智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

自适应差分隐私的联邦学习方案

高媛, 石润华, 刘长杰

引用本文:
高媛, 石润华, 刘长杰. 自适应差分隐私的联邦学习方案[J]. 智能系统学报, 2024, 19(6): 1395-1406.
GAO Yuan, SHI Runhua, LIU Changjie. Federated learning scheme with adaptive differential privacy[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(6): 1395-1406.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202306052

DOI: 10.11992/tis.202306052
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20240909.1851.020

自适应差分隐私的联邦学习方案

高媛, 石润华, 刘长杰
(华北电力大学 控制与计算机工程学院, 北京 102206)

摘　要: 差分隐私被广泛应用于联邦学习中, 以保障模型参数的安全, 但不够合理的加噪方式会限制模型准确度进一步提高。为此, 提出一种能够自适应分配隐私预算和计算学习率的联邦学习方案(differential privacy-federated learning adaptive gradient descent, DP-FLAGD), 通过自适应分配隐私预算找到梯度的正确下降方向, 并计算合适的学习率以达到最小的损失。同时, DP-FLAGD 方案能够为不同隐私需求的用户提供不同的隐私预算, 以满足其需求。为评估 DP-FLAGD 的有效性, 在广泛使用的 2 个数据集 MNIST(modiffe national institute of standard and technology)和 CIFAR-10 上进行相关实验, 实验结果表明, DP-FLAGD 方案在保证模型参数安全的同时, 能够进一步提高模型的准确率。

关键词: 联邦学习; 差分隐私; 自适应; 梯度下降; 卷积神经网络; 学习率; 梯度; 隐私预算
中图分类号: TP181　　**文献标志码**: A　　**文章编号**: 1673−4785(2024)06−1395−12

中文引用格式: 高媛, 石润华, 刘长杰. 自适应差分隐私的联邦学习方案 [J]. 智能系统学报, 2024, 19(6): 1395–1406.
英文引用格式: GAO Yuan, SHI Runhua, LIU Changjie. Federated learning scheme with adaptive differential privacy[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1395–1406.

Federated learning scheme with adaptive differential privacy

GAO Yuan, SHI Runhua, LIU Changjie
(School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China)

Abstract: Differential privacy is widely used in federated learning to ensure the security of model parameters. However, inappropriate methods for adding noise can limit the further improvement of model accuracy. A federated learning method with adaptive allocation of the privacy budget and calculation of the learning rate (DP–FLAGD) is proposed to address this problem. Through the adaptive allocation of the privacy budget, the right descending direction of the gradient can be identified, and the appropriate learning rate can be calculated to achieve minimal loss. Simultaneously, DP–FLAGD provides different privacy budgets for users with various privacy requirements. Experiments were conducted on two widely used datasets, namely MNIST and CIFAR-10, to evaluate the validity of DP–FLAGD. Experimental results show that the DP–FLAGD scheme can further improve model accuracy while ensuring the safety of model parameters.

Keywords: federated learning; differential privacy; adaptive; gradient descent; convolutional neural network; learning rate; gradient; privacy budget

目前, 一些配备了传感、计算和通信功能的智能设备, 如可穿戴设备和车内传感装备等非常流行, 用户在使用这些设备时, 会产生大量的数据[1]。设备利用这些数据来分析用户的行为和偏好, 可以为用户提供智能和个性化的服务。在传统的集成式学习方法中, 需要将不同设备上的数据传到一个统一中心来进行学习训练, 这违反了数据隐私保护要求[2]。现如今, 数据的隐私问题受到越来越多的关注。欧盟在 2018 年 5 月实施了通用数据保护条例(general data protection regulation, GDPR)[3], 该条例要求只能在有合法目的并在严格限制条件下才能够收集个人数据, 并且数

收稿日期: 2023−06−30.　**网络出版日期**: 2024−09−10.
基金项目: 国家自然科学基金面上项目(61772001).
通信作者: 石润华. E-mail: rhshi@ncepu.edu.cn.

据的完全控制权属于数据所有者。随着对数据隐私保护要求的提升，限制了不同物联网设备之间的数据共享，为物联网的快速发展带来了挑战。

为解决此问题，2016 年谷歌 AI 团队提出联邦学习（federated learning, FL）算法框架，应用于移动互联网手机终端的隐私保护。FL 是一种分布式的机器学习技术，其思想是多个拥有本地数据用户在本地进行训练，仅通过交换训练得到的模型参数或中间结果进行联合训练，实现“数据不动模型动”的新范式。虽然 FL 避免了用户的数据直接暴露，但有研究表明，传输的模型梯度或参数信息仍会泄露用户的一些个人信息。例如，Phong 等 [4] 展示了如何利用一小部分梯度数据泄露有用的数据信息。Fredrikson 等 [5] 研发了一类新的模型反演攻击，可以通过多次访问机器学习模型来学习个人的敏感信息。此外，Nasr 等[6] 提出的白盒隶属度推理攻击，可以利用随机梯度下降（stochastic gradient descent, SGD）算法的隐私漏洞，在仅知道模型参数的情况下跟踪训练数据记录。为解决在数据传输过程中造成的隐私泄露，相关研究[7-12] 中将 FL 框架与差分隐私（differential privacy, DP）机制相结合，为数据提供安全保障。

DP 是由 Dwork 等[13] 在 2006 年首次提出，用户评估一个隐私保护机制（算法）所能提供的隐私保护程度，其核心思想是给用户数据集或模型更新添加噪声，由于其严格的理论保证和较小的计算开销，在 FL 中得到了广泛的应用。尽管现有的方案已经能够很好地保证数据安全，但这些方案在联邦学习中的运用往往是简单的，即在联邦学习的每一轮都添加相同的噪声尺度，然而这种方法是存在一定弊端的。首先，模型的准确率依赖于预先设置的迭代轮数（T），若 T 设置得太小，则模型还未收敛就结束训练；若 T 设置得太大，则会造成每轮分配的隐私预算过小，噪声过大，影响模型精度。其次，在模型训练刚开始的时候，模型的准确率未趋于收敛，此时即使添加较大噪声，也不会对模型准确率造成较大影响；但当训练结果趋于收敛时，则应添加较小噪声。所以以往方案在每轮迭代中添加相同噪声的方案是不够合理的。同时，已有方案通常无法顾及数据拥有者的个人隐私偏好。为此，本研究提出了一种自适应加噪和计算学习率的联邦学习算法（differential privacy-federated learning adaptive gradient descent, DP-FLAGD），在模型刚开始训练时分配较小的隐私预算，即添加较大的噪声，而在模型训练后期分配较大的隐私预算，即添加较小的噪声，同时自适应地计算相应的学习率，使损失降到最小，并为具有不同隐私偏好的用户提供不同的隐私预算。

本研究主要贡献如下：

1）提出了一种自适应分配隐私预算的联邦学习算法，即在每一轮迭代中，通过更精细地分配隐私预算来改进原有的 DP 与 FL 相结合的算法。

2）提出自适应修改学习率的方案，在每一步更新中寻找更合适的学习率进行更新，以提供更高的准确度。

3）为每个用户提供了不同的隐私预算值，由于每个用户的隐私偏好不同，为每个用户添加相同程度的噪声方法是不合理的，因此，本研究为具有不同隐私偏好的用户提供了不同的隐私预算。

4）本研究在真实数据集上进行了相关实验，证明本方案在模型准确性和损失等方面优于现有的算法。

1　相关工作

虽然 FL 允许用户将原始数据保存在本地，但已有研究表明[14]，攻击者能够从交换的梯度等参数中推测出用户隐私信息。基于 DP 的联邦学习是一种为解决 FL 中所存在的安全问题而广泛使用的联邦学习框架，该框架通过在用户上传的梯度中添加噪声来保护本地数据的隐私。

Shokri 等[15] 提出了一种分布式的训练方法，将噪声注入到参数的梯度中，以保护神经网络的隐私。然而，这种方法注入的噪声大小和隐私预算都会与训练的周期数和共享参数的数量成比例地累积，因此，该方案可能会消耗不必要的大部分隐私预算，因为训练的轮数和多方之间共享参数的数量往往是较大的[16]。

为了改进这一点，基于 Abadi 等[17] 的组成定理，提出了一种时刻统计方法，该方法跟踪隐私预算支出并执行适用的隐私政策。然而，这仍然依赖于训练轮数，因为该方法在每个训练步骤中都会在参数的梯度中加入噪声。在隐私预算较小的情况下，只能使用少量的训练周期来训练模型。在实践中，当训练周期的数量需要很大时，这可能会影响模型的效用。

Phan 等 [18] 探索了一种针对差分隐私深度神经网络的方法。这项工作提出了深度私有自动编码器（dPAs），其中通过扰动自动编码器[19] 中的交叉熵误差来实现 DP。但其算法是专门为自动编码器设计的，应用了特定的目标函数。

以上方法对所有参数注入的噪声量都是相同

的, Phan 等 [20] 提出一种应用于神经网络的自适应拉普拉斯机制(adaptive laplace mechanism, AdLM), 能够根据每个噪声对模型输出的贡献, 自适应地将噪声注入到特征中的方案, 在与模型输出不那么相关的特性中添加较多的噪声, 而在那些与模型输出更为相关的特性上添加较少的噪声。

随机梯度下降(stochastic gradient descent, SGD)算法是一种常用的优化方法, 是解决各种问题的常用工具, 如模型拟合[21]。出于安全考虑, 人们推出了 DP 版本的 SGD 算法, 然而, 在这些方案中, 通常是将总的隐私预算均分到每轮迭代中, 这种方法所得到的准确率往往不会太高, 现有的 FL 与 DP 结合的方案[10] 中多采用的也是这种 SGD 算法进行迭代优化。基于此, Lee 等[21] 提出通过更加合理地分配隐私预算的方案来提高模型的准确度。在训练刚开始的时候, 梯度值预计会比较大, 此时不需要精确地测量, 可以添加较少的隐私预算; 然而, 随着一轮又一轮的迭代, 参数逐渐接近最优值时, 梯度减小, 此时需要更加精准地测量, 因此需要更多的隐私预算添加到其中。

上述这些优化算法都是将自适应加噪运用到神经网络之中, 而本研究所关注的是如何在联邦学习中将其进行应用, 因此, 这里创造性地提出了将 Lee 等[21] 的方案运用到联邦学习中, 以提高训练模型的准确度。

2　预备知识

2.1　联邦学习

联邦学习是为解决数据孤岛问题而提出的方法, 将训练数据分布在移动设备上, 通过聚合本地计算的更新来学习共享模型, 这种分散的方法就叫作联邦学习。用户集体由一个中央服务器进行协调, 每次用户只需将模型更新上传到服务器, 而无需上传数据本身, 从而为数据的隐私安全提供一定的保障[22]。

其具体算法如算法 1 所示, 具体来说可以分为 4 个步骤[14]。

1) 初始化: 服务器初始化全局模型参数, 并选择客户端, 将初始化的参数分发给客户端。

2) 本地训练: 客户端使用自己的本地数据集以及接收到的全局模型参数进行训练并更新参数。

3) 本地上传: 客户端将参数更新上传至服务器。

4) 全局聚合及更新: 服务器将各个客户端的参数更新进行聚合, 并将更新的全局参数分发给客户端。

重复步骤 2)~4)直至满足要求为止。

算法 1　联邦平均算法(FedAvg)

输入　K 个客户端(用 k 进行索引), 局部小批量大小 B, 本地训练迭代轮数 E, 学习率 α。

输出　全局模型参数 w。

服务器端:

1) 服务器初始化模型参数 w_0

2) for each round $t = 1, 2, \cdots$ do

3) $m \leftarrow (\mathrm{CK}, 1)$

4) $S_t \leftarrow$随机选择的 m 个用户

5) for each client $k \in S_t$ in parallel do

6) $w_{t+1}^k \leftarrow \mathrm{ClientUpdate}(k, w_t)$

7) $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$

客户端:

8) $B \leftarrow$将客户按数量大小为 B, 分成小批量

9) for each local epoch i from 1 to E do

10) for batch $b \in B$ do

11) $w = w - \alpha \nabla l(w; b)$

12) return w to server

典型的横向联邦学习系统的架构示例如图 1 所示。

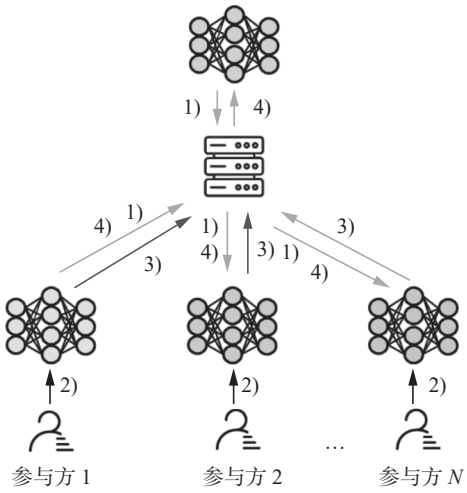


图 1　横向联邦学习系统示例

Fig. 1　Diagram of horizontal federated learning system

2.2　差分隐私

差分隐私解决了在了解到一个群体的有用信息后, 但不能获得该群体中某一人信息的悖论[13], 被广泛应用于解决联邦学习中的隐私问题。

定义 1($\varepsilon-\mathrm{DP}$)[23]　一个算法 A 满足 $\varepsilon-\mathrm{DP}$, 其中 $\varepsilon \geqslant 0$, 当且仅当对于任意 2 个只在 1 个元素上不同的数据集 D 和 D', 有

$$\forall T \subseteq \mathrm{Range}(A): \Pr[A(D) \in T] \leqslant \exp(\varepsilon)\Pr[A(D') \in T] \tag{1}$$

式中: $\mathrm{Range}(A)$ 表示算法 A 所有可能输出的集合, ε 为 (0, 1) 的任意一个数, Pr[] 表示概率。

定义 2（$(\varepsilon,\delta)-$DP） 一个算法A满足$(\varepsilon,\delta)-$DP，其中$\varepsilon\geqslant 0$，$\delta\geqslant 0$，当且仅当对于任意 2 个只在 1 个元素上不同的数据集D和D'，有

$$\Pr[A(D)\in T]\leqslant \exp(\varepsilon)\Pr[A(D')\in T]+\delta \tag{2}$$

定义 3（高斯机制） ε是 (0, 1) 的任意一个数；$\Delta_2(A)$是算法A的L_2灵敏度；N表示正态分布；σ为尺度参数，决定了分布的幅度；高斯机制返回的结果为$A(D)+N(0,\sigma^2)$，其中

$$\sigma\geqslant\frac{\Delta_2(A)}{\varepsilon}\sqrt{2\ln\frac{1.25}{\delta}} \tag{3}$$

满足$(\varepsilon,\delta)-$DP。

定义 4（$\rho-z$CDP） 对于一个输出$o\in\mathrm{Range}(A)$，算法A的隐私损失随机变量Z定义为

$$Z=\log\frac{\Pr[A(D)=o]}{\Pr[A(D')=o]} \tag{4}$$

$\rho-z$CDP 对Z的矩生成函数施加一个约束，并要求其集中在零附近。在形式上，需要满足

$$\exp(D_\alpha(M(D)\|M(D'))=\mathrm{E}[\exp(\alpha-1)\alpha\rho],\quad\forall\alpha\in(1,\infty) \tag{5}$$

式中：$\exp(D_\alpha(M(D)\|M(D'))$为$\alpha$-瑞丽散度，$D_\alpha$为$\alpha$散度，$M$为某种算法，$\alpha\in(-\infty,+\infty)$。

引理 1 假设 2 种机制分别满足ρ_1-zCDP和ρ_2-zCDP，则其组合满足$(\rho_1+\rho_2)-z$CDP，其中，ρ_1和ρ_2代表隐私参数。

引理 2 返回结果为$A(D)+N(0,\sigma^2)$的高斯机制满足$\Delta_2\left(A^{\frac{2}{2\sigma^2}}\right)-z$CDP。

引理 3 算法A满足$\rho-z$CDP，则A满足$(\varepsilon,\delta)-$DP对于任意的$\delta>0$，并且$\varepsilon=\rho+\sqrt{4\rho\log\left(\frac{1}{\delta}\right)}$。

3 自适应加噪联邦学习方法

3.1 算法描述

DP-FLAGD 算法的主要思想是保持预设总的隐私预算大小不变，在模型迭代的每一轮中，通过自适应计算隐私预算的方式，为梯度加入大小不同的噪声。

1）自适应加噪

在已有的方案中[24-25]，隐私预算ε被平均分配到每一轮本地迭代中，即$\varepsilon_1=\varepsilon_2=\cdots=\varepsilon_T=\frac{\varepsilon}{T}$，但这样的分配方式是存在一定问题的。首先，模型的准确率依赖于预先设定的迭代轮数T，如果T设置得太小，则还没有达到模型最优效果，迭代就已停止；而若T太大，则分配给每一轮的隐私预算ε_t的值就会太小，即添加的噪声太大，噪声太大也会影响到模型的准确度。因此，预算设定迭代轮数T的方案是不合理的。其次，在训练过程中为每一轮迭代平均分配隐私预算的方案也是不够合理的。在训练刚开始的时候，梯度参数往往不够精确，而当迭代进行几轮之后，参数趋于收敛，梯度下降的方向变得精确。因此，若每轮迭代都提供较小的隐私预算，则可能会影响到最后梯度参数的准确性；若每轮都提供较大的隐私预算，则可能会造成精度还未收敛，就因隐私预算耗尽而终止训练。无论是哪种情况，都会导致模型准确率较低。因此，本研究提出一种自适应分配隐私预算的方式，通过更合理地分配隐私预算来提高模型准确度。

在没有考虑隐私的设置中，随机优化方法使用从一小组随机选择的数据中计算近似梯度，而非精确梯度。因此，每个更新的梯度方向$-\tilde{g}$可能并不是一个下降方向，而是一个期望的下降方向，但在使用了 DP 的算法中，不能依赖于期望中的保证。当本研究提出的算法认定$-\tilde{g}_t$可能并不代表一个下降方向时，本方案将当前的隐私预算份额从ρ_t增加到更大的一个值ρ_{t+1}，即$\rho_{t+1}\leftarrow(1+\eta)\rho_t$，利用$\rho_{t+1}-\rho_t$再次计算噪声梯度，得到$-\tilde{g}_2$，利用 GRADAVG 算法将 2 次计算的噪声梯度合并，得到$\tilde{S}$。具体细节如算法 2 所示。重复此过程，直到找到一个下降方向。

$$\tilde{g}_2=\nabla f(w_t)+N\left(0,\frac{\nabla f^2}{2\rho_t}\right) \tag{6}$$

式中$\Delta_2(\nabla f^2)$是f的梯度的L_2灵敏度。

$$\tilde{g}_2=\nabla f(w_t)+N\left(0,\frac{\Delta_2(\nabla f^2)}{2(\rho_{t+1}-\rho_t)}\right)$$

$$\tilde{S}=\frac{\rho_t\tilde{g}_t+(\rho_{t+1}-\rho_t)\tilde{g}_2}{\rho_t+(\rho_{t+1}-\rho_t)} \tag{7}$$

通过计算可以得到：

$$\mathrm{E}\left[\widehat{S}_t\right]=\nabla f(w_t) \tag{8}$$

$$\mathrm{Var}\left(\hat{S}_t\right)=\frac{\rho_t^2\dfrac{\Delta_2(\nabla f)^2}{2\rho_t}+\dfrac{\Delta_2(\nabla f)^2}{2(\rho_{t+1}-\rho_t)}(\rho_{t+1}-\rho_t)^2}{\rho_{t+1}^2}=$$

$$\left(\frac{\Delta_2(\nabla f)^2\rho_t}{2}+\frac{\Delta_2(\nabla f)^2(\rho_{t+1}-\rho_t)}{2}\right)\Big/\rho_{t+1}^2=\frac{\Delta_2(\nabla f)^2}{2\rho_{t+1}} \tag{9}$$

算法 2 梯度平均算法（GRADAVG）

输入 $\rho_{\mathrm{old}},\rho_{\mathrm{ng}},g,\tilde{g},C_{\mathrm{grad}}$。

输出 全局模型参数$\tilde{S}$。

1) $\tilde{g}_2\leftarrow g+N\left(0,\left(\dfrac{C_{\mathrm{grad}}^2}{2(\rho_{\mathrm{ng}}-\rho_{\mathrm{old}})}\right)I\right)$

2) $\tilde{S}\leftarrow\dfrac{\rho_{\mathrm{old}}\tilde{g}+(\rho_{\mathrm{ng}}-\rho_{\mathrm{old}})\tilde{g}_2}{\rho_{\mathrm{ng}}}$

3) return $\tilde{S}$

4) end

2)自适应计算学习率

为了获得最小的损失, 本方案构建了一个集合$\Omega=\{f(w_t-\alpha\tilde{g}_t):\alpha\in\phi\}$, 其中$\phi$是预定义的学习率集。然后将集合$\phi$带入到 Noisy_Min[21] 算法中来选择使得函数$f(w_t-\alpha\tilde{g}_t)$值最小即损失值最小的α。在本研究的设置中, ϕ的第一个元素为 0, 即此时的学习率设置为 0。设i是由算法 Noisy_Min 返回的索引值, 当$i>0$时, 说明已找到使损失最小的学习率, 此时算法使用所选择的学习率α_i更新w_t; 而当$i=0$时, 说明$-\tilde{g}_t$可能并不代表一个下降方向, 因此将α设置为 0, 即不更新。此时需要重新分配较大的隐私预算, 并计算得到$\tilde{g}_2$, 再调用算法 2 得到$\tilde{S}$, 重复此过程, 至$i>0$, 即可以进行更新为止。自适应计算学习率的具体细节如算法 3 的第 9)~20) 所示。

3)为隐私需求不同的用户提供相应的隐私预算

由于每个用户的隐私偏好有所不同, 所以本研究采取了给隐私要求较高的用户提供了较小的隐私预算, 而给隐私保护要求较低的用户提供了较大的隐私预算的方式, 以满足不同用户的要求, 从而提高用户参与 FL 的积极性[26]。

DP-FLAGD 算法的具体细节如算法 4 所示。首先, 为不同隐私偏好的用户设置不同的隐私预算集合$\{\varepsilon\}_{k=0}^{K}$, 同时初始化模型参数$w_0$和学习率集$\phi$, 并将其广播给选中的 m 个参与者。其次, 参与者本地使用具有差分隐私保护的自适应梯度下降算法(DP-AGD)进行更新。用户根据ε_i计算相应的ρ, 当$\rho>0$时, 即隐私预算没有耗尽时, 将索引i初始化为 0, 计算并裁剪梯度获得g_t, 并为梯度参数添加噪声$\tilde{g}_t\leftarrow g_t+N(0,(C^2/2\rho_{\text{ng}})I)$, 得到$\tilde{g}_t$, 然后减掉相应的隐私预算。通过 Noisy_Min 算法获取使得损失最小的索引i的值, 若计算出的索引i的值为 0, 则说明$-\tilde{g}_t$可能并不代表一个下降方向, 此时重新分配隐私预算并得到新的$\tilde{g}_t$, 其具体细节如算法 4 的第 15)~18) 所示。如算法 4 的第 12)~13) 所示, 若计算得到的$i>0$, 则说明此时的$-\tilde{g}_t$代表下降方向, 此时进行更新, 即$w_{t+1}\leftarrow w_t-\alpha_i\tilde{g}_t$, 并将更新结果上传至服务器。最后, 服务器聚合来自不同客户端的参数更新, 并对其进行聚合, 将最终的更新结果发送给客户, 重复此过程, 直至隐私预算耗尽。

算法 3　具有差分隐私保护的自适应梯度下降算法(differential privacy-g adaptive gradient descent , DP-AGD)

输入　隐私预算ρ_{ng}、ε、δ_{tot}, 模型参数w, 步长集合ϕ, 隐私预算增长率η, 裁剪阈值 C, 用户数据$\{d_1,d_2,\cdots,d_n\}$,目标函数$f(w)=\sum_{i=1}^{n}l(w;d_i)$。

输出　模型参数 w。

1) $t\leftarrow 0$

2) $\rho\leftarrow$根据ε计算

3) while $\rho>0$ do

4) $i\leftarrow 0$

5) $g_t\leftarrow\sum_{i=1}^{n}\left(\nabla l(w_t;d_i)/\max\left(1,\frac{\|\nabla l(w_t)\|}{C}\right)\right)$

6) $\tilde{g}_t\leftarrow g_t+N(0,(C^2/2\rho_{ng})I)$

7) $\rho\leftarrow\rho-\rho_{ng}$

8) $\tilde{g}_t\leftarrow\tilde{g}_t/\|\tilde{g}_t\|_2$

9) while $i=0$ do

10) $\Omega=\{f(w_t-\alpha\tilde{g}_t):\alpha\in\phi\}$

11) $i\leftarrow\text{Noisy\_Min}(\phi)$

12) if $i>0$ then

13) if $\rho>0$ then $w_{t+1}\leftarrow w_t-\alpha_i\tilde{g}_t$

14) else

15) $\rho_{\text{old}}\leftarrow\rho_{\text{ng}}$

16) $\rho_{\text{ng}}\leftarrow(1+\eta)\rho_{\text{ng}}$

17) $\tilde{g}_t\leftarrow\text{GRADAVG}(\rho_{\text{old}},\rho_{\text{ng}},g_t,\tilde{g}_t,C)$

18) $\rho\leftarrow\rho-(\rho_{\text{ng}}-\rho_{\text{old}})$

19) end

20) end

21) $t\leftarrow t+1$

22) end

23) return w_t

算法 4　自适应加噪联邦学习算法(DP-FLAGD)

输入　隐私预算集合$\{\varepsilon\}_{k=0}^{K}$、$\delta$, 全局通信轮数 T, 联邦学习的参与者人数 K(k 为索引), 每轮的参与者数 m。

输出　模型参数 w。

服务器端:

1) 服务器初始化模型参数w_0和ϕ, 并广播给所有参与者

2) for $1\leqslant t\leqslant T$ do

3) $m\leftarrow(\text{CK},1)$

4) $S_t\leftarrow$随机选择的 m 个用户

5) for $i\in m$ do

6) $w_{t+1}^{k}\leftarrow\text{ClientUpdate}(\varepsilon_i/T,w_t)$

7) $w_{t+1}^{i}\leftarrow w_t+\sum_{i=1}^{N}\frac{n_k}{n}w_{t+1}^{k}$

客户端:

8) ClientUpdate(ε,w_t):

9) $\hat{w} \leftarrow w_t$

10) $w = \mathrm{DP-AGD}(\rho, w_t)$

11) $w_{t+1} = w - \hat{w}$

12) return w_{t+1}

3.2　隐私性分析

本方案使用 DP 对在 FL 中传输的数据进行隐私保护，通过为梯度参数添加噪声的方式来避免攻击者通过模型还原出原始训练数据。

本方案分别在 DP-AGD 算法的第 6) 步和 GRADAVG 算法的第 1) 步添加了高斯噪声，下面证明添加的噪声满足 $\rho-z$CDP，以在 DP-AGD 算法中添加的噪声为例。

定义 $f(y)=\log(\mathrm{P}[M(x)=y]/\mathrm{P}[M(x'=y)])$，其中假设 x、x' 相邻，设 $Y\sim M(x)$，$Z=f(Y)$，$\alpha\in(1,\infty)$。根据定义 4 有

$$\exp(D_\alpha(M(D)\|M(D'))=\mathrm{E}[\exp(\alpha-1)\alpha\rho] \tag{10}$$

所以

$$\mathrm{E}[\exp(\alpha-1)Z]=\mathop{\mathrm{E}}_{Y\sim M(X)}\left[\left(\frac{\mathrm{P}[M(x)=Y]}{\mathrm{P}[M(x'=Y)]}\right)^{\alpha-1}\right]=\exp((\alpha-1)D_\alpha(M(x)\|M(x')))\leqslant\exp((\alpha-1)(\rho\alpha)) \tag{11}$$

由马尔可夫不等式

$$\mathrm{P}[Z>\varepsilon]=\mathrm{P}[\exp((\alpha-1)Z)>\exp(\alpha-1)\varepsilon]\leqslant\frac{\mathrm{E}[\exp((\alpha-1)Z)]}{\exp(\alpha-1)\varepsilon}\leqslant\exp((\alpha-1)(\rho\alpha-\varepsilon)) \tag{12}$$

当 $\alpha=(\varepsilon+\rho)/2\rho>1$ 时，有

$$\mathrm{P}[Z>\varepsilon]\leqslant\exp(-(\varepsilon-\rho)^2/4\rho)\leqslant\delta \tag{13}$$

对于任何可测量的 $S\subset\gamma$，

$$\mathrm{P}[M(x)\in S]=\mathrm{P}[Y\in S]\leqslant\mathrm{P}[Y\in S\wedge Z\leqslant\varepsilon]+\mathrm{P}[Z>\varepsilon]\leqslant\mathrm{P}[Y\in S\wedge Z\leqslant\varepsilon]+\delta=\int_\gamma\mathrm{P}[M(x)=y]\cdot\mathrm{I}(y\in S)\cdot\mathrm{I}(f(y)\leqslant\varepsilon)\mathrm{d}y+\delta\leqslant\int_\gamma\exp(\varepsilon)\mathrm{P}[M(x')=y]\cdot\mathrm{I}(y\in S)\mathrm{d}y+\delta=\exp(\varepsilon)\mathrm{P}[M(x')\in S]+\delta \tag{14}$$

即 $\mathrm{P}[M(x)\in S]=\exp(\varepsilon)\mathrm{P}[M(x')\in S]+\delta$，$M$ 满足 $(\varepsilon,\delta)-\mathrm{DP}$。由于 $(\varepsilon,\delta)-\mathrm{DP}$ 在 $\delta>0$ 的情况下等效于 $z-\mathrm{CDP}$ [27]，所以该算法满足 $\rho-z$CDP。

与在 GRADAVG 算法中添加的噪声同理，由上述证明和引理 2 可以证明每次添加的噪声都是满足 $\rho-z$CDP 的。

隐私损失会在每次的迭代中累积，根据引理 1 可得到添加的噪声在迭代后满足 $(\rho_1+\rho_2+\cdots+\rho_n)-z$CDP 仍满足 $\rho-z$CDP。同时，本方案在 DP-AGD 算法的第 3)~13) 步中检查更新是否会导致隐私预算小于 0，若小于 0 即表示隐私预算已耗尽，则不再继续迭代。因此证明了本方案是严格满足 $\rho-z$CDP 的。

4　结果与分析

实验平台是 6 核 Inter GPU(2.90 GHz)、16 GB 内存、Win11 系统，代码采用 Python 实现。实验采用 MNIST（modiffe national institute of standard and technology）和 CIFAR-10 数据集。

MNIST 数据集由 7 万张手写体数字灰度图像组成，包括 6 万张训练图像和 1 万张测试图像。每张图像的大小是 28×28。标签是从 0~9 的数字，如图 2 所示。

图 2　MNIST 数据集示例

Fig. 2　Example figure of MNIST dataset

CIFAR-10 数据集由 10 类 RGB 彩色图像组成。图像大小为 32×32。这 10 个类别包括飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船和卡车。数据集共有 5 万个训练示例和 1 万个测试示例，如图 3 所示。

图 3　CIFAR-10 数据集示例

Fig. 3　Example figure of CIFAR-10 dataset

由于卷积神经网络（convolutional neural network, CNN）具有较好的特征提取能力，所以使用 CNN 作为神经网络体系结构。CNN 网络示意如图 4 所示。

将本方案 DP-FLAGD 与 FedAvg、LDP-

FL[27] 和 AdLM[28] 进行对比实验, 并记录结果。其中, AdLM 是自适应的 Laplace 机制模型, 对模型输出影响较大的特征加入更少的噪声, 而对模型输出影响较小的特征加入更多的噪声, 同时具有使噪声添加独立于训练过程的特性。

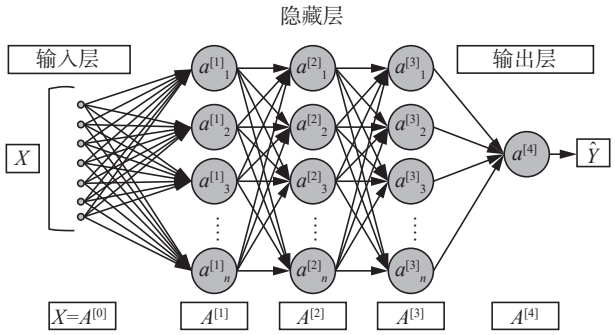


图 4　CNN 网络示意

Fig. 4　CNN network diagram

4.1　通信轮数对实验结果的影响

本组实验固定 2 个客户端, 本地训练两轮。分别记录了在独立同分布(IID)和非独立同分布(Non-IID)2 种情况下, DP-FLAGD、FedAvg、LDP-FL 和 AdLM 4 种算法在 MNIST 和 CIFAR-10 数据集上准确率随通信轮数变化的变化, 见表 1 和表 2。

表 1　4 种算法在 MNIST 数据集上的准确率比较

Table 1　Accuracy comparison of four algorithms on MNIST dataset　%

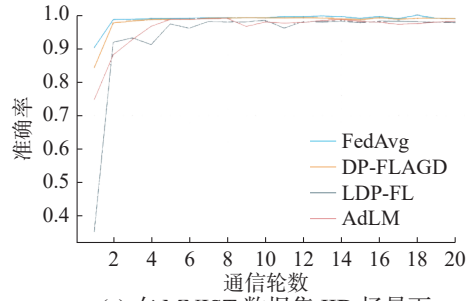
算法	准确率(IID)	准确率(Non-IID)
DP-FLAGD	99.02	95.88
FedAvg	99.11	98.30
LDP-FL	97.88	87.86
AdLM	**97.98**	**95.36**

表 2　4 种算法在 CIFAR-10 数据集上的准确率比较

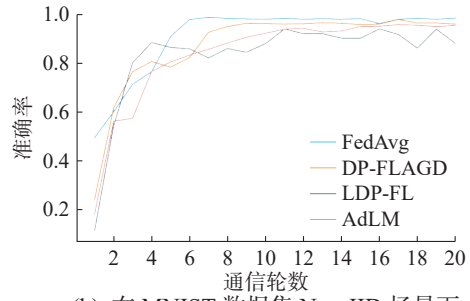
Table 2　Accuracy comparison of four algorithms on CIFAR-10 dataset　%

算法	准确率(IID)	准确率(Non-IID)
DP-FLAGD	53.86	34.04
FedAvg	57.32	41.71
LDP-FL	52.73	26.30
AdLM	**53.22**	**32.99**

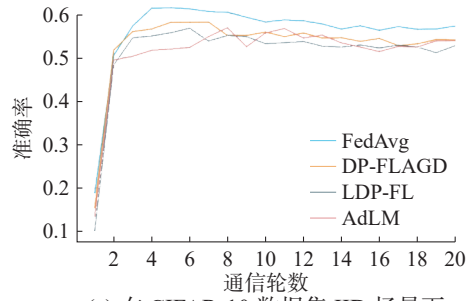
图 5(a) 给出了 4 种算法在 MINIST 数据集 IID 场景下准确率随通信轮数变化的趋势, 图 5(b) 给出了 4 种算法在 MNIST 数据集 Non-IID 场景下准确率随通信轮数变化的趋势。图 5(c) 给出了 4 种算法在 CIFAR-10 数据集 IID 场景下准确率随通信轮数变化的趋势, 图 5(d) 给出了 4 种算法在 CIFAR-10 数据集 Non-IID 场景下准确率随通信轮数变化的趋势。其中, FedAvg 算法前期随着通信轮数的增大, 准确率上升, 之后准确率几乎没有波动, 这是因为在该算法中, 没有噪声扰动梯度, 模型训练至后期逐渐趋于稳定。而 DP-FLAGD 算法、LDP-FL 算法以及 AdLM 算法因为有噪声的扰动, 从而准确率有所波动。最终 DP-FLAGD 的准确率接近 FedAvg, 证明了本研究提出的算法在保护隐私的同时, 也保证了训练的准确率。



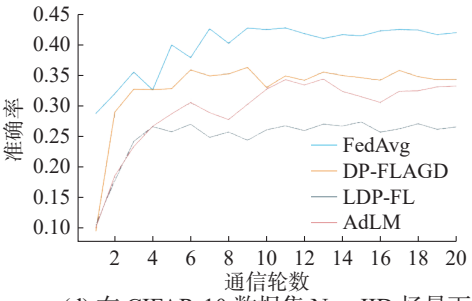
(a) 在 MNIST 数据集 IID 场景下



(b) 在 MNIST 数据集 Non-IID 场景下



(c) 在 CIFAR-10 数据集 IID 场景下



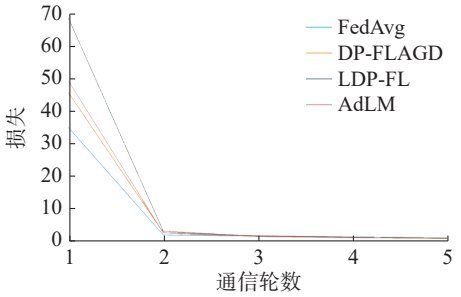
(d) 在 CIFAR-10 数据集 Non-IID 场景下

图 5　4 种算法在 2 个数据集上准确率随通信轮数的变化

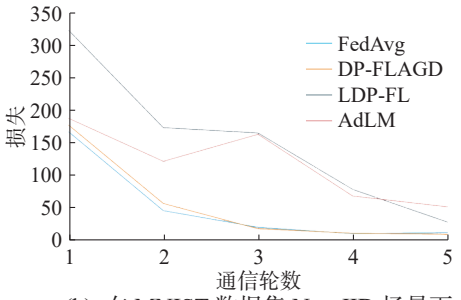
Fig. 5　Diagram of the accuracy variation of four algorithms with the number of communication rounds on two datasets

图 6(a) 给出了 4 种算法在 MINIST 数据集 IID 场景下损失随通信轮数变化的趋势, 图 6(b) 给出

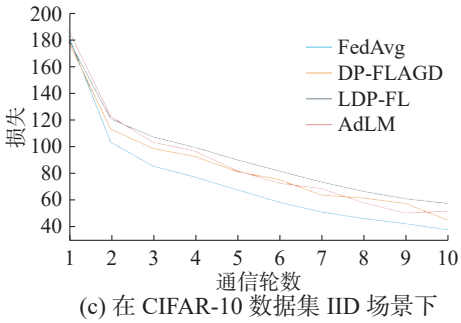
了 4 种算法在 MNIST 数据集 Non-IID 场景下损失随通信轮数变化的趋势。图 6(c) 给出了 4 种算法在 CIFAR-10 数据集 IID 场景下损失随通信轮数变化的趋势，图 6(d) 给出了 4 种算法在 CIFAR-10 数据集 Non-IID 场景下损失随通信轮数变化的趋势。由此看出，在 2 个数据集和 2 种场景下，DP-FLAGD 算法的表现都要优于 LDP-FL 算法和 AdLM 算法，同时损失与 FedAvg 算法较为接近，从而可证明本方案的有效性。



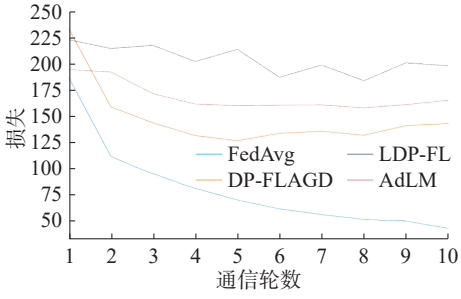
(a) 在MNIST 数据集 IID 场景下



(b) 在 MNIST 数据集 Non-IID 场景下



(c) 在 CIFAR-10 数据集 IID 场景下



(d) 在 CIFAR-10 数据集 Non-IID 场景下

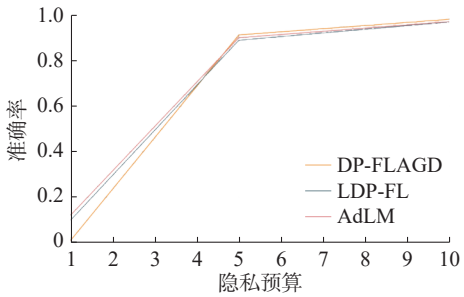
图 6　4 种算法在 2 个数据集上损失随通信轮数的变化

Fig. 6　Diagram of the loss variation of four algorithms with the number of communication rounds on two datasets

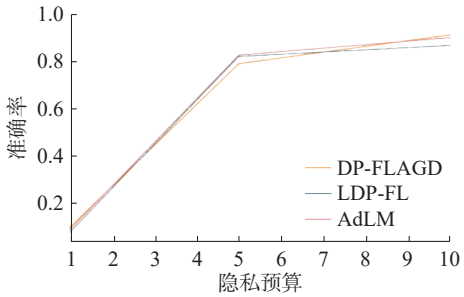
4.2　隐私预算对实验结果的影响

本组实验固定 MNIST 数据集全局训练 5 轮，CIFAR-10 数据集全局训练 10 轮。

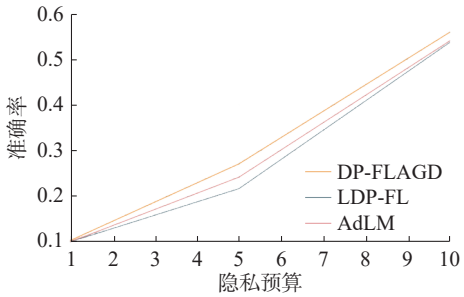
图 7(a) 给出了在不同隐私预算下，DP-FLAGD 算法和 LDP-FL 算法以及 AdLM 算法在 MNIST 数据集上的准确率，而图 7(b) 则给出了 Non-IID 的场景下的准确率，图 7(c) 和图 7(d) 分别是 2 种算法在 CIFAR-10 数据集上对应 IID 和 Non-IID 情境下的准确率。



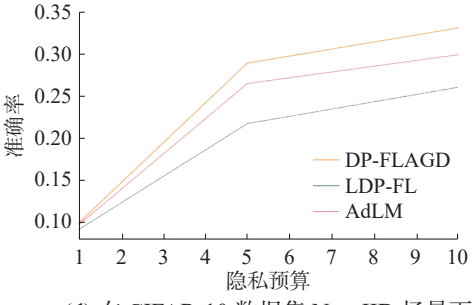
(a) 在MNIST 数据集 IID 场景下



(b) 在 MNIST 数据集 Non-IID 场景下



(c) 在 CIFAR-10 数据集 IID 场景下



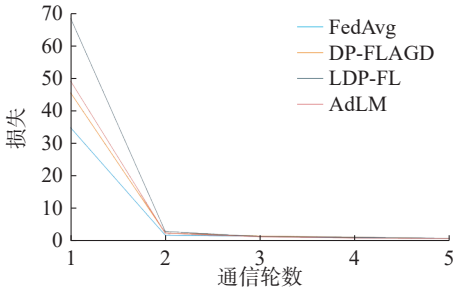
(d) 在 CIFAR-10 数据集 Non-IID 场景下

图 7　3 种算法在 2 个数据集上准确率随隐私预算的变化

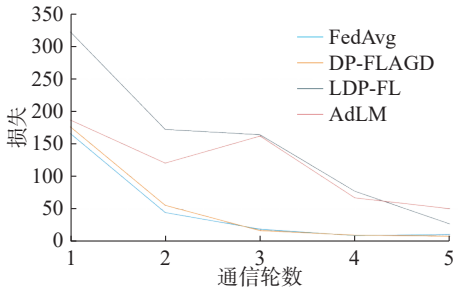
Fig. 7　Diagram of the accuracy variation of three algorithms with privacy budget on two datasets

由图 7 可以看出, 随着ε的增大, 准确率在增大, 这是因为ε越大, 添加的噪声就越小, 这符合 DP 机制。由于本研究提出的 DP-FLAGD 算法能够更加精细地分配隐私预算, 而 LDP-FL 和 AdLM 由于给模型加入的噪声过大, 造成准确率偏低, 所以可以看出此算法在 2 个数据集的 2 种场景下几乎都要优于现有的算法, 由此可证明本方案要优于现有的方案, 具有更高的准确性。

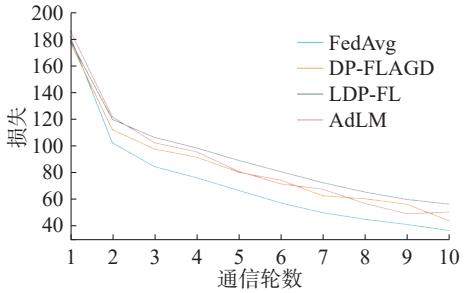
图 8~10 分别给出了在ε=10、5、1 的 3 种情况下, 4 种算法在 MNIST 和 CIFAR-10 数据集上对于 IID 和 Non-IID 情况下的损失值变化。



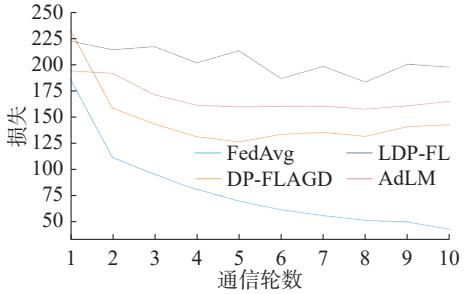
(a) 在 MNIST 数据集 IID 场景下



(b) 在 MNIST 数据集 Non-IID 场景下



(c) 在 CIFAR-10 数据集 IID 场景下

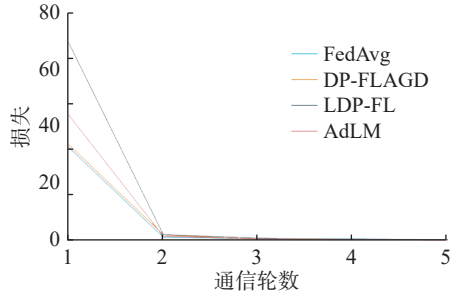


(d) 在 CIFAR-10 数据集 Non-IID 场景下

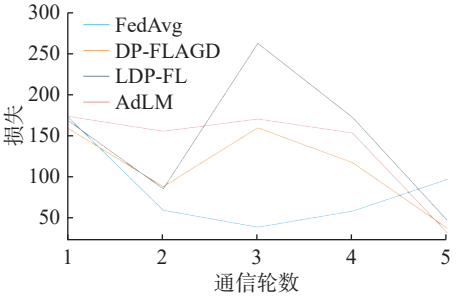
图 8　4 种算法在 2 个数据集上损失变化 (ε=10)

Fig. 8　Diagram of the loss variation of four algorithms on two datasets (ε=10)

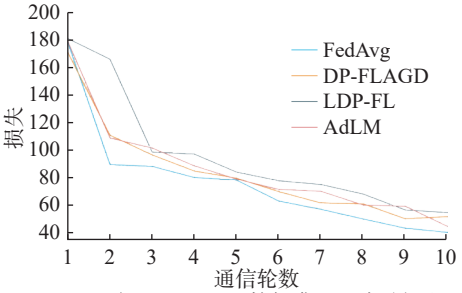
可以看出, 随着ε的增大, 损失也有所减小, 并且本方案的损失几乎始终小于 LDP-FL 算法和 AdLM 算法。



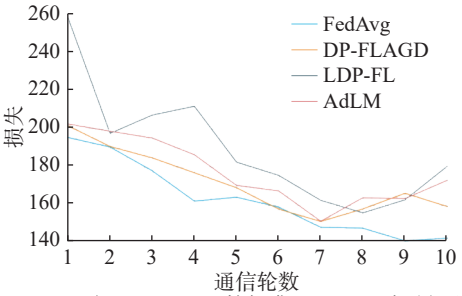
(a) 在 MNIST 数据集 IID 场景下



(b) 在 MNIST 数据集 Non-IID 场景下



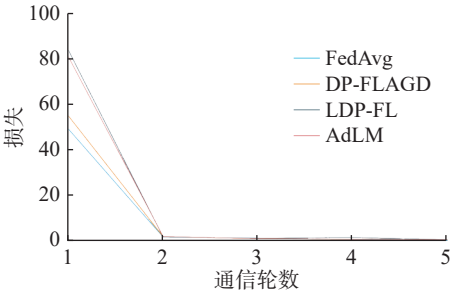
(c) 在 CIFAR-10 数据集 IID 场景下



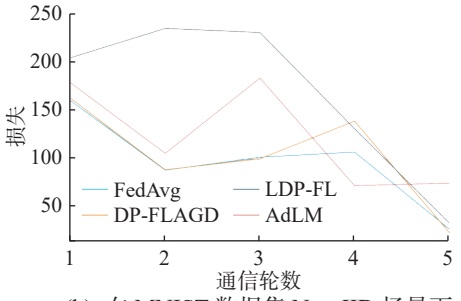
(d) 在 CIFAR-10 数据集 Non-IID 场景下

图 9　4 种算法在 2 个数据集上损失变化 (ε=5)

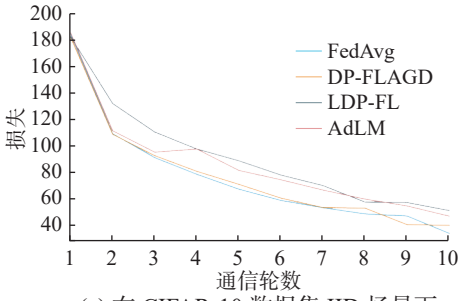
Fig. 9　Diagram of the loss variation of four algorithms on two datasets (ε=5)



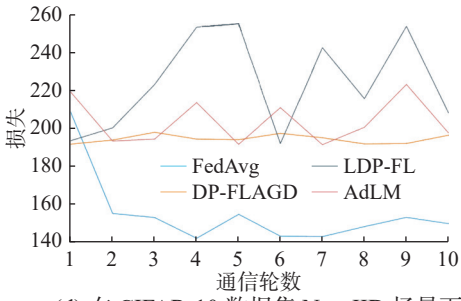
(a) 在 MNIST 数据集 IID 场景下



(b) 在 MNIST 数据集 Non-IID 场景下



(c) 在 CIFAR-10 数据集 IID 场景下



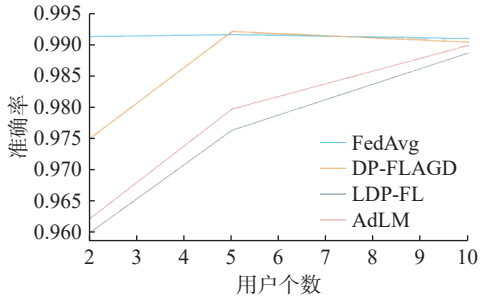
(d) 在 CIFAR-10 数据集 Non-IID 场景下

图 10　4 种算法在 2 个数据集上损失变化 (ε=1)

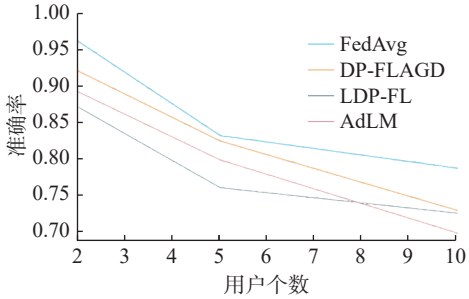
Fig. 10　Diagram of the loss variation of four algorithms on two datasets (ε=1)

4.3　参与用户的个数对实验结果的影响

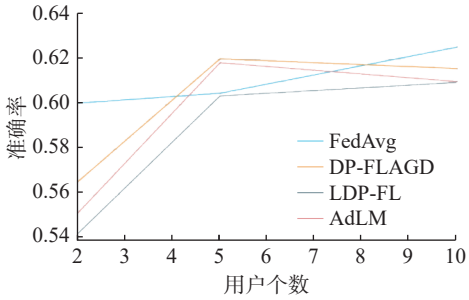
本组实验固定 MNIST 数据集全局训练 5 轮，CIFAR-10 数据集全局训练 10 轮。4 种算法在 2 个数据集上准确率随用户个数的变化如图 11 所示。由图 11 可以看出，在 IID 的场景下，准确率几乎是随着用户数量的增多而增加的，而在 Non-IID 场景下，随用户数量的增加，准确率却在下降，这是由于不同用户之间的数据是来自不同的数据分布模型所导致的。



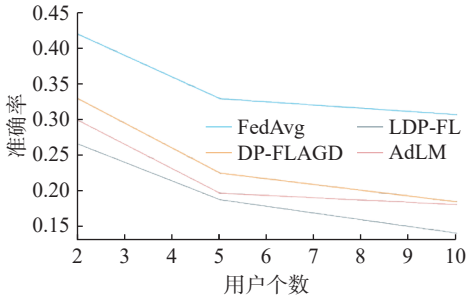
(a) 在 MNIST 数据集 IID 场景下



(b) 在 MNIST 数据集 Non-IID 场景下



(c) 在 CIFAR-10 数据集 IID 场景下



(d) 在 CIFAR-10 数据集 Non-IID 场景下

图 11　4 种算法在 2 个数据集上准确率随用户个数的变化

Fig. 11　Diagram of the accuracy variation of four algorithms with the number of clients on two datasets

总体来说，本研究提出的方法虽在 Non-IID 场景下准确率较 FedAvg 算法有一定的降低，但仍优于 LDP-FL 算法和 AdLM 算法。

5　结束语

在本研究中，为解决以往方案固定通信轮数和在每轮训练中平均分配隐私预算的不合理性，提出了一种自适应分配隐私预算和自适应计算学习率的 DP-FLAGD 方法。在训练刚开始的几轮添加较大噪声，而在梯度参数趋于精确时添加较小的噪声，当隐私预算耗尽时则停止训练。本研究所提出的方案有效解决了固定迭代轮数和均分隐私预算带来的一系列问题，并通过相应的实验证明了本方案的有效性。今后的工作将考虑如何提高算法在 Non-IID 场景下的准确性以及如何激励更多用户参与到 FL 之中。

参考文献:

[1] 徐树良, 王俊红. 结合无监督学习的数据流分类算法[J]. 模式识别与人工智能, 2016, 29(7): 665−672.
XU Shuliang, WANG Junhong. Classification algorithm combined with unsupervised learning for data stream[J]. Pattern recognition and artificial intelligence, 2016, 29(7): 665−672.

[2] ALWARAFY A, AL-THELAYA K A, ABDALLAH M, et al. A survey on security and privacy issues in edge-computing-assisted Internet of Things[J]. IEEE Internet of Things journal, 2021, 8(6): 4004−4022.

[3] VOIGT P, VON DEM BUSSCHE A. The EU general data protection regulation (GDPR)[M]. Cham: Springer International Publishing, 2017.

[4] PHONG L T, AONO Y, HAYASHI T, et al. Privacy-preserving deep learning via additively homomorphic encryption[J]. IEEE transactions on information forensics and security, 2018, 13(5): 1333−1345.

[5] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. Denver: ACM, 2015: 1322−1333.

[6] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning[C]//2019 IEEE Symposium on Security and Privacy. San Francisco: IEEE, 2019: 739−753.

[7] GEYER R C, KLEIN T, NABI M. Differentially private federated learning: a client level perspective[EB/OL]. (2017−12−20) [2021−01−01]. http://arxiv.org/abs/1712.07557.

[8] WEI Kang, LI Jun, DING Ming, et al. Federated learning with differential privacy: algorithms and performance analysis[J]. IEEE transactions on information forensics and security, 2020, 15: 3454−3469.

[9] LI Yiwei, CHANG T H, CHI C Y. Secure federated averaging algorithm with differential privacy[C]//2020 IEEE 30th International Workshop on Machine Learning for Signal Processing. Espoo: IEEE, 2020: 1−6.

[10] MCMAHAN H B, RAMAGE D, TALWAR K, et al. Learning differentially private recurrent language models [EB/OL]. (2017−10−18) [2021−01−01]. http://arxiv.org/abs/1710.06963.

[11] ZHAO Yang, ZHAO Jun, YANG Mengmeng, et al. Local differential privacy-based federated learning for Internet of Things[J]. IEEE internet of things journal, 2021, 8(11): 8836−8853.

[12] YANG Jia, FU Cai, LIU Xiaoyang, et al. Recommendations in smart devices using federated tensor learning[J]. IEEE internet of things journal, 2022, 9(11): 8425−8437.

[13] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and trends® in theoretical computer science, 2013, 9(3/4): 211−407.

[14] 梁文雅, 刘波, 林伟伟, 等. 联邦学习激励机制研究综述[J]. 计算机科学, 2022, 49(12): 46−52.
LIANG Wenya, LIU Bo, LIN Weiwei, et al. Survey of incentive mechanism for federated learning[J]. Computer science, 2022, 49(12): 46−52.

[15] SHOKRI R, SHMATIKOV V. Privacy-preserving deep learning[C]//2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton). Monticello: IEEE, 2015: 909−910.

[16] DWORK C, LEI Jing. Differential privacy and robust statistics[C]//Proceedings of the forty-first annual ACM symposium on Theory of computing. Bethesda: ACM, 2009: 371−380.

[17] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna: ACM, 2016: 308−318.

[18] PHAN N, WANG Yue, WU Xintao, et al. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction[J]. Proceedings of the AAAI conference on artificial intelligence, 2016, 30(1): 3175−3183.

[19] BENGIO Y. Learning deep architectures for AI[J]. Foundations and trends® in machine learning, 2009, 2(1): 1−127.

[20] PHAN N, WU Xintao, HU Han, et al. Adaptive Laplace mechanism: differential privacy preservation in deep learning[C]//2017 IEEE International Conference on Data Mining. New Orleans: IEEE, 2017: 385−394.

[21] LEE J, KIFER D. Concentrated differentially private gradient descent with adaptive per-iteration privacy budget[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018: 1656−1665.

[22] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from de-

centralized data[EB/OL]. (2016−02−17)[2021−01−01] http://arxiv.org/abs/1602.05629.

[23] 张兴, 陈昊. 差分隐私的高维数据发布研究综述[J]. 智能系统学报, 2021, 16(6): 989−998.
ZHANG Xing, CHEN Hao. A research review of high-dimensional data publishing based on a differential privacy model[J]. CAAI transactions on intelligent systems, 2021, 16(6): 989−998.

[24] ZHANG Jun, XIAO Xiaokui, YANG Yin, et al. PrivGene: differentially private model fitting using genetic algorithms[C]//Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2013: 665−676.

[25] TALWAR K, THAKURTA A, ZHANG Li. Nearly-optimal private LASSO[J]. Advances in neural information processing systems, 2015: 3025−3033.

[26] SUN Peng, CHE Haoxuan, WANG Zhibo, et al. Pain-FL: personalized privacy-preserving incentive for federated learning[J]. IEEE journal on selected areas in communications, 2021, 39(12): 3805−3820.

[27] BUN M, STEINKE T. Concentrated differential privacy: simplifications, extensions, and lower bounds[M]//Lecture Notes in Computer Science. Berlin: Springer Berlin Heidelberg, 2016: 635−658.

[28] MAHAWAGA ARACHCHIGE P C, BERTOK P, et al. Local differential privacy for deep learning[J]. IEEE internet of things journal, 2020, 7(7): 5827−5842.

作者简介：

高媛，硕士研究生，主要研究方向为差分隐私、联邦学习。E-mail: gaoyuanerer@163.com。

石润华，教授，博士生导师，博士，主要研究方向为经典\量子密码、量子计算、大数据与隐私保护。主持国家自然科学基金面上项目 2 项。发表学术论文 100 余篇。申请发明专利 40 项，其中已授权 30 余项。E-mail: rhshi@ncepu.edu.cn。

刘长杰，硕士，主要研究方向为联邦学习、入侵检测。E-mail: lcj@ncepu.cn。