设置对实验的整体影响。不同$\gamma$下模型准确率及收敛速度如图 4 所示。

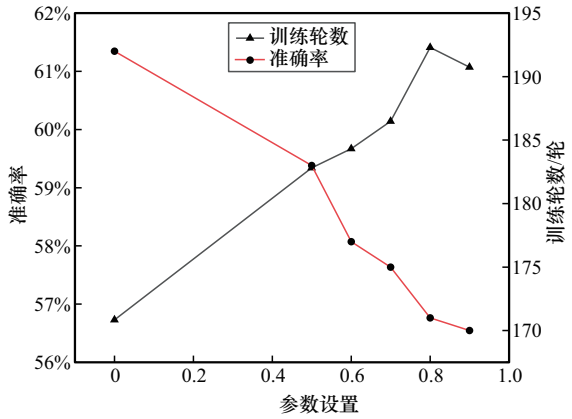



图 4　不同$\gamma$下模型准确率及收敛速度

从图 4 中可以看到，当 $\gamma$在 0.7~0.9 时，模型均能保持较高的收敛速度和精度。虽然较小的 $\gamma$会导致后期噪声衰减过慢，但算法整体并未出现发散。当$\gamma$为 0 时，选择了静态、固定地注入相同噪声，固定噪声强度无法感知梯度的实时演化规律。在训练后期，随着模型梯度范数的减小，恒定规模的噪声会产生较高的信噪比干扰，掩盖了真实的参数更新方向，从而导致模型准确率出现较大的下降，模型的收敛性能减弱。值得注意的是，虽然在$\gamma$为 0.8 之后收敛速度依旧在提高，但为了保证模型处于最佳精度，在实验中仍然选择$\gamma=0.8$作为实际数值。

为了验证层级权重，固定其他相关参数，分别对传统均匀权重（各层分配数值相同）与 DPBR-Adapt 采用的层级差异化权重策略进行实验。考虑到实验所采用的神经网络结构中，卷积层构成了特征表示的主要参数空间，其梯度范数的波动直接决定了全局模型的收敛方向。本文将敏感性分析的焦点集中于卷积层权重，以更清晰地刻画超参数与模型鲁棒性之间的演变关系。初始层级变化对模型准确率的影响如图 5 所示。

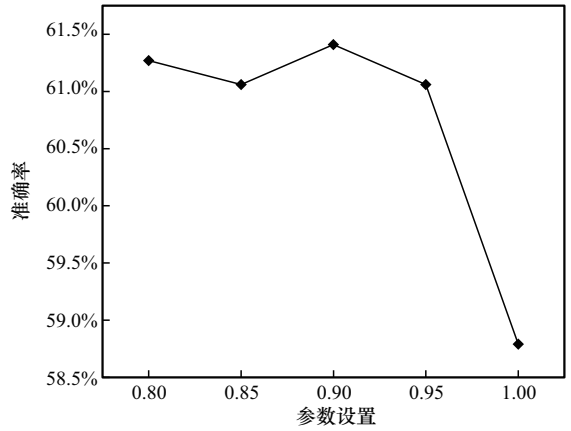



图5　初始层级变化对模型准确率的影响

从图5中可以看出，虽然权重数值的微调对结果影响微乎其微，但若取消层级差异（层级数都改为1），选择使用均匀分配，模型在同等隐私预算下的精度下降了3%~4%。这不仅证明了超参数选取的合理性，也验证了本文“层级自适应”机制的有效性和必要性。

双重过滤机制是DPBR-Adapt模型应对安全攻击的重要手段，验证该机制的贡献，对于理解模型如何在保证隐私性的同时构建鲁棒性防线至关重要。本文通过移除双重过滤模块进行对比实验，并评估该模块在不同攻击场景下对准确率的影响，包括Min-Max、ByzMean和ALIE攻击。不同筛选策略下的模型准确率如表8所示。

**表8　　不同筛选策略下的模型准确率**

| 过滤机制 | | 拜占庭攻击 | | |
|---|---|---|---|---|
| 距离 | 方向 | Min-Max | ByzMean | ALIE |
| √ | × | 40.88% | 43.77% | 37.13% |
| × | √ | 37.92% | 42.76% | 40.97% |
| √ | √ | **55.04%** | **60.73%** | **51.63%** |

由表8可以看出，单一维度的筛选策略在面对高级对抗攻击时表现出明显的局限性。距离筛选在拦截ByzMean等数值攻击时较为有效，但极易被ALIE等具备范数伪装能力的攻击绕过；方向筛选虽能捕捉部分语义偏离，但是在离散无固定方向的噪声环境下容易造成误判，从而影响模型性能。本文通过DPBR-Adapt双重过滤机制整合距离统计特征与方向几何特征，通过多维度的约束来保障模型的稳健收敛，从而实现防御性能的提升，充分验证了两类筛选准则在异常检测中的互补性与必要性。

## 5　结束语

本文DPBR-Adapt是一种具有层级自适应的联邦学习防御方案。为了解决联邦学习中隐私泄露和安全威胁的问题，DPBR-Adapt提出了一种深层次的隐私感知的鲁棒聚合方案，将层级重要性作为关键指数同时指导隐私分配和鲁棒性权重，形成了隐私–鲁棒性内在统一的协助防护机制，实现了自适应噪声添加过程与拜占庭筛选的异常检测互补增强的效果。研究结果证明，DPBR-Adapt在相同的隐私预算下，优于DP-FedSGD、DP-SIGNSGD、DP-BREM这3种对比模型，提供了更好的隐私–效用平衡和更强的拜占庭鲁棒性，在各类复杂场景下保持了高水平的准确性和稳定性。

尽管该方案在半诚实服务器下表现优异，但在处理大规模深层网络及极度Non-IID数据时仍面临计算负荷与公平性挑战。未来研究将重点突破不可信服务器下的安全聚合难题，结合可信执行环境与零知识证明等技术，在放宽信任假设的同时，进一步优化层级评估指标，提升方案在极端数据分布下的普适性与运行效率。

## 参考文献：

[1] Mcmahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]// Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). New York: PMLR, 2017: 1273-1282.

[2] Chen Y, Esmaeilzadeh P. Generative AI in medical practice: in-depth exploration of privacy and security challenges[J]. Journal of Medical Internet Research, 2024, 26: e53008.

[3] Hu R, Guo Y X, Gong Y M. Energy-efficient distributed machine learning at wireless edge with device-to-device communication[C]//Proceedings of the ICC 2022 - IEEE International Conference on Communications. Piscataway: IEEE Press, 2022: 5208-5213.

[4] Zeng Y J. Towards large-scale spectrum sensing and data analysis[D]. Madison: University of Wisconsin-Madison, 2022.

[5] Wei H L, Zhang H, AI-Haddad K, et al. Ensuring secure platooning of constrained intelligent and connected vehicles against Byzantine attacks: a distributed MPC framework[J]. Engineering, 2024, 33: 35-46.

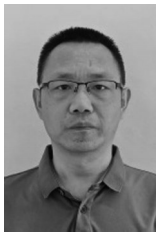
[6] 赵晓洁, 时金桥, 黄梅, 等. 联邦学习中的拜占庭攻防研究综述[J]. 通信学报, 2024, 45(12): 197-215.
Zhao X J, Shi J Q, Huang M, et al. Survey on Byzantine attacks and defenses in federated learning[J]. Journal on Communications, 2024, 45(12): 197-215.

[7] Yuan L Q, Wang Z R, Sun L C, et al. Decentralized federated learning: a survey and perspective[J]. IEEE Internet of Things Journal, 2024, 11(21): 34617-34638.

[8] Blanchard P, Mhamdi E M E, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]// Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS). Red Hook: Curran Associates, Inc., 2017: 119-129.

[9] Guerraoui R, Rouault S, et al. The hidden vulnerability of distributed learning in Byzantium[C]// Proceedings of the 35th International Conference on Machine Learning (ICML). New York: PMLR, 2018: 3521-3530.
[10] Dwork C, Roth A. The algorithmic foundations of differential privacy[J]. Foundations and Trends® in Theoretical Computer Science, 2014, 9(3/4): 211-487.
[11] Sun Z, Kairouz P, Suresh A T, et al. Can you really backdoor federated learning [C]// Proceedings of the 2nd Workshop on Federated Learning for Data Privacy and Confidentiality at NeurIPS. Vancouver: NeurIPS, 2019: 1-10.
[12] 康海燕, 冀源蕊. 基于本地化差分隐私的联邦学习方法研究[J]. 通信学报, 2022, 43(10): 94-105.
Kang H Y, Ji Y R. Research on federated learning approach based on local differential privacy[J]. Journal on Communications, 2022, 43(10): 94-105.
[13] Lyu L. DP-SIGNSGD: when efficiency meets privacy and robustness[C]// Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 3070-3074.
[14] Zhu H, Ling Q. Bridging differential privacy and Byzantine-robustness via model aggregation[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. Fremont: International Joint Conferences on Artificial Intelligence Organization, 2022: 2427-2433.
[15] Gu X L, Li M, Xiong L. DP-BREM: Differentially-private and Byzantine-robust federated learning with client momentum[C]// Proceedings of the 34th USENIX Security Symposium (USENIX Security 25). Berkeley: USENIX Association, 2025: 1-18.
[16] 周由胜, 高璟琨, 左祥建, 等. 基于自适应拜占庭防御的安全联邦学习方案[J]. 通信学报, 2024, 45(8): 166-179.
Zhou Y S, Gao J K, Zuo X J, et al. Secure federated learning scheme based on adaptive Byzantine defense[J]. Journal on Communications, 2024, 45(8): 166-179.
[17] Liu X, Liu Y, Liu J, et al. Adaptive differential privacy for deep learning based on layer-wise relevance propagation[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(1): 721-734.
[18] Dwork C, Roth A. The algorithmic foundations of differential privacy[M]. Boston: Now Publishers, 2014.
[19] Everitt B S, Skrondal A. The Cambridge Dictionary of Statistics[M]. Cambridge: Cambridge University Press, 2010.
[20] Luo M, Chen F, Hu D, et al. No fear of heterogeneity: classifier calibration for federated learning with non-IID data[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates, Inc., 2021: 5972-5984.
[21] Li T, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks[C]//Proceedings of Machine Learning and Systems. Austin: MLSys Press, 2020: 429-450.
[22] Chen Q, Wang H B, Wang Z L, et al. LLDP: a layer-wise local differential privacy in federated learning[C]//Proceedings of the 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Piscataway: IEEE Press, 2022: 631-637.
[23] Shejwalkar V, Houmansadr A. Manipulating the Byzantine: optimizing model poisoning attacks and defenses for federated learning[C]// Proceedings of the 28th Annual Network and Distributed System Security Symposium (NDSS). San Diego: Internet Society, 2021: 1-18.
[24] Xu J, Huang S L, Song L Q, et al. Byzantine-robust federated learning through collaborative malicious gradient filtering[C]//Proceedings of the 2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2022: 1223-1235.
[25] Baruch G, Baruch M, Goldberg Y. A little is enough: circumventing defenses for distributed learning[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates, Inc., 2019: 8635-8645.

[作者简介]

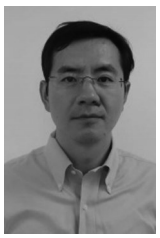


**胡荣磊**（1977– ），男，河北衡水人，博士，北京电子科技学院副研究员、硕士生导师，主要研究方向为隐私保护、联邦学习、区块链安全、物联网安全等。

**白晨阳**（2001– ），男，河南周口人，北京电子科技学院硕士生，主要研究方向为隐私保护、联邦学习等。

**魏占祯**（1971– ），男，青海西宁人，北京电子科技学院研究员级高级工程师，主要研究方向为网络安全。

**韩妍妍**（1982– ），女，黑龙江哈尔滨人，博士，北京电子科技学院副研究员、硕士生导师，主要研究方向为密码学中信息隐藏、秘密共享、可视密码等。



**段晓毅**（1979– ），男，贵州六盘水人，博士，北京电子科技学院副教授，主要研究方向为侧信道安全、人工智能应用与安全、无线网络安全。

**张浩**（1998– ），男，四川南充人，北京电子科技学院硕士生，主要研究方向为联邦学习、区块链等。