

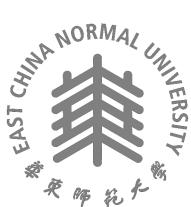
2022 届硕士专业学位研究生学位论文

分类号: _____

学校代码: 10269

密 级: _____

学 号: 51194501126



東華師範大學

East China Normal University

硕士专业学位论文

MASTER'S DISSERTATION (Professional)

**论文题目：基于联邦学习的隐私保护的技术
研究**

院 系: 软件工程学院

专业学位类别: 工程硕士

专业学位领域: 软件工程

论文指导教师: 曹珍富 教授

论文作者: 何慧娴

2021 年 09 月 10 日

Thesis (Professional) for Master's Degree in 2021

School Code: 10269

Student Number:51194501126

EAST CHINA NORMAL UNIVERSITY

TITLE: TECHNOLOGIES RESEARCH FOR PRIVACY PRESERVING BASED ON FEDERATED LEARNING

Department:	Software Engineering Institute
Major:	Software Engineering
Research Direction:	Privacy Preserving
Supervisor:	Professor ZhenFu Cao
Candidate:	HuiXian He

Nov 9, 2021

摘 要

随着人工智能的快速发展与移动设备的普及，需要多个参与方协作的应用场景不断涌现，分布式数据处理和分布式机器学习的作用日益凸显。比如分散在多个银行的金融数据、不同医院里的医疗记录、大平台下的每个用户的行为记录，以及智能电表、传感器或移动设备等产生的数据都需要分布式处理与挖掘。数据孤岛是分布式数据处理和分布式机器学习面临的重要挑战之一，作为解决数据孤岛的解决方案，联邦学习是一种很有前景的分布式计算框架，可以在多个分散的边缘设备上本地训练模型，而无需将其数据传输到服务器。随着公民隐私意识的提高和相关法律的完善，联邦学习中的隐私安全问题也日益受到人们的关注，且最新的研究工作表明已经能通过对模型的梯度参数进行攻击，还原用户的隐私数据，即仅通过保持数据的局部性来保护隐私是不够的，并且隐私保护技术在保护隐私的同时，还会牺牲模型精度。为此，本文使用差分隐私技术来保护联邦学习中用户的隐私，并针对分布式场景，分析模型训练过程中针对梯度下降算法的自适应干扰机制，实现提高模型精度的目的，并提出安全混洗模型，防止恶意服务器的攻击。本文主要工作包括如下几个方面：

本文主要的工作和贡献如下：

1. 在联邦学习差分隐私的场景下，本文提出了一种新型的、基于本地差分隐私的权重分配自适应干扰算法。在客户端本地训练的神经网络模型中，通过分析前向传播算法，计算每个属性类对于模型输出的贡献比，然后，我们设计了一个自适应噪声添加的方案，根据贡献率注入不同隐私预算的噪声。与传统的注入噪声的方法相比，我们在相同的隐私保护程度下最大限度地提高了

模型的准确性，减少噪声对模型输出结果的影响，提高模型精度。

2. 考虑到联邦学习中参数聚合器的攻击，本文提出了一种新的安全聚合机制，在本地客户端和中心服务器之间新增混洗器，在用户将参数上传到云服务器之前，先对参数进行混洗，模型参数的更新被匿名的发送到混洗器，通过对模型参数的拆分和混洗实现客户端匿名，并且证明了安全混洗模型的可行性。
3. 本文通过实验，展示了自适应本地差分隐私方案和安全混洗框架的结合，使得联邦学习的模型的精度和隐私预算达到平衡。

关键词： 联邦学习，隐私保护，本地差分隐私，安全混洗

ABSTRACT

With the rapid development of artificial intelligence and the proliferation of mobile devices, application scenarios that require the collaboration of multiple participants are emerging and the role of distributed data processing and distributed machine learning is becoming increasingly prominent. For example, financial data scattered across multiple banks, medical records in different hospitals, behavioural records of each user under a large platform, as well as data generated by smart meters, sensors or mobile devices all need to be processed and mined in a distributed manner.

Data silos are one of the key challenges facing distributed data processing and distributed machine learning. As a solution to address data silos, Federated Learning is a promising distributed computing framework that can train models locally on multiple decentralised edge devices without transferring their data to servers. With the increasing awareness of privacy among citizens and the improvement of related laws, privacy security in federation learning is also a growing concern, and recent research work has shown that it has been possible to restore users' private data by attacking the gradient parameters of the model, i.e. it is not enough to protect privacy by keeping the data local, and privacy-preserving techniques can protect privacy at the expense of model accuracy.

To this end, this paper uses differential privacy techniques to protect user privacy in federation learning, and for distributed scenarios, analyses the adaptive interference mechanism against the gradient descent algorithm during model training to achieve the goal of improving model accuracy, and proposes a secure shuffle framework to prevent

attacks by malicious servers.

The main work of this paper includes the following aspects:

1. In a federal learning differential privacy scenario, this paper presents a novel, local differential privacy-based adaptive interference algorithm for weight assignment. In a client-side locally trained neural network model, the contribution ratio of each attribute class to the model output is calculated by analysing the forward propagation algorithm, and then we develop an adaptive noise addition scheme that injects noise with different privacy budgets according to the contribution ratio. Compared with the traditional method of injecting noise, we maximise the accuracy of the model with the same degree of privacy protection, reduce the impact of noise on the model output results and improve the model accuracy.
2. Considering the attacks on parameter aggregators in federation learning, this paper proposes a new secure aggregation mechanism by adding a new mashup between the local client and the central server, where parameters are mashup before users upload them to the cloud server, and updates to model parameters are sent anonymously to the mashup, achieving client anonymity through splitting and mashup of model parameters, and demonstrating the secure The feasibility of mashup models is demonstrated.
3. In this paper, we experimentally demonstrate the combination of an adaptive local differential privacy scheme and a secure mashup framework that allows a federally learned model to balance accuracy and privacy budgets.

Keywords: *Federated learning, Privacy preserving, Local differential privacy , Security aggregation*

目录

第一章 绪 论	1
1.1 研究背景及意义	1
1.2 问题和挑战	4
1.2.1 数据异构	4
1.2.2 高昂的通信代价	5
1.2.3 安全性和隐私威胁	5
1.3 国内外研究现状	6
1.3.1 攻击模型的研究现状	6
1.3.2 隐私保护的研究现状	7
1.4 本文工作与主要贡献	9
1.5 本文组织结构	10
1.6 本章小结	10
第二章 基础知识	11
2.1 深度神经网络	11
2.2 联邦学习	14
2.2.1 基本介绍	14
2.2.2 模型框架	16
2.3 差分隐私	16
2.3.1 基本定义	17
2.3.2 相关概念	18
2.3.3 实现机制	19

2.4	联邦学习中的差分隐私	20
2.5	本章小结	22
第三章	联邦学习中的自适应本地差分机制	23
3.1	引言	23
3.2	自适应差分的 SGD 算法	25
3.2.1	层间依赖传播算法	26
3.2.2	自适应噪声添加	28
3.2.3	梯度范数裁剪	29
3.2.4	解析高斯机制	30
3.3	隐私性证明	32
3.4	隐私预算分析	34
3.5	本章总结	36
第四章	联邦学习的安全混洗模型	37
4.1	引言	37
4.2	安全混洗模型	38
4.2.1	客户端抽样	40
4.2.2	混洗器	40
4.3	隐私放大效应	42
4.4	隐私性证明	43
4.5	模型收敛性分析	45
4.6	本章总结	46
第五章	实验与评估	48
5.1	基准数据集介绍	48
5.2	实验环境与配置	49
5.3	实验设计	49
5.3.1	联邦学习模型	49
5.3.2	神经网络模型	50
5.4	自适应扰动方案的实验评估	51

5.5	安全混洗算法的实验评估	54
5.6	结果分析	57
5.7	本章小结	58
第六章	总结与展望	59
6.1	总结	59
6.2	展望	60
	参考文献	62
	致谢	69
	发表论文和科研情况	71

插图

1.1	联邦学习模型概况	3
2.1	深度神经网络结构图	12
2.2	前馈神经网络结构图	13
2.3	差分隐私的相邻数据集示意图	17
3.1	层间依赖传播算法	27
4.1	联邦学习中的安全模型框架	40
4.2	联邦学习安全模型中执行参数拆分混淆的混淆器	42
5.1	卷积神经网络结构图	51
5.2	梯度固定加噪方法下模型准确率随隐私预算变化情况	52
5.3	梯度自适应扰动方法下模型精度、损失随隐私参数的变化趋势	53
5.4	不同隐私预算的自适应干扰模型的准确率	54
5.5	DP-SGD、DLPP、ACDP 在模型准确率和隐私预算上的对比	54
5.6	安全混淆模型中参与混淆的本地客户端数量对联合模型精度的影响	55
5.7	安全混淆模型中通信轮数和客户端采样比对联合模型精度的影响	56
5.8	训练参数对模型精度的影响	56

List of Algorithms

1	基于自适应差分隐私的随机梯度下降算法	26
2	解析高斯算法	31
3	联邦学习中的安全模型算法: $\mathcal{A}_{\text{csdp}}$	39
4	混淆器中的拆分混淆算法	42

第一章 緒論

1.1 研究背景及意义

随着机器学习的不断发展和壮大，我们一方面惊叹于它的成就，比如 Alpha GO 击败了围棋世界冠军——柯洁、面部识别技术帮助我们抓住了躲藏多年的逃犯、大型工业企业也大力应用机器学习技术推动生产力的快速发展；另一方面，我们也认识到，机器学习还有巨大的发展潜力，例如：在医疗建设方面，构建基于大量病例的医疗救助诊断系统；在金融建设方面，运行基于大量商业行为数据的信用风险控制模型，帮助高价值的企业融资，并基于整个产业链的数据提供个性化的产品分配和营销策略。我们在各行各业真正见证了人工智能（AI）的巨大潜力，以及已经开始期待在许多应用中使用更复杂、更尖端的人工智能技术，包括无人驾驶、医疗、金融等。今天，人工智能技术几乎在各方面都大显身手。传统的人工智能系统依赖于集中管理的训练数据集，建立在大量数据上，从数据中学习特征，从而完成复杂的任务，甚至是人类也难以完成的操作。

人工智能的基础是大数据，而大多数训练数据是来自于不同组织的个人或机构。在一个深度学习的项目中，可能涉及到多个领域，需要采集不同公司、不同机构、不同部门的数据进行融合。（比如研究用户的消费爱好和水平，可能需要采集各个消费平台、银行、商店等多个机构的数据），然而在现实生活中，数据是分散在各地的，很难进行整合。于是诞生了集中式深度学习，它是通过收集数据并将其发送到一个能看到并控制所有数据的中央服务器，完成所有训练数据的整合。这个中心位置不仅要有强大的计算机集群来训练和创建深度学习模型，还要处理敏感数据并防止数据被用于其他目的。此外，敏感数据的处理方式必须不损害用户

的隐私。集中式的深度学习需要大量的数据去训练模型，达到较好的训练效果。比如，大量的互联网公司从数百万的用户那里收集数据，然后利用这些数据进行深度学习，实现智能推荐、语音识别、面部识别等。然而在集中式深度学习中，中央服务器是半可信的，半可信是指服务器是诚实但好奇的（Honest but Curious），在处理一些敏感数据时，用户也不知道他们的数据将被用于何处，用户隐私安全存在一定的威胁。

而且，这些数据的采集很可能涉及到用户的隐私。在 2018 年，中国互联网协会收到用户举报发现，腾讯音乐等多家应用软件以“通过深度学习向用户提供更好的服务”为由，长期收集并保存大量的用户个人数据，如照片、地址、电话等，甚至将这些包含了用户大量个人隐私的数据用作其他途径，为企业谋取更多利益。资料显示，许多应用软件在数据收集方面存在大量的安全漏洞，比如，过度用户手机中的“通讯录”、“位置”、“麦克风”等信息，未经授权访问用户的本地数据，导致千万级的用户资料泄露。

随着越来越多的涉及数据泄漏和隐私侵权事件的发酵，人们的隐私意识的普遍提高，越来越的用户关注自己的隐私信息是否在未经个人许可，或者出于商业和政治目的被他人或机构利用。更多的用户拒绝向互联网企业提供“收集数据”的权限，关闭了“通讯录”、“短信”、“位置”等访问权限。同时，相关的隐私法律法规不断完善，中国出台的《网络安全与数据合规》白皮书中明确要求加强用户个人信息保护，2018 年欧洲联盟会出台的《通用数据保护条例》强调保护用户的隐私和数据的使用安全性。随着个人意识和国家政策的关注，在大数据和人工智能领域数据采集和使用的过过程中，保护用户隐私和数据的机密显得越来越重要。

人工智能的力量是基于大数据的，在数据监管和隐私保护的要求下，传统的集中式深度学习系统难以收集到模型训练所需要的数据，进而无法提供更专业的网络服务。大数据的基础没有了，人工智能的未来也就岌岌可危。那么能否创建一个深度学习框架，使人工智能系统能够更有效和准确地集体使用数据，同时满足隐私性、安全性和监管要求，并解决数据孤岛的问题。如何才能做到这一点呢？

为了解决这个问题，google 在 2016 年率先提出了联邦学习的概念^[3]，它提供了一个具有隐私保护功能的分布式深度学习框架，该框架通过分布式的方式使成千上万的参与者协作，共同迭代训练一个联合的深度学习模型。这种机制允许参与者之间共享训练模型，但是训练数据在联合训练过程中仅保存在参与者的本地，确保了每个参与者的隐私。

如图1.1所示，联邦学习的基本工作流程如下：

- **初始化：**所有用户在他们的设备上都有一个预先分配的神经网络模型，并且可以自愿加入联邦学习协议，指定相同的深度学习和模型训练目标。
- **本地训练：**在一个给定的通信回合中，联邦学习参与者首先从中央服务器下载全局模型参数，然后使用他们的私人训练模式训练模型，更新本地模型（即模型参数），并将这些更新发送到中央服务器。
- **中央参数聚合：**中央服务器汇总此次通信回合中所有参与者上传的模型参数，并对其进行聚合求得全局模型的参数，然后更新全局模型。
- **迭代更新：**迭代地执行上述步骤直至全局模型参数满足收敛条件，最终得到最优的全局模型。

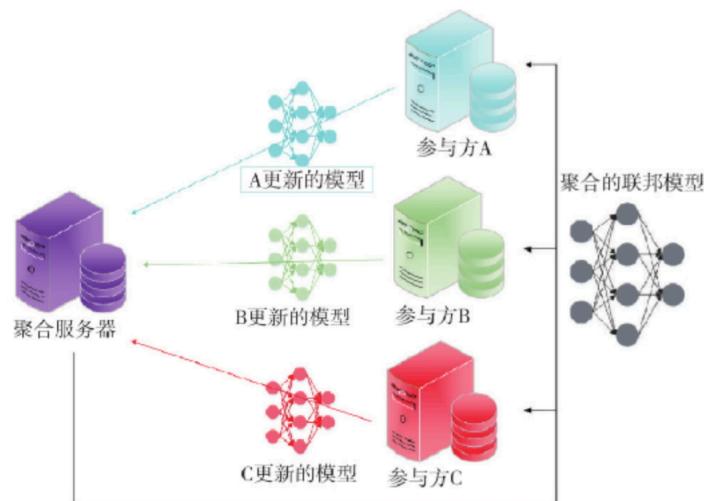


图 1.1: 联邦学习模型概况

联邦学习在隐私敏感的场景（包括金融、工业和许多其他与数据相关的场景）中展现出巨大的前景，这是因为它具有独特的优势，多个参与者无需共享本地数据却能训练出统一的全局模型，保护了本地数据的隐私性^[4]。联邦学习解决了数据聚合的问题，并允许一些机器学习模型和算法在各机构和部门之间进行独立设计和训练。在一些移动设备上的机器学习模型应用中，联邦学习展示出良好的性能和稳健性。此外，对于一些没有足够的私人数据来训练准确的本地模型的用户（客户）来说，深度学习模型和算法的性能可以通过联邦学习得到显著改善。

1.2 问题和挑战

1.2.1 数据异构

由于联邦学习的重点是通过以分布式方式从所有参与的客户端设备中学习本地数据来获得高质量的全局模型，所以它无法捕捉每个设备的个人信息，导致推理或分类性能下降。此外，传统的联邦学习要求所有参与的设备同意使用一个共同的模型来共同训练，这在复杂的现实世界物联网应用中是不现实的。研究人员对联邦学习在实际应用中面临的问题总结如下^[5]：

- (1) 设备的异质性：由于客户端设备的硬件条件（CPU、内存）、网络连接（3G、4G、5G、WiFi）和电源（电池）的变化，联邦学习网络上每个设备的通信、存储和计算能力都可能有差异。受限于网络和设备，不能保证在任何时候所有设备都能参与学习。此外，设备可能会受到意外事件的影响，如断电或断网，这可能会导致暂时的断网。这种异质性的系统结构影响了联邦模型的整体学习战略。
- (2) 统计的异质性：在整个网络中，设备通常以不同的方式产生和收集数据，而且不同设备的数据量、特征等会有很大的不同，所以联邦学习网络中的数据不是独立和相同的分布（非 IID）。目前的深度学习算法主要是基于 IID 数据。因此，非 IID 数据的异质属性给建模、分析和评估带来了重大挑战。联邦学习的参数聚合 FedAvg 方法可以解决非均匀同分布数据的问题，但是当数据分布偏态很严重的时候 FedAvg 的性能退化严重，一方面其性能比中心化的方法差好多，另一方面它

只能学习到 IoT 设备粗粒度的特征而无法学习到细粒度的特征。

(3) 模型的异质性：每个客户根据其应用场景要求定制不同模型。

1.2.2 高昂的通信代价

在联邦学习过程中，根据存储在几十甚至几百万个远程客户端设备上的数据来学习一个全局模型。原始数据被储存在本地的客户端设备上，在训练过程中这些远程设备必须不断地与中央服务器互动，以完成全局模型的构建。通常情况下，整个联盟学习网络可能涉及大量的设备，而网络通信可能比本地计算慢几个数量级，因此高通信成本成为联邦学习的关键瓶颈。

1.2.3 安全性和隐私威胁

联邦学习解决了传统集中式深度学习所面临的大规模数据收集等问题，减少了数据在收集过程中所遇到的隐私泄露风险，节省了传输数据所占用的通信资源。但是，联邦学习中的共享参数以及传输数据的无线链路仍然可能泄露数据隐私。

在联邦学习系统中，攻击方可能是内部攻击者，比如中央服务器、本地客户端；也有可能是外部攻击者。他们试图影响、破坏联邦学习模型的准确性，恶意推导本地客户端的训练数据。外部攻击主要通过本地客户端与中央服务器之间的通信信道发起。

有一些恶意参与者会发送无效的模型参数更新到中央服务器，破坏全局模型的训练。比如，这些恶意参与方作为本地客户端参加训练，修改本地的训练数据，对本地数据注入一些有毒的数据，进行投毒攻击，从而损害全局模型的准确性，操纵模型的预测结果。

在训练过程中，局部模型更新和全局模型参数的结合过程，提供了关于训练数据的隐藏知识，用户的个人信息很有可能泄露给不受信任的服务器或其他恶意用户。例如，白盒推理攻击和黑盒推理攻击通过客户端上传的参数恶意的窃取用户的训练数据生成的样本原型。

1.3 国内外研究现状

尽管联邦学习框架提供了隐私保护的机制，还是有各种类型的攻击方式可以攻击联邦学习系统，从而破坏联邦学习系统安全和参与方的隐私。本节将重点讨论关于联邦学习的攻击模型和隐私保护的研究现状。

1.3.1 攻击模型的研究现状

各类攻击模型阻碍了深度学习技术的发展，也会极大地威胁到人们的隐私敏感信息。无论是模型并行化还是数据并行化，分布式学习系统在用户数据隐私性方面相对于集中式学习存在一定的优势。但文献^[6]的作者发现，在分布式联邦学习系统中，参与者需要多次的联合迭代过程才能完成全局模型的收敛，参与者的参数也需要多次的训练、上传和共享，这些参数中包含的参与者训练集的相关信息，用户的信息可以通过计算用户上传的多个参数得到。因此，有许多外部攻击者或者恶意服务器通过用户上传的参数恢复出原始的样本试例。

模型反演攻击^[7]利用用户上传的参数信息，以一种很简单的方式攻击用户数据：一旦用户的网络模型经过训练并达到收敛，攻击者就可以通过调整网络模型权重的梯度，获得网络模型中所有表示类的逆向工程试例。在模型反演攻击中，攻击者无需接触目标信息的标签类，攻击模型仍然能够恢复原始样本试例。这一攻击模型表明，任何经过精确训练的深度学习网络，无论是以何种方式进行训练收敛，都可以泄露深度网络中区分不同标签类的信息。但是参数中包含的信息有限，模型反演攻击方式很难攻击卷积神经网络等复杂深度网络模型，在模型进行了一定的隐私保护后，攻击也基本失效。

生成对抗攻击（GAN 攻击）：目前研究人员也利用诸多安全模型对深度学习网络的训练数据集进行保护，但 Hitaj 等人^[8]发现，一个联邦学习框架非常容易受到系统内参与者发起的主动攻击。他们首次提出了基于 GAN 的模型重建攻击，攻击者为本地客户端。在训练阶段，攻击者冒充为本地的无害用户，训练 GAN 模型，模拟产生其他用户的训练数据产生的原型样本，之后通过不断添加假的训练样本，

攻击可以逐渐影响整个学习过程，使受害者暴露出更多关于被攻击者的目标类的敏感信息。除了客户端发起的 GAN 攻击，服务器也能通过 GAN 攻击。恶意服务器最初假装是一个为用户提供联邦学习服务的正常服务器，其主要目标是重建被攻击用户的训练样本。

投毒攻击：在联邦学习框架中，攻击者可能试图修改、删除或插入恶意信息到训练数据中，以破坏原始数据分布，改变学习算法的逻辑。两种常见的中毒攻击的例子包括标签反转攻击^[9] 和后门攻击^[10]。标签反转攻击是指恶意用户反转样本标签，并在训练数据中加入预定义的攻击点，导致训练后的模型偏离预测的界限。与标签反转攻击不同，后门攻击要求攻击者用精心设计的训练数据，利用特定的隐藏模式来训练目标的深度神经网络（DNN）模型。这些模型被称为”反馈回路”，可以干扰学习模型，并在预测阶段产生与真实情况截然不同的结果。

成员推理攻击：给定一个数据点和一个预训练过的模型，判断该数据点是否被用于训练该模型。在联邦学习中，每轮迭代的梯度都被发送给了服务器，在成员推理攻击中，中央服务器有能力推断一个特定数据点是否在本地训练集中。在一些情况下，它可以直接导致隐私泄露。例如，发现特定患者的临床记录用于训练与疾病相关的模型会泄露该患者患有疾病的事实。在实践中，Melis 等 [7] 证明了一个恶意攻击者可以准确判断一个特定位置档案是否被用于一个性别分类器在 FourSquare 位置数据集上的训练，准确率达 0.99。

1.3.2 隐私保护的研究现状

随着深度学习中攻击模型增多，研究人员开始关注训练网络模型时存在的隐私安全问题。在联邦学习中，存在着无数与隐私有关的挑战学习中的隐私问题。除了保证隐私之外，也要保证确保通信成本的低廉和高效。有许多关于联邦学习的隐私定义^{[11][12][13]}，主要分为全局隐私和局部隐私。在本地局部隐私中，每个客户端发送一个不同的隐私值，该值被安全的加密的上传到中央服务器。在全局隐私中，服务器在最终输出中添加不同的隐私噪音。安全多方计算、同态加密和差分隐私是最常见的保证联邦学习中的安全和隐私的技术。在分布式环境下，常用密码学中的

同态加密（Homomorphic Encryption）和本地差分隐私（Local Differential Privacy, LDP）技术来解决分布式数据收集中的隐私保护问题，保证数据收集者不能拥有任何个体用户数据的准确值，但是仍能获取用户数据的一些基本统计信息。

安全多方计算（Secure Multi-Party Computation）是由姚期智在 1982 年提出。多个参与者在不泄露各自隐私数据情况下，利用隐私数据参与保密计算，共同完成某项计算任务。目前，在安全多方计算领域，主要用到的是技术是秘密共享、不经意传输、混淆电路、同态加密、零知识证明等关键技术。

同态加密是一种加密形式，允许在加密之后的密文上直接进行计算，且计算结果解密后和明文的计算结果一致的加密算法，利用同态加密技术可以实现让解密方只能获知最后的结果，而无法获得每一个密文的消息，可以提高信息的安全性。如果对密文进行加法（或乘法）运算后解密，与明文进行加法（或乘法）运算的结果相等，则称这种加密算法为加法（乘法）同态。如果同时满足加法和乘法同态，则称为全同态加密。在联邦学习中，因为只需要对中间结果或模型进行聚合，一般使用的同态加密算法为 PHE（加法同态加密算法），在加密机制下进行本地客户端和云服务器的参数交换，保护用户数据隐私^[20]，例如在 FATE 中使用的 Paillier 即为加法同态加密算法。

差分隐私方法的主要原理是向数据添加噪音，或使用概括方法来掩盖某些敏感属性，使至多相差 1 条数据的 2 个数据集的查询结果概率不可区分，以保护用户的隐私。在联邦学习框架中，通过在本地模型和全局模型中对相关训练参数添加噪音，进行扰动，使对手无法获得真实的模型参数，进而防御模型反演攻击、成员推理攻击等。在深度学习中，差分隐私可以作为一种局部隐私保护方案来保护用户梯度的隐私，Abadi 等人^[23] 提出了一种隐私保护的深度学习方法，主要通过添加噪音来扰乱少量步骤后的局部梯度，将差分隐私机制与模型训练中的随机梯度下降算法（SGD）相结合。令人担忧的是，现有的差分隐私保护方案很难权衡隐私保护预算的成本和联邦学习模型的有效性，因为较高的隐私保护预算可能对一些大规模的攻击（如基于 GAN 的攻击）不是很有用^[24]，而较低的隐私保护预算可能阻

碍模型的局部收敛。而且与安全多方计算等密码学技术相比，差分隐私无法保证参数传递过程中的机密性。

总的来说，安全多方计算基于复杂的计算协议，同态加密的运算成本非常高，而差分隐私破坏了数据的可用性，很难在模型性能和隐私成本上达到平衡，当前的研究方向主要集中在对数据集和神经网络中的参数的加密和隐私保护机制上，较少关注到模型整体框架等过程。目前的联邦学习中的隐私保护方法还有许多不足，不能在隐私性和模型可用性上都达到一个相对满意的效果，此外，大部分方法是基于统一的、固定的参数设置，会导致模型迭代过程中累积大量隐私损失，使模型性能大幅下降。因此，在联邦学习场景下，保护用户隐私的同时维持模型准确性仍需大量的研究。

1.4 本文工作与主要贡献

针对联邦学习中隐私性和模型精度的双重指标，本文提出了本地自适应差分隐私算法和安全混洗框架，主要的工作和贡献包含以下三个方面：

- (1) 在联邦学习差分隐私的场景下，本文提出了一种新型的、基于本地差分隐私的权重分配自适应干扰算法。在客户端本地训练的神经网络模型中，通过分析前向传播算法，计算每个属性类对于模型输出的贡献比，然后，我们设计了一个自适应噪声添加的方案，根据贡献率注入不同隐私预算的噪声。与传统的注入噪声的方法相比，我们在相同的隐私保护程度下最大限度地提高了模型的准确性，减少噪声对模型输出结果的影响，提高模型精度。
- (2) 考虑到联邦学习中参数聚合器的攻击和针对参数传播信道的攻击，本文提出了一种新的安全聚合机制，在本地客户端和中心服务器之间新增混洗器，在用户将参数上传到云服务器之前，先对参数进行混洗，模型参数的更新被匿名的发送到混洗器，通过对模型参数的拆分和混洗实现客户端匿名，并且证明了安全混洗模型的可行性。

(3) 本文通过实验，展示了自适应本地差分隐私方案和安全混洗框架的结合，在较低的隐私预算下，维持了联邦学习模型的精度。

1.5 本文组织结构

本文一共六章，主要内容的组织安排如下：

第一章对本文研究内容：联邦学习的研究背景、国内外研究现状进行了阐述，介绍了目前联邦学习中的隐私保护的研究现状和发展方向。

第二章详细介绍本文研究内容所涉及的一些理论基础与背景知识，包含了联邦学习的相关概念，差分隐私的基础知识和神经网络的基本结构。

第三章描述了本文所提出的本地自适应差分隐私算法的设计和实现，根据神经网络前向传播算法，分析属性值的贡献度，根据属性值自适应添加高斯噪声，然后采用解析高斯机制分析添加的噪声大小，并证明了在自适应差分隐私机制下的联邦学习算法的隐私性。

第四章在上一章的基础之上，提出了一种联邦学习安全混洗模型，混洗器对客户端上传的梯度进行采样后，然后拆分混洗，再将混洗模型和自适应本地差分隐私保护方法结合在分布式系统中，提高系统学习效果。并且证明了安全混洗模型的隐私性和收敛性。

第五章为实验部分，基于本文提出的隐私保护框架，我们在三个基准数据集的进行了实验和讨论，并与之前的差分隐私联邦学习框架进行对比实验。

第六章是对本文的一个内容总结和展望，首先对本文的研究内容进行了概括，并对现有的不足进行总结，对未来的研究和改进方向进行了展望。

1.6 本章小结

这一章节为绪论，主要介绍的是本文章的研究背景以及意义，对当下联邦学习中的应用以及存在的问题与挑战行了介绍和总结、讨论了联邦学习中隐私威胁和隐私保护的国内外研究现状，并对文章的主要工作和文章的章节进行了介绍。

第二章 基础知识

我们在本章节中介绍了本文研究所需要的一些基本知识，有助于更好的理解之后章节的内容。

2.1 深度神经网络

深度神经网络基于模块化思想，通过在多个层次上部署多个神经元并通过逐层训练的手段调整神经元间的连接权值，从而实现原始特征数据进行多次非线性变换，对于任何有限给定输入/输出数据的拟合，最终获取到稳定的特征用于后续的问题分析。

神经网络的设计来源于人脑的结构，是人脑处理信息方式的一个简化模型。人类的大脑是人中枢神经系统中的主要部分，这些神经元像网状物一样相互连接。来自外部环境的刺激或来自感觉器官的输入通过感受器之后进入传入神经，神经元一层一层兴奋（激活）后，传导到神经中枢（大脑或脊髓），相当于输出层，神经中枢根据信号的类型做出不同的判断（分类），然后再下达命令，将信号传递到输出神经。不同的信号，大脑都可以进行学习和分辨，而这一通用的模型，就是神经网络。如果你能把几乎任何传感器接入到大脑中，大脑的学习算法就能找出学习数据的方法，并处理这些数据。

如图2.1所示，神经网络的基本单元是神经元，由数千甚至数百万个简单的神经元组成，这些神经元密集地相互连接，神经元按层排列。每一层有多个神经元，层与层之间是“前馈传播”的，也就是说，网络中的数据只在一个方向上移动。一个单独的神经元可能与它前面一层的几个神经元相连，它从这些神经元接收数据；与它后面一层的几个神经元相连，它向这些神经元发送数据。神经元之间不存在

同层连接，也不存在跨层连接。

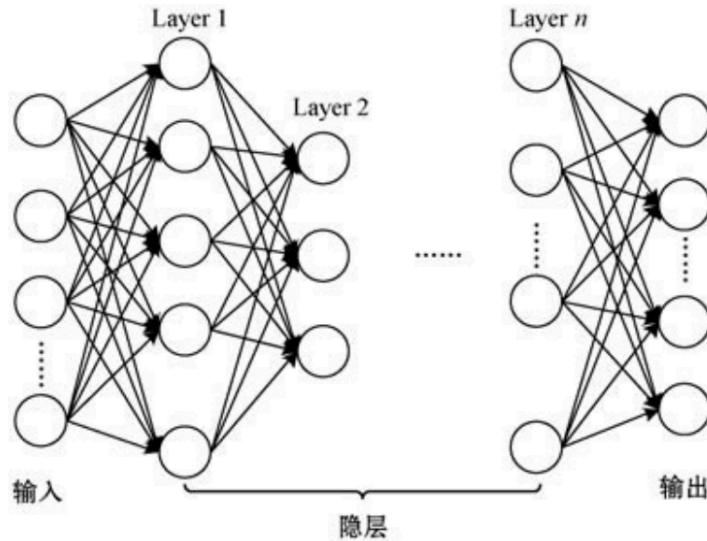


图 2.1: 深度神经网络结构图

神经网络通常有三个部分：一个输入层，主要用于获取输入的信息；一个或多个隐藏层，主要进行特征提取，调整权重让隐藏层的神经单元对某种模式形成反应；以及一个输出层，对接隐藏层并输出模型结果，调整权重以对不同的隐藏层神经元刺激形成正确的反应。当一个神经网络被训练时，其所有的权重和阈值最初都被设置为随机值。训练数据被送入输入层-并通过后续隐藏层，以复杂的方式相乘和相加，直到最后到达输出层，从根本上改变了数据。在训练过程中，不断调整网络的权重和阈值，直到具有相同标签的训练数据持续产生类似的输出。

所谓的前向传播算法就是：将上一层的输出作为下一层的输入，并计算下一层的输出，一直到运算到输出层为止。如图2.2所示，假设现有输入层的训练数据为 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, $x_i \in R^d$, $y_i \in R^l$, 即输入样本由 d 个属性描述，输出是 l 维实值向量。

假设隐藏层神经元个数为 q 个， θ_j 表示输出层神经元的阈值。隐藏层第 h 个神经元的阈值用 γ_h 表示。输入层第 i 个神经元与隐藏层第 h 个神经元之间的连接权为 v_{ih} 。隐藏层第 h 个神经元与输出层第 j 个神经元之间的连接权为 w_{hj} 。隐藏层第 h 个神经元接收到的输入为 $\alpha_h = \sum_{i=1}^d v_{ih} x_i$ ，输出层第 j 个神经元接收到的

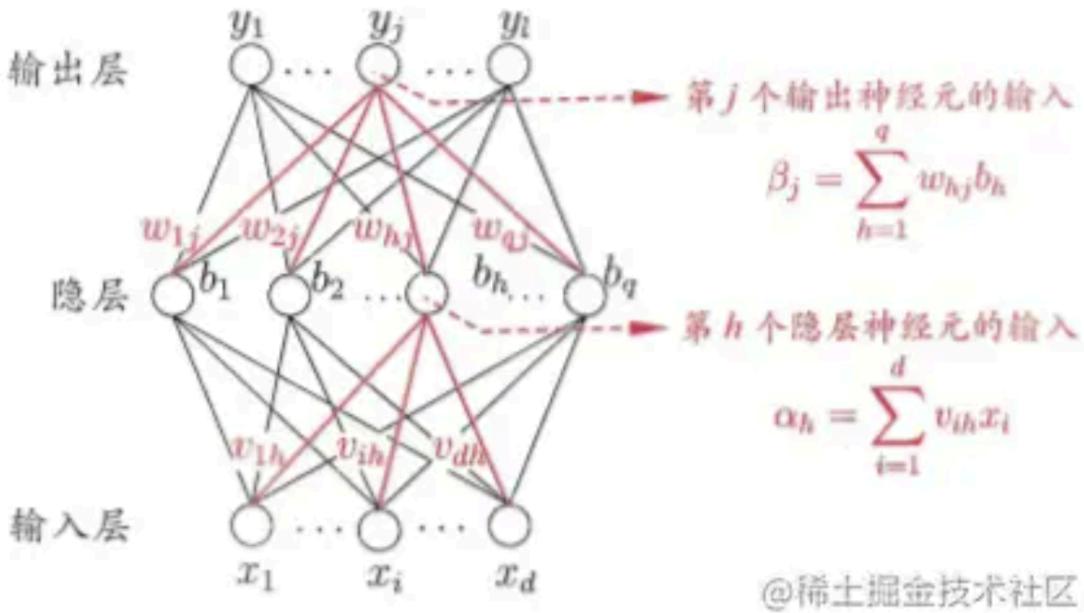


图 2.2: 前馈神经网络结构图

输入为 $\beta_j = \sum_{h=1}^q w_{hj} b_h$ 。其中 b_h 为隐藏层层第 h 个神经元的输出。假设隐层和输出层神经元都使用 Sigmoid 激活函数。对训练例 (x_k, y_k) , 假定神经网络的输出为 $\hat{y}_k = (\hat{y}_1^k, \hat{y}_2^k, \dots, \hat{y}_l^k)$, 即神经网络的预测输出表达式为:

$$\hat{y}_j^k = f(\beta_j - \theta_j) \quad (2.1)$$

那么如何评估所提神经网络输出预测值与真实值之间的差异程度呢? 这里提出损失函数 L , 文中采用均方差损失函数, 表示为:

$$L(\theta, x) = \frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2 \quad (2.2)$$

2.2式中: θ 为待训练的神经网络权重系数; x 表示目标值; y 表示预测值输出, 标 i 表示样本标签。深度神经网络算法训练的目的就是使得损失函数 L 最小。而对于复杂的神经网络而言, 最小化损失函数 L 通常采用随机梯度下降 (stochastic gradient descent, SGD) 算法来完成。即每次迭代过程中随机进行批量抽取训练样本 (记为 B), 并计算损失函数 L 的偏导数 $g_B = \frac{1}{|B|} \sum_{x \in B} \nabla_{\theta} L(\theta, x)$, 然后沿着负梯度方向 $-g_B$ 朝向局部最小值进行更新权重系数 θ 。

2.2 联邦学习

2.2.1 基本介绍

深度学习的成功应用需要建立在大量数据的基础之上，才能完成人们指派的学习任务。然而，近年来数据泄漏和隐私侵权事件不断发生，用户开始更加关注他们的隐私信息是否未经自己的许可，或被他人出于商业或者政治目的而被利用。人们逐渐地意识到，在人工智能的构建与使用的过程中保护用户隐私和数据机密的重要性。

大部分拥有的训练数据是由不同组织的个人、部门产生并拥有的，传统机器学习的做法是收集数据并传输到一个中心服务器，服务器可以看见并控制所有的数据，因此这个中心点不仅需要拥有高性能的计算集群来训练和建立机器学习模型，而且还需要处理敏感数据，避免泄漏用户隐私。然而，这种方法需要用户对服务器的完全信任，这已经不再有效或适用了。在这样的情况下，数据拥有者倾向于将自己的数据保留在自己的手中，进而会形成各自孤立的数据孤岛，至此大量数据的基础已经消失，人工智能的未来将面临绝境。作为回应，2016年谷歌^[25]率先提出联邦学习概念，旨在建立高质量分布式学习的框架。在联邦学习系统中，数据所有者（参与者）不需要彼此共享原始数据，也不需要依赖单个可信实体（中心服务器）来进行机器学习模型的分布式训练。相反，参与者通过在自己的本地数据上执行本地训练算法，并且只与参数服务器共享模型参数，来共同协作训练联邦模型。在每轮训练中，参数聚合节点会随机选择合适的节点加入到训练池中。那些被选中的本地节点通常是保持充电且无线网络可用。然后参数聚合节点平均所有已提交者的权重并作为下一轮回合的初始化模型。重复此过程直至终止条件。

根据用户维度和模型特征维度的重合去分类，将联合学习分为水平联邦学习、纵向联邦学习和联合迁移学习^[26]。

- **水平联邦学习：**当两个数据集的用户属性重叠较多而用户重叠较少的情况下，我们对数据集进行横向切割（即按用户维度切割），取出两边用户属性相同

但用户不完全相同的那部分数据用于训练。这种方法被称为横向联合学习。例如，两家银行位于不同的地区，有来自各自地区的用户群，而且它们之间的联系非常少。然而，他们的业务活动非常相似，因此他们的用户特征也是一样的。在这个阶段，我们可以使用跨部门的联合学习来建立一个联合模型。2016年，谷歌提出了一个在安卓手机上更新模型的联合数据建模系统：模型参数在本地不断更新，并在各个用户使用安卓手机时上传到安卓云端，使拥有数据的每一方都能建立一个具有相同特征维度的联合模型。

- **纵向联邦学习：**在两个数据集中用户重叠较多，而用户属性重叠较少的情况下，我们将数据集纵向切开（即按特征维度），选择数据集中两边用户相同但用户属性不完全相同的部分进行训练。这种方法被称为纵向的联合学习。例如，有两个不同的组织，一个是在一个地方的银行，另一个是在同一个地方的电子商务公司。他们的用户群很可能包括该地的大部分人口，所以有很大的用户交集。然而，由于银行储存的是用户的收入和支出以及信用评分的数据，而电子商务公司储存的是用户的浏览和购买历史的数据，他们的用户档案并没有那么紧密的联系。长期的联邦学习是在一个加密的空间里将这些不同的功能结合起来，以提高模型的性能。渐渐地，人们发现可以在这个联合系统之上建立若干机器学习模型，如逻辑回归、树状结构和神经网络模型。
- **联合迁移学习：**联合迁移学习是通过使用迁移学习模型来弥补数据或标签的差距，而不是对数据进行切分，两个数据集中的用户和用户属性几乎没有重叠。这种方法被称为混合式学习迁移。这里举一个例子，考虑两个不同的组织，一个是中国的银行，另一个是美国的电子商务公司。由于地理上的限制，这两个机构的用户群重叠的地方很少。由于它们是不同类型的组织，数据的特点也没有太多的重叠。在这种情况下，为了保证有效的联邦学习，有必要引入反式学习，以克服单变量数据量小和标注样本小的问题，提高模型的效率。

2.2.2 模型框架

本文我们提出的方案是基于典型的分布式横向协作学习系统架构，即各个参与者的本地数据集特征空间相同，但样本不同。通过中心服务器，各个参与者相互协作，在保护个人本地敏感数据的同时，有效地提高本地学习效果。通常这种系统包含以下步骤：

- 步骤 1：中央服务器初始化联合训练模型，然后将初始参数传递给每一个本地客户端。
- 步骤 2：客户端在本地数据上使用中央服务器传递的模型参数进行模型训练。
- 步骤 3：中央服务器汇总此次通信回合中所有参与者上传的模型参数，并对其进行聚合求得全局模型的参数，然后更新全局模型。
- 步骤 4：客户端更新各自的本地模型，重复步骤 2-4 直至中心服务器的联合模型收敛。

2.3 差分隐私

差分隐私作为一种隐私保护方法是为一个用户服务的，因为根据隐私的定义，隐私泄露只是与特定用户有关的信息泄露，而一组用户的统计特征不包括在隐私信息中。如果一个对象在数据库中的存在或不存在，或其价值的变化不会对搜索结果产生重大影响，那么该对象的隐私信息就会受到保护，这就是差分隐私（DP）概念的起源。差分隐私首先被应用于数据查询，为了更好地说明数据集之间的差异，定义了相邻数据集的概念：两个数据集只差一个信息或只差一个数值不同的记录^[28]。因此，查询数据库相关信息的攻击者将无法以任何概率确定 X_n 是否存在于数据集中，而成员 X_n 被认为是相对安全的。

2.3.1 基本定义

对于一个有限域 Z , $z \in Z$ 为 Z 中的元素, 从 Z 中抽样所得 z 的集合组成数据集 D , 其样本量为 n , 属性的个数为维度 d 。对数据集 D 的各种映射函数被定义为查询 (Query), 用 $F = \{f_1, f_2, \dots\}$ 来表示一组查询, 算法 M 对查询 F 的结果进行处理, 使之满足隐私保护的条件, 此过程称为隐私保护机制。设数据集 D 和 D' , 具有相同的属性结构, 两者的对称差记作 $D\Delta D'$, $|D\Delta D'|$ 表示 $D\Delta D'$ 中记录的数量。若 $|D\Delta D'| = 1$, 则称 D 和 D' 为邻近数据集 (Adjacent Dataset)。

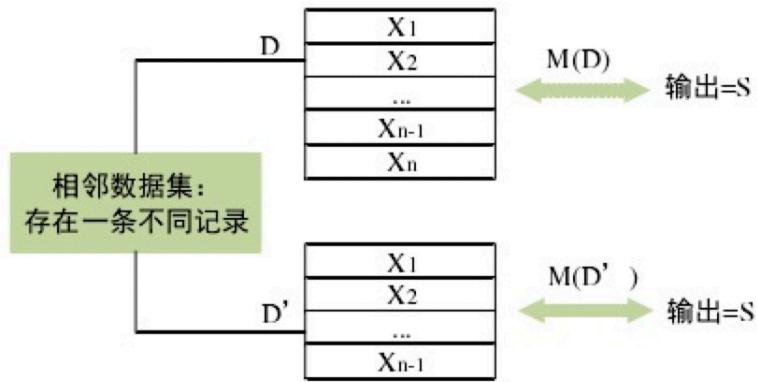


图 2.3: 差分隐私的相邻数据集示意图

定义 2.3.1 (成立条件). 若随机算法 $M : D \rightarrow R$ 满足 $(\varepsilon, \delta) - DP$, 当且仅当相邻数据集 d, d' 对于算法 M 的所有可能输出子集 $S \in R$ 满足不等式^[40] :

$$\Pr[M(d) \in S] \leq e^\varepsilon \Pr[M(d') \in S] + \delta$$

其中, ε 表示隐私预算参数, ε 越小意味着隐私预算越低, 信息泄露越少, 隐私保护的强度越高。添加项 δ 代表允许以概率 δ 打破 $\varepsilon - DP$ 的可能性, 其值通常选择为小于 $1/|D|$. 当 $\delta = 0$ 时, 定义转化为 $\varepsilon - DP$, 这时机制提供了更加严格的隐私保护。隐私预算参数决定着隐私保护强度, 针对传统数据库保护, 当 $\varepsilon \in (0, 1)$ 时认为隐私保护强度是有效的, 但应用在深度学习领域, $\varepsilon \in (0, 10)$ 都认为是可以被接受的合理范围。

2.3.2 相关概念

差分隐私保护可以通过在查询函数的返回值中加入适量的干扰噪声来实现。加入噪声过多会影响结果的可用性，过少则无法提供足够的安全保障。敏感度是决定加入噪声量大小的关键参数，它指删除数据集中任一记录对查询结果造成最大改变。在差分隐私保护方法中定义了两种敏感度，即全局敏感度 (Global Sensitivity) 和局部敏感度 (Local Sensitivity)。

定义 2.3.2 (全局敏感度). 设有函数 $f : D \rightarrow R^d$, 输入为一数据集, 输出为一 d 维实数向量。对于任意的邻近数据集 D 和 D' ,

$$GS_f = \max_{D, D'} \|f(D) - f(D')\|_1$$

称为函数 f 的全局敏感度。

函数的全局敏感度由函数本身决定，不同的函数会有不同的全局敏感度，一些函数具有较小的全局敏感度（例如计数函数，其全局敏感度为 1），因此只需加入少量噪声即可掩盖因一个记录被删除对查询结果所产生的影响，实现差分隐私保护。

定义 2.3.3 (局部敏感度). 对于一个查询函数 $f : D \rightarrow R^d$, 其中 D 为一个数据集, R^d 为 d 维实数向量, 是查询的返回结果。对于给定的数据集 D 和它的任意邻近数据集 D' , 有 f 在 D 上的局部敏感度为: $LS_f(D) = \max_{D'} \|f(D) - f(D')\|_1$

局部敏感度由函数及给定数据集中的具体数据共同决定。由于利用了数据集的数据分布特征，局部敏感度通常要比全局敏感度小得多。敏感度代表了查询函数针对相邻数据集的输出的最大不同，或者说量化评估了最坏情况下单个样本对整体数据带来的不确定性大小。敏感度函数仅与查询函数的类型有关，为扰动的添加提供了依据。但是，由于局部敏感度在一定程度上体现了数据集的数据分布特征，如果直接应用局部敏感度来计算噪声量则会泄露数据集中的敏感信息。

全局差分隐私技术旨在实现这样一个目标：如果替换数据集中的任意样本的效果足够小，则查询结果不能被用来探索数据集中任何样本的更多信息^[29]。作为一种优势，这种技术比局部差分隐私技术更准确，因为它不需要向数据集添加大量的噪声。局部差分隐私技术被引入以去除全局差分隐私中所要求的受信任的中央机构^[30]。与全局差分隐私技术相比，局部差分隐私技术不需要可信的第三方^[31]。其缺点是，噪声总量比全局差分隐私技术大得多。

可量化性、可组合性和后处理不变性是差分隐私最重要的三个性质。可量化性指的是差分隐私算法在计算特定随机化过程时，可以透明化、精准量化所施加的扰动，即上文提及的隐私预算。这样使用者就可以清楚地知道算法的隐私保护力度；组合性可以将相互独立的差分隐私算法进行组合；差分隐私的后处理不变性，确保了即使对算法的结果进行进一步处理，只要不引入额外信息，后处理就并不会削弱算法的隐私保护力度。通过组合定理，人们可以利用基础的差分隐私算法设计出复杂的满足差分隐私保证的系统，这也是差分隐私的重要优势之一。

在差分隐私部署过程中常常不仅仅在一处添加噪声，也仅仅针对数据集进隐私预算的分配有序列组合性和并行组合性两种组合特性：

定理 2.3.4 (串行组合). 给定 \mathbf{n} 个随机算法 $M_i (1 \leq i \leq n)$ 满足 $\varepsilon_i - DP$ ，那么针对一个数据库 D 而言，在 D 上的算法序列组合可以提供 $\varepsilon - DP$ ，其中 $\sum_{i=1}^n \varepsilon_i = \varepsilon$ 。

定理 2.3.5 (并行组合). 对于数据库 D ，当其被划分成 n 个不相交的子集 $\{D_1, D_2, \dots, D_n\}$ ，在每个子集上应用算法 M_i ，每个算法提供 $\varepsilon_i - DP$ ，则在序列 $\{D_1, D_2, \dots, D_n\}$ 上整体满足 $(\max \{\varepsilon_1, \dots, \varepsilon_n\}) - DP$

2.3.3 实现机制

在实践中为了使一个算法满足差分隐私保护的要求，对不同的问题有不同的实现方法，这些实现方法称为“机制”。拉普拉斯机制 (Laplace Mechanism)、指数机制 (ExponentialMechanism) 与高斯机制是三种最基础的差分隐私保护实现机制。其中，Laplace 机制和高斯适用于对数值型结果的保护，指数机制则适用于非数值型

结果。

在中心化差分隐私中，最为常用的扰动机制是拉普拉斯 (Laplace) 机制，该机制可以后期处理聚合查询（例如，计数、总和和均值）的结果以使它们差分私有。Laplace 分布是统计学中的概念，是一种连续的概率分布。

定理 2.3.6 (拉普拉斯机制). 如果随机变量的概率密度函数分布为：

$$f(x | \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) = \frac{1}{2b} \begin{cases} \exp\left(-\frac{\mu-x}{b}\right) & x < \mu \\ \exp\left(-\frac{x-\mu}{b}\right) & x \geq \mu \end{cases}$$

其中, D 表示数据集, $f(D)$ 表示的是查询函数, Y 表示的是 Laplace 随机噪声, $M(D)$ 表示的是最后的返回结果。 $M(D) = f(D) + Y$ 如果噪声 $Y \sim L(0, \frac{\Delta f}{b})$ 满足 $(\epsilon, 0)-$ ，则表示服从拉普拉斯分布的随机噪声。因此，当隐私预算确定时，敏感度越大，引入的噪声量越大。

对于非数值型的查询结果或数据，通常使用指数机制来随机选择离散的输出结果来满足差分隐私。指数机制整体的思想就是，当接收到一个查询之后，不是确定性的输出一个 R_i 结果，而是以一定的概率值返回结果，从而实现差分隐私。而这个概率值则是由打分函数确定，得分高的输出概率高，得分低的输出概率低。

定理 2.3.7 (指数机制). 指数机制满足差分隐私，如果：

$$A(D, u) = \left\{ p : \Pr[p \in O] \propto \exp\left(\frac{\epsilon u(D, p)}{2\Delta u}\right) \right\}$$

其中 Δu 为打分函数 $u(D, p)$ 的全局敏感性。由式2.3.7可知，打分越高被选择输出的概率越大。^[53]

与拉普拉斯机制类似高斯机制对输入的所有维度施加高斯噪声干扰 $N(0, \sigma^2)$ 。

定理 2.3.8 (高斯机制). 对于任意 $\epsilon \in (0, 1)$ 与 $c^2 > 2 \ln(1.25/\delta)$ ，参数满足 $\sigma \geq c\Delta_2 f / \epsilon$ 的高斯机制为 (ϵ, δ) -差分隐私。

2.4 联邦学习中的差分隐私

传统的联邦学习中使用差分隐私的主要流程如下所示：

- 本地计算: 客户端 i 根据本地数据库 \mathcal{D}_i 和接受的服务器的全局模型 w_G^t 作为本地的参数, 即 $w_i^t = w_G^t$, 进行梯度下降策略进行本地模型训练得到 w_i^{t+1} (t 表示当前 round)。
- 模型扰动: 每个客户端产生一个随机噪音 n, n 是符合高斯分布的, 使用 $\bar{w}_i^{t+1} = w_i^{t+1} + n$ 扰动本地模型 (这里注意 w 是一个矩阵, 那么 n 就对矩阵的每一个元素产生噪音)。
- 模型聚合: 服务器使用 FedAVG 算法聚合从客户端收到的 $\bar{w}_i t + 1$ 得到新的全局模型参数 w_G^{t+1} , 也就是扰动过的模型参数。
- 模型广播: 服务器将新的模型参数广播给每个客户端。
- 本地模型更新: 每个客户端接受新的模型参数, 重新进行本地计算。

上述的差分隐私技术将原始数据集中到一个数据中心, 然后发布满足差分隐私的相关统计信息, 我们称其为中央化差分隐私 (centralized differential privacy) 技术。因此, 中央化差分隐私对于敏感信息的保护始终基于一个前提假设: 可信的第三方数据收集者, 即保证第三方数据收集者不会窃取或泄露用户的敏感信息。然而, 在实际应用中, 即使第三方数据收集者宣称不会窃取和泄露用户的敏感信息, 用户的隐私依旧得不到保障。由此可知, 在实际应用中想要找到一个真正可信的第三方数据收集平台十分困难, 这极大地限制了中央化差分隐私技术的应用。鉴于此, 在不可信第三方数据收集者的场景下, 本地化差分隐私 (local differential privacy)^{[32][33]} 技术应运而生, 其在继承中央化差分隐私技术定量化定义隐私攻击的基础上, 细化了对个人敏感信息的保护。具体来说, 其将数据的隐私化处理过程转移到每个用户上, 使得用户能够单独地处理和保护个人敏感信息, 即进行更加彻底的隐私保护。目前, 本地化差分技术在工业界已经得到运用: 苹果公司将该技术应用在操作系统 IOS10 上以保护用户的设备数据, 谷歌公司同样使用该技术从 Chrome 浏览器采集用户的行为统计数据^[34]。

2.5 本章小结

本章对论文需要使用的一些基础理论知识进行了讨论。主要介绍了联邦学习系统的学习协议以及差分隐私的基本概念、定义和定理，分布式联邦学习系统是本论文主要使用的系统架构，所提的攻击模型和隐私对策都是基于该分布式联邦学习系统。本章同时也介绍了差分隐私及其变体的概念、实现机制。最后介绍了联邦学习中各个神经网络的基本结构和随机梯度下降算法。

第三章 联邦学习中的自适应本地差分机制

3.1 引言

与传统的集中式深度学习相比，联邦学习通过分布式训练在一定程度上缓解了隐私泄漏的问题。然而，许多研究表明，攻击者仍然可以通过模型训练的梯度损害用户的隐私 [13]。文献 [20] 表明，深度学习技术可以“记忆”模型中的训练数据信息。在这种情况下，敌方一旦通过白盒推理攻击或者黑盒推理攻击访问模型，就可以推演出客户端本地的训练数据。

在传统的集中式隐私保护方案中，数据管理者倾向于给每个用户的数据以相同的隐私预算。同样的隐私预算忽略了用户之间的差异。有些用户希望有更好的隐私保护。而有些用户对某些数据的隐私不敏感。在这种情况下，由于联邦学习模型是分布式结构，从一个大数据库到许多小数据库，所以对于每个用户来说。他们只需要关心他们自己的隐私。他们可以设置不同的隐私预算方案，而不是传统的统一分配，然后在最坏的情况下注入噪音。所以我们需要注入更少的整体噪音。

机器学习中模型的优化问题可以概括为 ERM（经验风险最小化）问题：

$$\arg \min_{\theta \in \mathcal{C}} \left(F(\theta) := \frac{1}{m} \sum_{i=1}^m F_i(\theta) \right) \quad (3.1)$$

从隐私保护的角度讲，我们只要截断了从原始输入到输出，在其中加入一道隐私保护屏障，具体在哪一步截断则对应于不同的方法。差分隐私保护机器学习的方法具体有以下几种：

- **输入扰动：** 输入扰动是在获取的训练数据上直接添加噪声，之后的模型训练和优化都是基于加噪后的训练数据。

- **输出扰动：**输出扰动沿袭了拉普拉斯机制最简单的思路，即考虑函数输出的敏感度来添加噪声，那么在 ERM 公式中我们只需要考虑 argmin 函数输出的敏感度，基于这个敏感度来添加拉普拉斯噪声即可得到一个简单的满足差分隐私的 ERM 方法。
- **梯度扰动：**梯度扰动是在执行最小化损失函数的过程中，设计满足差分隐私的算法。
- **目标扰动：**目标扰动是在模型的目标函数中添加一个随机量，以使得最终模型的输出满足随机性。

基于输入的扰动和输出的扰动基本可以视为一个黑匣子模型，简单直接。但是这种添加噪声的方式无法对训练过程中数据的相互依赖性和输出有效性作出有用的、紧密的描述。在输入数据中加入过多的噪声，可能会影响模型训练的收敛性。在输出参数中加入过于保守的噪声，也就是根据最坏的攻击情况去添加噪声，可能会影响模型的实用性。因此本文采用一种更加复杂的方法来分析训练过程中训练数据对模型输出的贡献比率，然后根据每一层神经网络对模型输出的贡献率，在梯度上自适应添加噪声。

基于梯度加噪的差分隐私保护方法作为主流的差分隐私应用于深度学习模型的方法之一，方案的目标是满足差分隐私条件下实现最优的模型可用性。文献 [1] 提出了一个 $(\epsilon_c + \epsilon_d)$ -差分隐私版本的随机梯度下降算法。在模型的每一次迭代过程中，对梯度添加高斯噪声，并通过差分隐私的组合性和隐私放大效果，得到完全隐私损失的上界。

本章提出的隐私保护方案是基于本地客户端的本地数据维度的，从以下三个方面展开研究：第一，通过在本地模型训练的梯度下降算法过程中针对不同层的贡献比自适应添加噪声；第二，采用解析高斯机制，计算对其梯度施加的噪声大小；第三，使用差分隐私的组合定理和后处理定理分析模型整体的隐私预算和性质。

在本文中，我们家是认为中央参数服务器是半可信的，一个“诚实但好奇”的

实体。也就是说，服务器将遵循与所有用户的协议。然而，通过利用完全访问用户梯度的便利，它也试图在训练过程中获得关于客户端的额外的信息。出于这个原因，我们的提出的自适应加噪机制目的是保护发送到服务器的本地梯度不被推断出任何关于用户的额外信息，并且尽量维持原有模型的精度。

3.2 自适应差分的 SGD 算法

算法1详细描述了在本地客户端训练过程中，在 SGD 算法中添加自适应差分隐私，并使用解析高斯机制衡量所添加的噪声大小。首先，我们采用先验组合机制计算 eps_{iter} 和 δ_{iter} （算法第 5 行）。每个客户端对训练数据进行采样，并计算他们的隐私预算 δ_u 。如果 $\delta_u > \delta$ ，用户将终止采样和训练，并且不上传其梯度信息（算法第 7-10 行）。否则，用户将用一个随机样本计算梯度（算法第 11-12 行）。然后使用解析机制对梯度进行剪辑并注入适当的噪声。最后，服务器对用户的梯度进行平均，并更新模型参数 w 。该算法有四个主要部分：自适应差分隐私，梯度范数裁剪，隐私预算累积，以及解析高斯计算噪声量。

在本节接下来的四个部分，我们将详细描述如何在神经网络的随机梯度下降算法中自适应添加噪声、梯度剪裁以及使用解析高斯机制衡量添加的噪声大小。

Algorithm 1 基于自适应差分隐私的随机梯度下降算法

```

1: 输入: 预估迭代次数  $T$ , 学习率  $\alpha$ , 梯度裁剪阈值  $C$ , 目标损失函数  $l$ , 解析高斯机制噪声
 $(\Delta, \varepsilon, \delta)$ 
2: 输出: 模型梯度
3: 初始化模型权重  $w$ 
4: while  $\exists \delta_u < \delta$  do
5:    $n=0$ 
6:    $grad=0$ 
7:   计算  $eps_{iter}$ ,  $\delta_{iter}$ 
8:   for each  $u \in Users$  do
9:     计算  $\delta_u$ 
10:    if  $\delta_u > \delta$  then
11:      continue
12:    end if
13:    从客户端数据集中随机采样
14:     $gt_u = \nabla l(w, x)$ 
15:     $gt_u = gt_u / \max\left(1, \frac{\|gt_u\|}{C}\right)$ 
16:     $n++$ 
17:  end for
 $w = w - \alpha * grad/n$ 
18: end while

```

3.2.1 层间依赖传播算法

如图3.1为神经网络的训练结构图。神经网络的模型结构可以简单分为输入层、隐藏层、输出层。在每一层下，都有很多神经元构成这一层的基本结构。输入层只有一个参数：激活值。输出层（包括隐藏层）神经元有三个参数：

- 权重：指的是和输入层某个神经元的紧密关系。联系越紧密这个值越大。
- 激活值：输出层的激活值是经过计算得到的，简单的计算就是把输入层的激活值乘以权重大后相加。
- 偏置：与线性方程 $y=ax+b$ 中的 b 的意义一致，偏置的存在能更好的拟合数据

这是实际应用中最常见的神经网络类型。第一层是输入，最后一层是输出。如果有多个隐藏层，我们称之为“深度”神经网络。他们计算出一系列改变样本相似性的变换。各层神经元的活动是前一层活动的非线性函数。

每个用户在本地用原始数据进行训练，在神经网络中进行前向传播操作，得到本地模型的输出。输入层的前向传播是神经网络中前向传播算法的第一步。

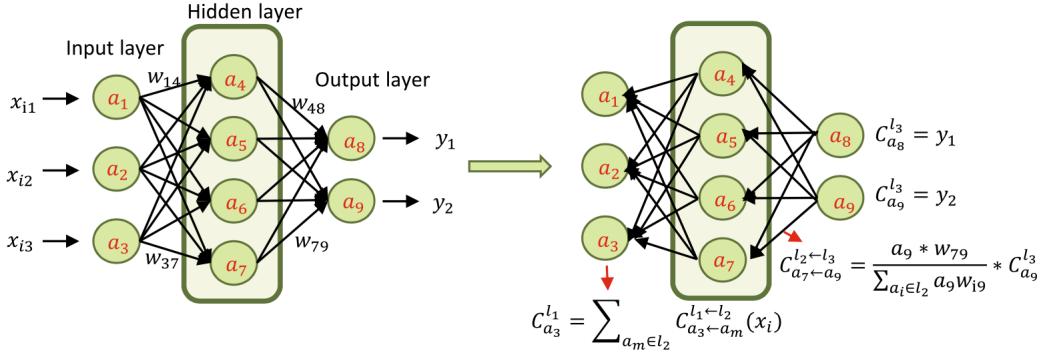


图 3.1: 层间依赖传播算法

根据矩阵层之间的线性相关性，神经元 a_i 在第 k 层的贡献 $C_{a_i}^{l_k}(x_i)$ 等于连接到神经元 a_i 的相邻层的贡献之和：

$$C_{a_i}^{l_k}(x_i) = \sum_{a_j \in l_{k+1}} C_{a_i \leftarrow a_j}^{l_k \leftarrow l_{k+1}}(x_i) \quad (3.2)$$

比如，在图3.1中，存在：

$$C_{a_7}^{l_2}(x_i) = \sum_{a_j \in l_3} C_{a_7 \leftarrow a_j}^{l_2 \leftarrow l_3}(x_i) = C_{a_7 \leftarrow a_8}^{l_2 \leftarrow l_3}(x_i) + C_{a_7 \leftarrow a_9}^{l_2 \leftarrow l_3}(x_i) \quad (3.3)$$

其中，“ \leftarrow ”表示两部分之间的连接关系。具体来说，“ $l_2 \leftarrow l_3$ ”是指神经网络中第二层和第三层之间相邻层的连接关系。那么对于第 k 个输出层：

$$C_{a_i}^{l_k}(x_i) = f(x_i, \omega_i^r) \quad (3.4)$$

因此，神经元 a_j 对于输出层的贡献等于模型的输出。第 k 层的神经元 a_j 对于第 $k-1$ 层的神经元 $C_{a_i \leftarrow a_j}^{l_{k-1} \leftarrow l_k}(x_i)$ 等于：

$$C_{a_i \leftarrow a_j}^{l_{k-1} \leftarrow l_k}(x_i) = \begin{cases} \frac{a_i w_{i,j}}{\sum_{a_i \in l_{k-1}} a_i w_{i,j}} C_{a_j}^{l_k}(x_i) & \sum_{a_i \in l_{k-1}} a_i w_{i,j} \neq 0 \\ \mu & \sum_{a_i \in l_{k-1}} a_i w_{i,j} = 0 \end{cases} \quad (3.5)$$

其中 μ 是一个无限接近于零，但大于零的数字。从上述公式中，我们可以认为每一层的贡献是相等的，而且贡献是逐层传递的。根据以上公式的推导，我们能得到神经网络模型中每一层以及每个神经元的贡献值。

3.2.2 自适应噪声添加

在第二章中介绍了关于神经网络的结构，

$$y = a(\mathbf{x} * \omega + b) \quad (3.6)$$

公式3.6是学习模型中每个隐藏神经元的转化过程。其中 \mathbf{x} 代表输入向量， y 是输出， b 和 ω 分别代表偏置项和权重矩阵。 $a()$ 是一个激活函数，用于结合线性变换和非线性变换。 $y = a(\mathbf{x} * \omega + b)$ 是线性变换部分。

由于神经网络的结构，上一层的输出是下一层的输入，由此我们可以得出，原始的训练数据只被第一隐层的线性变换所利用。直观地说，为了得到一个具有隐私保护的学习模型，我们可以在第一层隐藏层的数据中注入噪声。正如 Phan 等人^[36]提到的，对于线性变换有一种传统的方法，即向原始数据注入具有相同隐私预算的噪声，但是这容易导致隐私预算增加，并且使原始数据失真过多。因此，本文提出一种自适应噪声添加算法，针对每个梯度计算其贡献值，根据贡献值进行梯度裁剪并添加噪声。

首先，我们引入了两个调整因素 f 和 p 。其中， f 代表一个阈值，用于决定属性对模型结果输出的贡献是高还是低，其值由用户定义，即贡献超过阈值 f 的属性类对输出的贡献更大。然后，我们向所有这些属性注入自适应拉普拉斯噪声。当贡献率低于阈值 f 时，对这些属性进行概率选择。也就是说，我们选择概率为 $1 - p$ 的原始数据，并对一些概率为 p 的属性注入自适应拉普拉斯噪声。该公式如下：

$$\tilde{x}_{i,j} = \begin{cases} \ddot{x}_{i,j} & \beta \geq f \\ \bar{x}_{i,j} & \beta < f \end{cases} \quad (3.7)$$

其中 β 代表贡献率： $\beta = \frac{|\ddot{C}_j|}{\sum_{j=1}^u |\ddot{C}_j|}$ ，当 $\beta < f$ 时，我们有：

$$\bar{x}_{i,j} = \begin{cases} \dot{x}_{i,j} \text{ with probability } p \\ x_{i,j} \text{ with probability } 1 - p \end{cases} \quad (3.8)$$

f 和 p 是超参数，用户可以根据自己的情况来调整。

也就是说，隐私预算 ϵ_l 是根据贡献率: $\epsilon_j = \frac{u * |\tilde{C}_j|}{\sum_{j=1}^u |\tilde{C}_j|} * \epsilon_l$ 。按比例分配给每个属性类。自适应噪声按以下方式注入属性中:

$$x'_{i,j} = x_{i,j} + \frac{1}{|D_i^t|} \text{Lap} \left(\frac{GS_l}{\epsilon_j} \right) \quad (3.9)$$

在不丧失一般性的情况下，调整因子 f 和 p 的值与系统的准确性和隐私水平有关。即 f 越小， p 越大。越高的秘密水平，准确性越低，反之亦然。

我们采用层间相关性传播算法计算每一层对模型输出的贡献比率。每个用户都在本地对原始数据在神经网络中进行前向传播的训练，这可以获得一个新的数据操作，从而获得本地模型的输出。根据相邻层之间的线性关系，在第 k 层的神经元的贡献 $C_{a_i}^{l_k}(x_i)$ 等于连接到神经元 a_i 的相邻层的贡献之和:

$$C_{a_i}^{l_k}(x_i) = \sum_{a_j \in l_{k+1}} C_{a_i \leftarrow a_j}^{l_k \leftarrow l_{k+1}}(x_i) \quad (3.10)$$

例如图3.1所示，我们有:

$$C_{a_7}^{l_2}(x_i) = \sum_{a_j \in l_3} C_{a_7 \leftarrow a_j}^{l_2 \leftarrow l_3}(x_i) = C_{a_7 \leftarrow a_8}^{l_2 \leftarrow l_3}(x_i) + C_{a_7 \leftarrow a_9}^{l_2 \leftarrow l_3}(x_i) \quad (3.11)$$

其中，“ \leftarrow ”表示两部分之间的连接关系。” $l_2 \leftarrow l_3$ ”是指深度神经网络中第 2 层和第 3 层之间相邻层的连接关系。当第 k 层为输出层时，我们有:

$$C_{a_i}^{l_k}(x_i) = f(x_i, \omega_i^r) \quad (3.12)$$

3.2.3 梯度范数裁剪

在模型的每一轮迭代过程中，算法将计算添加了高斯噪声的梯度 $g = \nabla f + N(0, \Delta^2 \sigma^2 I)$ ，方差是 σ^2 。对梯度注入的噪声量 $\Delta^2 \sigma^2$ 会根据用户个体对于梯度 g 在二范数下的最大全局敏感度，即 Δ 。由于梯度的大小没有一个先验的界限，我们用二范数的固定值来裁剪每个梯度。

用户上传的梯度向量将可以改写为 $g = g / \max \left(1, \frac{\|g\|}{C} \right)$ ，其中 C 为裁剪阈值。参数裁剪行为确保了梯度值小于 i 一定的阈值，即当 $\|g\| \leq C$ 时，那么 g 保持不变；

当 $\|g\| > C$ 时, 它按比例缩小为 C 。可以注意到, 这种形式的梯度裁剪是获得全局灵敏度的常用的方法。

但是参数 $clipc$ 的值如果太小, 那么裁剪后的噪声会较小, 算法添加的噪声较少时可能会破坏梯度估计的无偏性; 但是如果不对梯度进行裁剪, 大量的噪声添加到每个梯度会导致模型的可用性大大降低。在模型训练前期, 梯度所包含的数据信息更多, 因此可以对应添加更多的高丝噪声, 使用较大的 $clipc$ 的值, 使得梯度裁剪后的模型偏差更小; 而在模型训练后期, 梯度所包含的数据信息相对较小了, 如果还使用相同的 $clipc$, 会引入很多不必要的噪声。

因此我们根据训练轮数和层间贡献率动态调整裁剪的值。在每次迭代中, 该算法使用方差为 $sigma^2$ 的高斯机制来计算噪声梯度 $g = \nabla f + N(0, \sigma^2 I)$ 。噪声 $sigma^2$ 的大小取决于一个个体在 l_2 规范下对 g 的最大影响, 即 $Delta$ 。由于对梯度的大小没有先验的约束, 我们以 l_2 规范对每个梯度进行剪辑。因此, 梯度向量 g 被 $g = g / \max\left(1, \frac{\|g\|}{C}\right)$ 取代, 以达到剪裁阈值 C 。这种剪裁保证了如果 $\|g\| \leq C$, 那么 $mathrm{g}$ 将被保留, 而如果 $\|g\| > C$, 它将被缩减为准则 C 。

3.2.4 解析高斯机制

高斯机制是差分隐私数据分析算法的基本组成部分。差分隐私的定义可以理解为: 如果删除或者替换数据集中的个体对输出分布的影响可以忽略不计, 那么对私有数据的计算不会泄漏数据集中个体的敏感信息。在第二章的基础知识中, 我们简要介绍了传统的高斯机制。它的定义如下:

定义 3.2.1. 对于任意 $\varepsilon \in (0, 1)$ 与 $c^2 > 2 \ln(1.25/\delta)$, 高斯噪声参数满足 $\sigma \geq c \Delta_2 f / \varepsilon$ 的高斯干扰机制为 (ε, δ) -差分隐私。

在此基础上我们自然的有两个疑问: 一是是否定义中的参数 σ 是使算法满足 (ε, δ) -差分隐私的最小值, 即是否可以施加更小的干扰来达到相同的差分隐私保护效果; 二是如果隐私预算 ε 大于 1 会发生什么。

Balle[文献]等人分析了传统的高斯机制在高隐私保护力度、隐私损失分析、低

隐私保护力度三个方面中存在的问题，提出了一种改进的解析高斯机制（Analytic Gaussian Mechanism）。传统的高斯机制施加的噪声干扰过大，方差公式在高隐私预算下过紧，并且无法适用于隐私预算大于 1 的低隐私保护力度场景。解析高斯机制针对这些局限性提出了解决方案。

解析高斯机制的核心思想是使用高斯累积分布函数的计算，来代替传统高斯机制的尾部约束近似。它的定义如下：

定义 3.2.2. 令 $f : \mathbb{N}^{N'} \rightarrow \mathbb{R}^d$ 表示一个 ℓ_2 敏感度为 Δ_2 的函数，对任意 $\varepsilon \geq 0$ 与 $\delta \in [0, 1]$ ，当且仅当 σ 满足下列不等式时，含参数 σ 的高斯机制满足 (ε, δ) -差分隐私：

$$\Phi\left(\frac{\Delta_2}{2\sigma} - \frac{\varepsilon\sigma}{\Delta_2}\right) - e^\varepsilon \Phi\left(-\frac{\Delta_2}{2\sigma} - \frac{\varepsilon\sigma}{\Delta_2}\right) \leq \delta \quad (3.13)$$

其中， $\Phi(t) = (1 + \text{erf}(t/\sqrt{2}))/2$ ， erf 是高斯误差函数，即 $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\eta^2} d\eta$ 。传统的差分隐私高斯机制的干扰大小 σ 可以直接计算，解析高斯机制的实现如下算法：

Algorithm 2 解析高斯算法

```

1: 输入:  $f, x, \Delta, \varepsilon, \delta$ 
2: 输出: 干扰后的  $f$ 
3: 令  $\delta_0 = \Phi(0) - e^\varepsilon \Phi(-\sqrt{2\varepsilon})$ 
4: if 如果  $\delta \geq \delta_0$  then
5:   定义  $B_\varepsilon^+(v) = \Phi(\sqrt{\varepsilon v}) - e^\varepsilon \Phi(-\sqrt{\varepsilon(v+2)})$ 
6:   计算  $v^* = \sup \{v \in \mathbb{R}_{\geq 0} : B_\varepsilon^+(v) \leq \delta\}$ 
7:   令  $\alpha = \sqrt{1+v^*/2} - \sqrt{v^*/2}$ 
8: else
9:   定义  $B_\varepsilon^-(u) = \Phi(-\sqrt{\varepsilon u}) - e^\varepsilon \Phi(-\sqrt{\varepsilon(u+2)})$ 
10:  计算  $u^* = \inf \{u \in \mathbb{R}_{\geq 0} : B_\varepsilon^-(u) \leq \delta\}$ 
11:  令  $\alpha = \sqrt{1+u^*/2} + \sqrt{u^*/2}$ 
12: end if
13: 令  $\sigma = \alpha\Delta/\sqrt{2\varepsilon}$ 
14: 输出:  $f(x) + \mathcal{N}(0, \sigma^2 I)$ 

```

3.3 隐私性证明

自适应差分 SGD 算法对线性变换函数进行了扰动，该函数满足 $(\epsilon_c + \epsilon_l)$ 差分隐私。证明如下：

在贡献中添加的扰动为：

$$\ddot{C}_j(x_i) = C_j(x_i) + \text{Lap}\left(\frac{GS_c}{\epsilon_c}\right), j \in [1, u] \quad (3.14)$$

它是满足 ϵ_c -差分隐私的。

贡献 GS_c 的敏感度为：

$$\begin{aligned} GS_c &= \frac{1}{|D|} \sum_{j=1}^u \left\| \sum_{x_i \in D} C_{x_{i,j}}(x_i) - \sum_{x'_i \in D'} C_{x'_{i,j}}(x'_i) \right\|_1 \\ &= \frac{1}{|D|} \sum_{j=1}^u \left\| C_{x_{n,j}}(x_n) - C_{x'_{n,j}}(x'_n) \right\|_1 \\ &\leq \frac{2}{|D|} \max \sum_{j=1}^u \|C_{x_{i,j}}(x_i)\|_1 \\ &\leq \frac{2u}{|D|} \end{aligned} \quad (3.15)$$

其中， u 和 $|D|$ 分别表示贡献的数量和元组，然后可以得到：

$$\begin{aligned} \frac{\Pr(\ddot{C}(D))}{\Pr(\ddot{C}(D'))} &= \frac{\prod_{j=1}^u \exp\left(\frac{\epsilon_c \left\| \frac{1}{|D|} \sum_{x_i \in D} C_j(x_i) - \ddot{C}_j(x_i) \right\|_1}{GS_c}\right)}{\prod_{j=1}^u \exp\left(\frac{\epsilon_c \left\| \frac{1}{|D'|} \sum_{x'_i \in D'} C_j(x'_i) - \ddot{C}_j(x'_i) \right\|_1}{GS_c}\right)} \\ &= \prod_{j=1}^u \exp\left(\frac{\epsilon_c}{|D|GS_c} \|C_j(x_n) - C_j(x'_n)\|_1\right) \\ &\leq \prod_{j=1}^u \exp\left(\frac{\epsilon_c}{|D|GS_c} \max \|C_j(x_n)\|_1\right) \\ &= \exp\left(\epsilon_c \frac{\max_{x_i \in D} \sum_{j=1}^u \|C_j(x_n)\|_1}{|D|GS_c}\right) \\ &\leq \exp(\epsilon_c) \end{aligned} \quad (3.16)$$

因此，添加噪声后的贡献值是满足 ϵ_c -差分隐私的。

假设两个相邻的批次 D_i^t 和 $D_i^{t'}$, 其最后一个元组 x_n 和 x'_n 不同, $z(D_i^t)$ 和 $z(D_i^{t'})$ 分别为线性变换函数。

一般来说, 我们把偏置项视为第一类数据属性, 即: $x_{i,0} = b_i$ 。线性转换可以改写为: $\ddot{\mathbf{z}}_{x \in D_i^t}(\omega) = \ddot{\mathbf{x}} * \omega$ 。线性变换的敏感性 GS_l 如下:

$$\begin{aligned}
GS_l &= \sum_{a_i \in l_1} \sum_{j=1}^u \left\| \sum_{x_i \in D_i^t} x_{i,j} - \sum_{x'_i \in D_i^{t'}} x'_{i,j} \right\|_1 \\
&= \sum_{a_i \in l_1} \sum_{j=1}^u \|x_{n,j} - x'_{n,j}\|_1 \\
&\leq \sum_{a_i \in l_1} \sum_{j=1}^u \max_{x_i \in D_i^t} \|x_{n,j}\|_1 \\
&\leq \sum_{a_i \in l_1} u
\end{aligned} \tag{3.17}$$

其中, $a_i \in l_1$ 是指第一隐藏层 l_1 中的神经元 a_i , u 是数据元组 $x_i \in D_i^t$ 中的属性数。它包括两个调整因素: f 和 p , 它们可以过滤多余的噪声。之后的属性的一般表达式如下:

$$\begin{aligned}
\tilde{x}_{i,j} &= [(1-f) + f * p] * \ddot{x}_{i,j} + f * (1-p) * x_{i,j} \\
&= [(1-f) + f * p] \left[x_{i,j} + \text{Lap} \left(\frac{GS_l}{\epsilon_j} \right) \right] + [f * (1-p)] x_{i,j} \\
&= x_{i,j} + [(1-f) + f * p] \left[\text{Lap} \left(\frac{GS_l}{\epsilon_j} \right) \right]
\end{aligned} \tag{3.18}$$

然后我们可以得到:

$$\begin{aligned}
\frac{\Pr(\ddot{\mathbf{z}}_{D_i^t}(\omega))}{\Pr(\ddot{\mathbf{z}}_{D_i^{t'}}(\omega))} &= \frac{\prod_{a_i \in l_1} \prod_{j=1}^u \exp\left(\frac{\epsilon_j \left\| \sum_{x_i \in D_i^t} x_{i,j} - \sum_{x_i \in D_i^{t'}} \tilde{x}_{i,j} \right\|_1}{GS_l}\right)}{\prod_{a_i \in l_1} \prod_{j=1}^u \exp\left(\frac{\epsilon_j \left\| \sum_{x'_i \in D_i^{t'}} x'_{i,j} - \sum_{x'_i \in D_i^t} \tilde{x}'_{i,j} \right\|_1}{GS_l}\right)} \\
&\leq \prod_{a_i \in l_1} \prod_{j=0}^u \exp\left(\frac{\epsilon_j}{GS_l} \left\| \sum_{x_i \in D_i^t} x_{i,j} - \sum_{x'_i \in D_i^{t'}} x'_{i,j} \right\|_1\right) \\
&\leq \prod_{a_i \in l_1} \prod_{j=0}^u \exp\left(\frac{\epsilon_j}{GS_l} \max_{x_i \in D_i^t} \|x_{n,j}\|_1\right) \\
&\leq \exp\left(\epsilon_l \frac{\sum_{a_i \in l_1} u \left[\sum_{j=1}^u \frac{|\tilde{C}_j|}{\sum_{j=1}^u |\tilde{C}_j|} \right]}{GS_l}\right) \\
&= \exp(\epsilon_l)
\end{aligned} \tag{3.19}$$

根据上述推倒证明可知，在联邦学习的神经网络中添加自适应噪声后，所上传的梯度是满足 $(\epsilon_c + \epsilon_l)$ 差分隐私的。在满足差分隐私的基础上，在下一节我们会给出隐私损失累积函数计算隐私成本。

3.4 隐私预算分析

对于本章所提出的在随机梯度下降算法的上进行差分隐私保护算法，除了确保算法运行的准确率以外，另一个重要的问题就是评估算法训练时的数据隐私损失成本。为此，提出隐私损失累积函数的概念来进行每次迭代过程访问训练数据的隐私损失以及随着训练进展时的累积隐私损失。为不失一般性，令 $\sigma = \frac{\sqrt{2 \log(1.25/\delta)}}{\varepsilon}$ ，文献 [36] 严格证明，对于抽样概率 $q = \frac{\ell}{N}$ 且 $\varepsilon < 1$ ，则对于完整样本而言，每次迭代过程都是 $(O(q\varepsilon), q\varepsilon)$ -差分隐私的。但文献并未对迭代过程以及噪声强度对差分隐私损失的影响展开研究，故无法对噪声强度以及剪切阈值 C 进行有依据的选取。故首先需要研究迭代过程对差分隐私的影响机制。

事实上，若令 $\sigma \geq c_2 \frac{q\sqrt{T \log(1/\delta)}}{\varepsilon}$ ，则同样应用文献 [36] 方法，可以严格证明算法对于任意的 $\varepsilon < c_1 q^2 T$ 都是 $(O(q\varepsilon\sqrt{T}), \delta)$ -差分隐私的，其中 c_1 和 c_2 为常数。与

文献^[36]相比,本文算法能够在相同迭代步骤下,大幅度降低 ε 的数值,对数据的隐私性保护更高。进一步地,对于两个相邻的数据集 $d, d' \in D$ 和映射机制 M ,引入一个辅助输入变量 aux 和输出 $o \in R$,定义映射机制 M 在输出 o 处的隐私损失为:

$$c(o; M, \text{aux}, d, d') \triangleq \log \frac{\Pr[M(\text{aux}, d) = o]}{\Pr[M(\text{aux}, d') = o]} \quad (3.20)$$

对于所提差分隐私SGD算法而言,神经网络各层权重系数的参数值与每次迭代过程中的差分隐私机制有着紧密的关联,从而对于给定的映射机制 M ,在第 λ 次迭代过程的隐私损失定义为:

$$\alpha_M(\lambda; \text{aux}, d, d') \triangleq \log \mathbb{E}_{o \sim M(\text{aux}, d)} [\exp(\lambda c(o); M(d, d'))] \quad (3.21)$$

进一步地,映射机制 M 的损失边界值定义为:

$$\alpha_M(\lambda) \triangleq \max_{\text{aux}, d, d'} \alpha_M(\lambda; \text{aux}, d, d') \quad (3.22)$$

其满足以下特性:

- 组合特性:给定一个机制 M ,由一组子机制顺序 $\{M_1, M_2, \dots, M_k\}$ 组成,并满足 $M_i : \prod_{j=1}^{i-1} R_j \times D \rightarrow R_i$,从而总隐私损失边界满足:

$$\alpha_M(\lambda) \leq \sum_{i=1}^k \alpha_{M_i}(\lambda) \quad (3.23)$$

- 差分隐私边界: $\forall \varepsilon > 0$,映射机制 M 是 (ε, δ) 差分隐私的,当且仅当:

$$\delta = \min_{\lambda} \exp(\alpha_M(\lambda) - \lambda \varepsilon) \quad (3.24)$$

上述2条性质确定了深度神经网络算法每次迭代的隐私损失以及所能够达到侵犯数据隐私容忍度的最大迭代次数。特别地,在附加高斯噪声的情况下,不妨令 μ_0, μ_1 分别为 $N(0, \sigma^2)$ 和 $N(0, \sigma^2)$ 的概率密度函数,而 μ 为两个高斯密度函数的混合概率密度函数,即 $\mu = (1-q)\mu_0 + q\mu_1$ 。依据式3.21-3.24可推导得 $\alpha(\lambda) = \log \max(E_1, E_2)$,其中:

$$E_1 = \mathbb{E}_{z \sim \mu_0} \left[\left(\frac{\mu_0(z)}{\mu(z)} \right)^\lambda \right] \quad (3.25)$$

$$E_1 = \mathbb{E}_{z \sim \mu_0} \left[\left(\frac{\mu_0(z)}{\mu(z)} \right)^\lambda \right] \quad (3.26)$$

3.5 本章总结

联邦学习以分布式学习技术为基础，使参与者彼此通过一定的方式（如中心服务器）联合起来训练一个神经网络。在这个过程中，参与者不需要将自己的隐私数据暴露出来便可以参与协作训练，可以克服参与者本地数据集较小、数据样本比较单一、隐私泄露等缺点。虽然基本的分布式协作深度学习没有直接暴露参与者的隐私数据集，但是恶意攻击者仍然可以通过共享的参数等信息获得一定的隐私信息。

本章详细介绍了基于梯度自适应加噪的差分隐私保护模型对于模型准确度的影响。其中梯度下降作为一种常见的深度学习优化方法，将梯度进行噪声扰动是最早被提出、也是目前相对主流的差分隐私加噪方案之一。我们设计了一个自适应噪声添加的方案，根据贡献率注入不同隐私预算的噪声。与传统的注入噪声的方法相比，我们在相同的隐私保护程度下最大限度地提高了模型的准确性。然后我们采用解析高斯机制计算对其梯度施加的干扰大小，最后分析了模型整体的隐私预算。然而，客户端的匿名性不足以防止侧信道链接攻击，例如，如果客户端在每次迭代中同时上传了大量的权重更新，云仍然可以将它们链接在一起。因此下一章将针对一种训练轮数无关的安全聚合模型进行研究。

第四章 联邦学习的安全混洗模型

4.1 引言

上一章节中我们提出的方案是针对本地差分隐私，在我们的威胁模型中，对手可能是一个用户或者第三方，并且对手还可能是除其他用户和第三方外的中心服务器，这是一个相当强的威胁模型假设，因为除了用户本身，所有对象都是不可信的，用户在上传信息时需要经过满足差分隐私的噪声扰动。然而，强大的隐私也带来了模型可用性的问题，特别是当用户的数据量特别小时，聚合所有用户的参数会带来大量的噪声，从而降低了模型的精度。

上一章节中所提出的本地自适应差分隐私方案是通过在客户端将梯度上传至参数服务器前，对梯度添加自适应噪声，尽管方案采用了本地差分技术减少一定程度的隐私预算，但不可避免的会降低联邦学习模型的准确性以及学习效率。正如^[37] 所指出的，一个复杂的隐私保护系统将多个局部差异化的算法进行组合，从而导致这些算法的隐私成本增长。也就是说，隐私预算为 ϵ_1 和 ϵ_2 的局部差异化算法的组合会消耗的隐私预算总和为 $\epsilon_1 + \epsilon_2$ 。使用联邦学习训练的联合模型需要客户在多次迭代中向中央服务器上传梯度更新。如果在迭代训练过程中的每一次迭代都应用本地自适应差分隐私，隐私预算就会累积起来，从而导致总隐私预算的爆炸。现有的本地差分隐私协议对于多维聚集 FL 可能是不可行的，局部噪声带来的误差会随着维度系数的增加而加剧^[38]，从而大大降低模型的精度。而且，当参与一次迭代的客户端数量达到上千人时，会导致聚合任务升级成一个高维任务，隐私预算暴增。而且，值得关注的是，不同的用户有不同的隐私需求，不同的用户上传的梯度对于联合模型的贡献比也有差异，因此本章将提出混合差分隐私技术，构

造一个全新的可信第三方——混洗器，与本地差分隐私相结合，实现的方案能提高全局模型的精度，也保证在更低的隐私成本下达到相同的隐私预算。

在本章节中我们提出了一个在联邦学习中的安全混洗器，本地客户端使用本地差分隐私进行加密，然后所有参与者将梯度传到一个安全混洗器，安全混洗器打乱次序，通过子抽样来增加隐私放大效果，再采用梯度稀疏化的技术筛选对联合模型贡献较高的 top-k 梯度，发送给中央服务器进行聚合。安全混洗器作为一个可信第三方，独立于服务器并专门用于梯度的子采样、混洗、稀疏化。这个模型通过子采样和混洗两者的结合达到隐私放大效应，从而提高了整体联邦学习模型的精度。当本地差分隐私添加更多的噪音时，对于同样的中央服务器能达到相同水平的隐私预算。

我们将在本章节详细的描述该框架中各个模块的设计和实现过程。

4.2 安全混洗模型

如图4.1所示，该框架主要由本地客户端、混洗器和中央服务器 3 部分组成：

- 本地客户端：基于第三章的本地自适应差分隐私方案，在模型训练的梯度下降算法中对梯度进行自适应的扰动，得到满足 $(\epsilon_c + \epsilon_l)$ 差分隐私的梯度。
- 混洗器：一个半诚信的第三方。首先动态采样本地客户端上传的梯度，然后借助现有的安全混洗协议在对数据一无所知的情况下，对子采样后的梯度完成安全的混洗操作，通过隐私放大效应满足 ϵ_0 -LDP，达到客户端匿名机制，最后将混洗后的结果发送至中央服务器。
- 中央服务器：一个诚实但好奇的第三方。服务器接受混洗器上传的梯度并进行聚合，然后更新全局模型。

假设现在有 m 个本地客户端，每个客户端表示为 $i \in [m]$ ，有本地数据集 $\mathcal{D}_i = \{d_{i1}, \dots, d_{ir}\} \in \mathbb{S}^r$ ，由 r 个数据集合构成。 $F_i(\theta)$ 表示在客户端 i 的本地数据集 \mathcal{D}_i 上进行训练，对于模型梯度 $\theta \in \mathbb{R}^d$ 进行衡量的损失函数，其中 $F_i(\theta) =$

$\frac{1}{r} \sum_{j=1}^r f(\theta; d_{ij})$, $f(\theta; \cdot) : \mathcal{C} \rightarrow \mathbb{R}$ 是凸函数。中央服务器的目标是找到一个最佳的模型参数向量 $\theta^* \in \mathcal{C}$ 使得损失函数 $\min_{\theta \in \mathcal{C}} (F(\theta) = \frac{1}{m} \sum_{i=1}^m F_i(\theta))$ 最小, 其中隐私性满足单个客户端的隐私预算, 也就是满足 ϵ_0 -LDP。在3中, 首先我们从 m 个客户端中随机挑选 k 个客户端, 表示为集合 \mathcal{U}_t , 其中 $k \leq m$ 。每个客户端 $i \in \mathcal{U}_t$ 从本地数据集中抽样 \mathcal{S}_{it} 个样本训练模型, 计算梯度 $\nabla_{\theta_t} f(\theta_t; d_{ij})$ 。第 i 个客户端采用基于第三章的自适应本地差分隐私方案, 添加噪声、裁剪梯度, 然后将梯度发送给混淆器。混淆器对收到的梯度进行拆分混淆, 然后发送给中央服务器。最后, 中央服务器对混淆后的梯度进行聚合求均值, 更新全局模型。

Algorithm 3 联邦学习中的安全模型算法: $\mathcal{A}_{\text{csdp}}$

```

1: 输入: 数据集  $\mathcal{D} = \bigcup_{i \in [m]} \mathcal{D}_i$ ,  $\mathcal{D}_i = \{d_{i1}, \dots, d_{ir}\}$ , loss function  $F(\theta) = \frac{1}{mr} \sum_{i=1}^m \sum_{j=1}^r f(\theta; d_{ij})$ , 本地差分隐私预算  $\epsilon_0$ , 梯度范数阈值  $C$ , 模型学习率  $\eta_t$ 
2: 初始化:  $\theta_0 \in \mathcal{C}$ 
3: for  $t \in [T]$  do
4:   客户端采样: 混淆器从  $k$  个客户端中随机采样  $i \in \mathcal{U}_t$  个客户端
5:   for 客户端  $i \in \mathcal{U}_t$  do
6:     梯度选择: 客户端 i 从  $s$  个样本空间中随机采样  $\mathcal{S}_{it}$  个梯度
7:     for 样本  $j \in \mathcal{S}_{it}$  do
8:        $\mathbf{g}_t(d_{ij}) \leftarrow \nabla_{\theta_t} f(\theta_t; d_{ij})$ 
9:        $\mathbf{g}_t(d_{ij}) \leftarrow \mathbf{g}_t(d_{ij}) / \max \left\{ 1, \frac{\|\mathbf{g}_t(d_{ij})\|_p}{C} \right\}^3$ 
10:       $\mathbf{q}_t(d_{ij}) \leftarrow \mathcal{R}_p(\tilde{\mathbf{g}}_t(d_{ij}))$ 
11:    end for
12:    客户端 i 将  $\{\mathbf{q}_t(d_{ij})\}_{j \in \mathcal{S}_{it}}$  发送给混淆器
13:  end for
14:  混淆器: 混淆器对于  $\{\mathbf{q}_t(d_{ij}) : i \in \mathcal{U}_t, j \in \mathcal{S}_{it}\}$  中的权重进行拆分混淆, 然后上传给中央服务器
15:  中央服务器聚合梯度:  $\bar{\mathbf{g}}_t \leftarrow \frac{1}{ks} \sum_{i \in \mathcal{U}_t, j \in \mathcal{S}_{it}} \mathbf{q}_t(d_{ij})$ 
16:  梯度下降:  $\theta_{t+1} \leftarrow \prod_{\mathcal{C}} (\theta_t - \eta_t \bar{\mathbf{g}}_t)$ 
17: end for
18: 输出: 最终全局模型参数  $\theta_T$ 

```

第三章我们提出了本地自适应差分隐私的方案, 中心化和本地化最大的区别是在客户端进行扰动还是在中央服务器进行扰动。本地化差分隐私是用户的数据在自己的本地化扰动后, 将扰动的值发送给聚合器, 聚合器收集大量的数据后再反推频数或者均值。而对于中心化差分隐私, 我们是信任聚合器的, 因为我们自

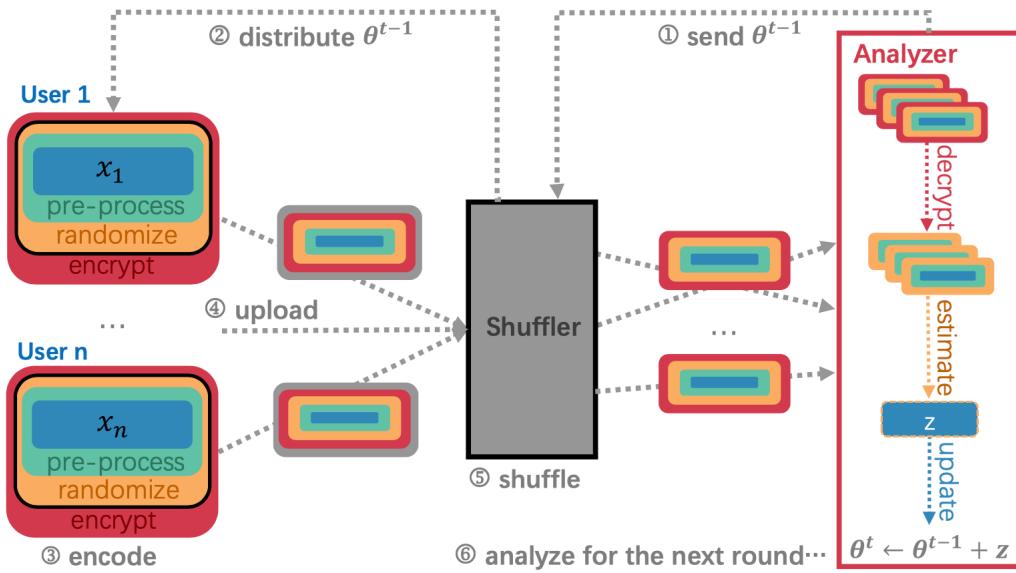


图 4.1: 联邦学习中的安全模型框架

己真实的数据直接发送给聚合器，聚合器收集大量的数据再扰动，再返回频数/均值等。

4.2.1 客户端抽样

假设在空间 \mathcal{U} 中我们有一个数据集 $\mathcal{D}' = \{U_1, \dots, U_{r_1}\} \in \mathcal{U}^{r_1}$ ，其中包含 r_1 个样本元素。如定义4.2.1所示，本文定义一个子采样程序：首先采样一个客户端数据集 $\mathcal{D}' \in \mathcal{U}^{r_1}$ ，再从中采样一个子集作为客户端的本地训练数据。

定义 4.2.1 (子采样). 定义一个抽样程序 $\text{samp}_{r_1, r_2} : \mathcal{U}^{r_1} \rightarrow \mathcal{U}^{r_2}$ ，其中 $r_2 \leq r_1$ ：从输入的数据集 $\mathcal{D}' \in \mathcal{U}^{r_1}$ 中以随机概率抽选一个子数据集 \mathcal{D}'' ，数据集 \mathcal{D}' 中的每个元素在数据集 \mathcal{D}'' 中出现的概率为 $q = \frac{r_2}{r_1}$ 。

4.2.2 混洗器

先前的研究工作 (H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. arXiv preprint arXiv:1710.06963, 2017.) 表明，在联邦学习模型中，假如在某个时间段数据是被适

当的匿名化，并将数据之间的耦合信息拆分后，模型整体的隐私保障可以得到极大的改善。在第三章中的隐私保护方案是基于本地客户端训练数据的，而面对恶意的中央服务器甚至是恶意的第三方攻击者时，无法保障每个客户端的隐私。

因此在本章中，我们针对客户端上传的梯度，进行参数的拆分混淆，通过在 LDP-FL 的本地模型更新上添加匿名性，打破从中央服务器接收的数据与特定客户端之间的联系，并在每次迭代中从同一客户端发送的梯度更新中将信息解耦。

客户端的匿名性可以通过现有的多种机制来实现，这取决于中央服务器在特定场景下如何跟踪客户端。作为一个典型的保护隐私的最佳做法，每个客户对服务器有一定程度的匿名性，以使客户的个人身份识别与他们的权重更新脱钩。例如，如果服务器通过 IP 地址追踪客户，每个客户可以通过使用网络代理、VPN 服务 [Belesi, 2016]、公共 WiFi 接入和 Tor[Dingledine 等人, 2004] 来采用一个无法追踪的 IP 地址。再比如，如果服务器通过软件生成的元数据（如 ID）来追踪客户，每个客户可以在向服务器发送元数据之前将其随机化。

但是，我们认为，客户端的匿名性不足以防止侧信道链接攻击。例如，如果客户端在每次迭代中同时上传了大量的权重更新，中央服务器仍然可以将它们连接在一起。因此，我们设计了混淆器，以打破来自相同客户的模型权重更新之间的联系，并将其放置于客户端上传梯度更新至中央服务器之间，使中央服务器更难结合多个客户端的同步更新来推断任何客户的更多信息。

如下图所示，我们的混淆器通过以下步骤对客户端上传的梯度参数进行混淆，然后上传给中央服务器：

- 权重分割：每个客户端都对其本地模型的权重进行分割，但给每个分割后的元素贴上一个 id，以表明其在网络结构中的权重位置。
- 权重混淆：对于所有客户端分割后的权重采用随机扰动机制进行混淆。

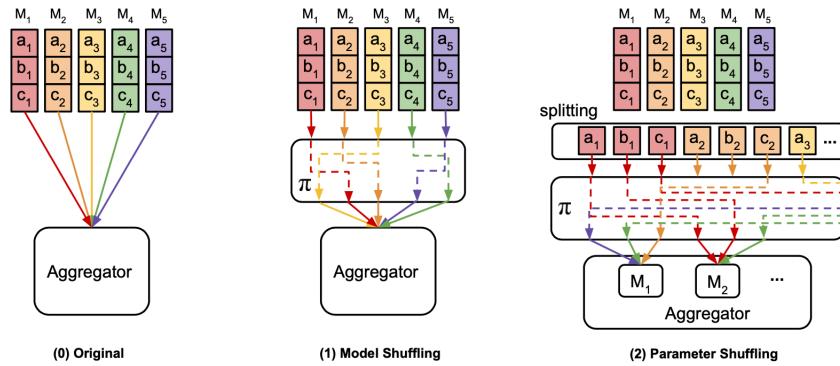


图 4.2: 联邦学习安全模型中执行参数拆分混淆的混淆器

Algorithm 4 混淆器中的拆分混淆算法

- 1: **Input:** 本地客户端添加自适应扰动后的权重 W_{l+1}^s
- 2: 对权重 W_{l+1}^s 进行分割, 给每个元素分配 id
- 3: for 每个元素 W_{l+1}^s do
- 4: 用一个唯一的 id 标记元素的位置
- 5: 随机采样 top-k 的权重
- 6: 上传给中央服务器

4.3 隐私放大效应

隐私放大 (privacy amplification) 是本章所提出的安全框架中混淆器对隐私效果增强的理论分析, 基于该理论, 可将现有的本地化差分隐私方法直接应用在安全框架上。由第三章的隐私性证明可知, 用户在本地端通过自适应扰动后的数据满足 $\varepsilon_1 - \text{LDP}$, 将参数上传至混淆器进行拆分混淆后, 所获取的数据满足 $\varepsilon_c - \text{DP}$ 。从 ε_1 到 ε_c 的转变可通过隐私放大约论证明。 ε_1 对应于较大的数值, 表示较低的隐私性; ε_c 对应于较小的数值, 表示较高的隐私性。因此经过混淆器后, 隐私性得到了增强, 具体的方案设计和证明会在下文详细讲述。

差分隐私的主要方法就是扰动 (perturbation) 和采样 (sampling)。差分隐私的扰动方案, 就是对输入数据、中间数据或者输出数据进行扰动, 加入噪音, 使其满足 ϵ_c -差分隐私。对于输入数据扰动的典型方案就是随机响应 (Randomized Response), 对于输出数据扰动的典型方案就是拉普拉斯算法 (Laplace algorithm)。中间数据可以看做前面一个子阶段的输出, 也可以看做是后面子阶段的输入, 因此可以灵活选

择输入或者输出扰动的算法。随机响应算法 (Randomized Response) 是最典型的本地化差分隐私算法。最早于 1965 年被 Warner 提出, 用于对隐私保护。随机响应技术主要包括两个步骤: 扰动性统计和校正。

为了获取更好的隐私放大的效果, 前人针对具体的差分隐私机制提出了更精确的隐私放大定理。其中基于随机响应机制提出的隐私放大定理是应用比较广泛的。

假设有 n 个可信用户参与计算, 每个用户拥有 1 个值 x_i , 其对应的取值范围为 $\{v_1, v_2, \dots, v_k\}$. 当用户发布该值时, 用户可能以 $1 - \lambda$ 的概率发布真实值, 以 λ 的概率从 $\{v_1, v_2, \dots, v_k\}$ 中随机输出 1 个值。混淆器接收这些数据后, 打乱值的顺序。令中央服务器收集到的每个值 $v_j (j \in \{1, 2, \dots, k\})$ 的数量为 n'_j , 通过式 4.1, 中央服务器可以估计每个值 j 的真实数量, 表示为 n_j 。

$$n_j = n'_j / (1 - \lambda) - n\lambda / (k(1 - \lambda)) \quad (4.1)$$

假设当 $k=2$ 时, 上述机制则是经典的随机扰动机制, 如果 $v1=“yes”, v2=“no”$, 则该机制解答的是 yes/no 的问题。而当 $k>2$ 时, 称其为 k 值随机扰动。

4.4 隐私性证明

在算法 3 中, 每个本地客户端采用第三章的满足 $(\epsilon_c + \epsilon_l)$ 的自适应本地差分隐私算法, 由差分隐私的强组合性可保证算法 \mathcal{A}_{csdp} 在每次迭代中对每个样本 d_{ij} 都能保证 ϵ_0 的本地差异隐私。因此本节只需要分析采样和混淆操作的隐私性。

定理 4.4.1. 算法 3 是满足 (ϵ, δ) - 差分隐私的, 当对于任意 δ , $\delta > 0$, 并且有:

$$\epsilon = \mathcal{O} \left(\epsilon_0 \sqrt{\frac{qT \log(2qT/\delta) \log(2/\delta)}{n}} \right)$$

假设在联邦学习模型中, 需要迭代的次数为 $t \in [T]$ 。 $\mathcal{M}_t(\theta_t, \mathcal{D})$ 表示在时刻 t 对于数据集 \mathcal{D} 和模型参数为 θ_t 的差分隐私机制, θ_{t+1} 表示模型的输出。因此, 在数据集 $\mathcal{D} = \bigcup_{i=1}^m \mathcal{D}_i \in \mathfrak{S}^n$ 上的差分隐私机制定义如下:

$$\mathcal{M}_t(\theta_t; \mathcal{D}) = \mathcal{H}_{ks} \circ \text{samp}_{m,k}(\mathcal{G}_1, \dots, \mathcal{G}_m) \quad (4.2)$$

其中, $\mathcal{G}_i = \text{samp}_{r,s}(\mathcal{R}(\mathbf{x}_{i1}^t), \dots, \mathcal{R}(\mathbf{x}_{ir}^t))$ 并且 $\mathbf{x}_{ij}^t = \nabla_{\theta_t} f(\theta_t; d_{ij}), \forall i \in [m], j \in [r]$ 。 \mathcal{H}_{ks} 表示在 ks 个数据样本上进行混淆操作, $\text{samp}_{a,b}$ 表示从有 a 个元素的集合中随机抽样 b 个元素的操作。

接下来我们给出 \mathcal{M}_t 的隐私性证明:

假设客户端 $i \in [m]$ 的本地数据集为 $\mathcal{D}_i = \{d_{i1}, d_{i2}, \dots, d_{ir}\} \in \mathfrak{S}^r$, $\mathcal{D} = \bigcup_{i=1}^m \mathcal{D}_i$ 表示总体数据集。根据公式4.2, $\mathcal{Z}(\mathcal{D}^{(t)}) = \mathcal{H}_{ks}(\mathcal{R}(\mathbf{x}_1^t), \dots, \mathcal{R}(\mathbf{x}_{ks}^t))$ 表示在本地客户端进行本地差分隐私后输出的 ks 个权重集合上进行混淆后的权重。任取 $\tilde{\delta} > 0$, 当 $\epsilon_0 \leq \frac{\log(ks/\log(1/\tilde{\delta}))}{2}$ 时, 算法 \mathcal{Z} 满足 $(\tilde{\epsilon}, \tilde{\delta}) - \text{DP}$ 差分隐私, 可得:

$$\tilde{\epsilon} = \mathcal{O} \left(\min \{ \epsilon_0, 1 \} e^{\epsilon_0} \sqrt{\frac{\log(1/\tilde{\delta})}{ks}} \right) \quad (4.3)$$

当 $\epsilon_0 = \mathcal{O}(1)$ 时, 有 $\tilde{\epsilon} = \mathcal{O} \left(\epsilon_0 \sqrt{\frac{\log(1/\tilde{\delta})}{ks}} \right)$ 。

令 $\mathcal{T} \subseteq \{1, \dots, m\}$ 表示在时刻 t 选取的 k 个客户端。对于 $i \in \mathcal{T}$, $\mathcal{T}_i \subseteq \{1, \dots, r\}$ 表示在时刻 t 客户端 i 所抽样的 s 条数据样本。对于任意的 $\mathcal{T} \in \binom{[m]}{k}$ 和 $\mathcal{T}_i \in \binom{[r]}{s}, i \in \mathcal{T}$, 有 $\bar{\mathcal{T}} = (\mathcal{T}, \mathcal{T}_i, i \in \mathcal{T})$, $\mathcal{D}^{\mathcal{T}_i} = \{d_j : j \in \mathcal{T}_i\}$ for $i \in \mathcal{T}$, and $\mathcal{D}^{\bar{\mathcal{T}}} = \{\mathcal{D}^{\mathcal{T}_i} : i \in \mathcal{T}\}$ 。 \mathcal{T} 和 $\mathcal{T}_i, i \in \mathcal{T}$ 为抽样产生的任意子集, 其中的随机性由客户端抽样和数据集抽样所决定。算法 \mathcal{M}_t 可以等价的表示为 $\mathcal{M}_t = \mathcal{Z}(\mathcal{D}^{\bar{\mathcal{T}}})$ 。

假设现有数据集: $\mathcal{D}' = (\mathcal{D}'_1) \bigcup (\bigcup_{i=2}^m \mathcal{D}_i) \in \mathfrak{S}^n$, 其中数据集 $\mathcal{D}'_1 = \{d'_{11}, d_{12}, \dots, d_{1r}\}$ 和 \mathcal{D}_1 为相邻数据集, 它们的第 d_{11} 条和第 d'_{11} 条数据样本不同。如果 \mathcal{M}_t 是满足 $(\bar{\epsilon}, \bar{\delta}) - \text{DP}$ 差分隐私的, 那么对于算法 \mathcal{M}_t 所选的任意子集 \mathcal{S} 都应该满足:

$$\Pr[\mathcal{M}_t(\mathcal{D}) \in \mathcal{S}] \leq e^{\bar{\epsilon}} \Pr[\mathcal{M}_t(\mathcal{D}') \in \mathcal{S}] + \bar{\delta} \quad (4.4)$$

$$\Pr[\mathcal{M}_t(\mathcal{D}') \in \mathcal{S}] \leq e^{\bar{\epsilon}} \Pr[\mathcal{M}_t(\mathcal{D}) \in \mathcal{S}] + \bar{\delta} \quad (4.5)$$

由于式4.4和4.5是对称的，因此只需要证明其中一条。下文给出式4.4的证明：

令 $q = \frac{ks}{mr}$ ，我们给出条件概率的定义：

$$\begin{aligned} A_{11} &= \Pr \left[\mathcal{Z}(\mathcal{D}^{\bar{T}}) \in \mathcal{S} \mid 1 \in \mathcal{T} \text{ and } 1 \in \mathcal{T}_1 \right] \\ A'_{11} &= \Pr \left[\mathcal{Z}(\mathcal{D}'^{\bar{T}}) \in \mathcal{S} \mid 1 \in \mathcal{T} \text{ and } 1 \in \mathcal{T}_1 \right] \\ A_{10} &= \Pr \left[\mathcal{Z}(\mathcal{D}^{\bar{T}}) \in \mathcal{S} \mid 1 \in \mathcal{T} \text{ and } 1 \notin \mathcal{T}_1 \right] = \Pr \left[\mathcal{Z}(\mathcal{D}'^{\bar{T}}) \in \mathcal{S} \mid 1 \in \mathcal{T} \text{ and } 1 \notin \mathcal{T}_1 \right] \\ A_0 &= \Pr \left[\mathcal{Z}(\mathcal{D}^{\bar{T}}) \in \mathcal{S} \mid 1 \notin \mathcal{T} \right] = \Pr \left[\mathcal{Z}(\mathcal{D}'^{\bar{T}}) \in \mathcal{S} \mid 1 \notin \mathcal{T} \right] \end{aligned} \tag{4.6}$$

令 $q_1 = \frac{k}{m}$, $q_2 = \frac{s}{r}$, 那么 $q = q_1 q_2$, 然后可以得到：

$$\Pr[\mathcal{M}_t(\mathcal{D}) \in \mathcal{S}] = qA_{11} + q_1(1 - q_2)A_{10} + (1 - q_1)A_0 \tag{4.7}$$

$$\Pr[\mathcal{M}_t(\mathcal{D}') \in \mathcal{S}] = qA'_{11} + q_1(1 - q_2)A_{10} + (1 - q_1)A_0 \tag{4.8}$$

因此，我们可以得到：

$$A_{11} \leq e^{\tilde{\epsilon}} A'_{11} + \tilde{\delta} \tag{4.9}$$

$$A_{11} \leq e^{\tilde{\epsilon}} A_{10} + \tilde{\delta} \tag{4.10}$$

式4.8成立，因此 \mathcal{M}_t 是满足 $(\bar{\epsilon}, \bar{\delta})$ -差分隐私的。

4.5 模型收敛性分析

回顾第二章的基础知识，在随机梯度下降算法的每次迭代中，中央服务器将当前的参数向量发送给所有本地客户端，客户端收到后在本地数据集上进行模型训练，计算随机梯度并上传给中央服务器，然后中央服务器计算收到的梯度的平均值/平均数并更新参数向量。因此在本节中，我们分析采用采样和混洗算法后模型的收敛性。

在算法3中，在每一轮迭代过程中，中央服务器聚合上传的 ks 个加噪后的梯度，如算法3的第 15 行所示，中央服务器进行聚合后得到结果： $\bar{\mathbf{g}}_t \leftarrow \frac{1}{ks} \sum_{i \in \mathcal{U}_t, j \in \mathcal{S}_{it}} \mathbf{q}_t(d_{ij})$,

然后通过随机梯度下降算法更新全局模型参数: $\theta_{t+1} \leftarrow \prod_C (\theta_t - \eta_t \bar{\mathbf{g}}_t)$ 。其中, $\mathbf{q}_t(d_{ij}) = \mathcal{R}_p(\nabla_{\theta_t} f(\theta_t; d_{ij}))$ 。

既然随机机制 \mathcal{R}_p 是无偏的, 那么平均梯度 $\bar{\mathbf{g}}_t$ 也是无偏的, 也就是说, 我们有 $\mathbb{E}[\bar{\mathbf{g}}_t] = \nabla_{\theta_t} F(\theta_t)$, 其中期望是相对于客户端和数据点的随机抽样以及机制 \mathcal{R}_p 的随机性而言的。

令 $F(\theta)$ 为凸函数, 考虑这样一个随机梯度下降算法: $\theta_{t+1} \leftarrow \prod_C (\theta_t - \eta_t \mathbf{g}_t)$, \mathbf{g}_t 满足 $\mathbb{E}[\mathbf{g}_t] = \nabla_{\theta_t} F(\theta_t)$ 并且 $\mathbb{E}\|\mathbf{g}_t\|_2^2 \leq G^2$ 。当确定 $\eta_t = \frac{D}{G\sqrt{t}}$, 可以得到:

$$\mathbb{E}[F(\theta_T)] - F(\theta^*) \leq 2DG \frac{2 + \log(T)}{\sqrt{T}} = \mathcal{O}\left(DG \frac{\log(T)}{\sqrt{T}}\right) \quad (4.11)$$

由文献的证明可知, 算法3的输出 θ_T 满足:

$$\mathbb{E}[F(\theta_T)] - F(\theta^*) \leq \mathcal{O}\left(\frac{LD \log(T) \max\left\{d^{\frac{1}{2}-\frac{1}{p}}, 1\right\}}{\sqrt{T}} \left(1 + \sqrt{\frac{cd}{qn}} \left(\frac{e^{\epsilon_0} + 1}{e^{\epsilon_0} - 1}\right)\right)\right) \quad (4.12)$$

其中, 存在 $\sqrt{1 + \frac{cd}{qn}} \left(\frac{e^{\epsilon_0} + 1}{e^{\epsilon_0} - 1}\right)^2 \leq \left(1 + \sqrt{\frac{cd}{qn}} \left(\frac{e^{\epsilon_0} + 1}{e^{\epsilon_0} - 1}\right)\right)$ 。

当 $\sqrt{\frac{cd}{qn}} \left(\frac{e^{\epsilon_0} + 1}{e^{\epsilon_0} - 1}\right) \leq \mathcal{O}(1)$ 时, 我们恢复了没有隐私性的虚构 SGD 的收敛率。而当 $\sqrt{\frac{cd}{qn}} \left(\frac{e^{\epsilon_0} + 1}{e^{\epsilon_0} - 1}\right) \geq \Omega(1)$ 时, 可以推导出:

$$\mathbb{E}[F(\theta_T)] - F(\theta^*) \leq \mathcal{O}\left(\frac{LD \log(T) \max\left\{d^{\frac{1}{2}-\frac{1}{p}}, 1\right\}}{\sqrt{T}} \sqrt{\frac{cd}{qn}} \left(\frac{e^{\epsilon_0} + 1}{e^{\epsilon_0} - 1}\right)\right) \quad (4.13)$$

如果我们在算法3中设置学习率为 $\eta_t = \frac{D}{G\sqrt{t}}$, 其中 $G^2 = L^2 \max\left\{d^{1-\frac{2}{p}}, 1\right\} \left(1 + \frac{cd}{qn} \left(\frac{e^{\epsilon_0} + 1}{e^{\epsilon_0} - 1}\right)^2\right)$ 。那么:

$$\mathbb{E}[F(\theta_T)] - F(\theta^*) \leq \mathcal{O}\left(\frac{LD \log(T) \max\left\{d^{\frac{1}{2}-\frac{1}{p}}, 1\right\}}{\sqrt{T}} \sqrt{\frac{cd}{qn}} \left(\frac{e^{\epsilon_0} + 1}{e^{\epsilon_0} - 1}\right)\right) \quad (4.14)$$

其中, 当 $p \in \{1, \infty\}$ 时, $c = 4$ 否则 $c = 14$ 。

4.6 本章总结

本章节我们针对联邦学习模型的整体框架进行了隐私性改进, 提出了安全混洗模型, 在本地客户端和中央服务器之间加设混洗器, 通过对本地客户端进行随

机抽样，将上传的梯度进行拆分混洗，增加隐私放大效果。然后筛选对联合模型贡献较高的 top-k 梯度，发送给中央服务器进行聚合。并对方案进行了隐私性证明，表明此安全混洗算法可以保证 $(\bar{\epsilon}, \bar{\delta})$ 的差分隐私，然后对此方案在中央服务器上的随机梯度下降算法进行了收敛性的分析，证明在凸函数上，梯度 \mathbf{g}_t 满足 $\mathbb{E}[\mathbf{g}_t] = \nabla_{\theta_t} F(\theta_t)$ 时模型能达到全局收敛。因此，本章的方案能在保持模型可用性的前提下，减少隐私预算。

第五章 实验与评估

之前的章节中，我们描述了联邦学习的本地自适应差分隐私和安全聚合框架的设计和实现过程。在本节的内容中，我们选取了一些基准的数据集在该验证框架上进行实验评估。本实验是关于联邦学习系统的隐私保护方案。本章的实验主要针联邦深度学习系统训练样本的攻击模型，保护联邦学习系统中参与者的共享梯度信息，避免梯度参数泄露隐私和恶意服务器获取客户端的信息，进而保护参与者本地训练样本。在实验室环境下，通过多 GPU 虚拟化设置模拟分布式联邦学习系统，并且将差分隐私保护方案和混淆器配置在模拟分布式联邦学习系统中，同时在系统中设置攻击模型，评估满足保护算法的系统学习准确率、攻击模型成功率以及隐私保护预算。

5.1 基准数据集介绍

我们选用了以下三个数据集评估了我们的联邦学习隐私保护框架：

- (1) 手写体数字识别数据集 (MNIST) 是用于分类任务的经典数据集，来源于美国国家标准与技术研究所。总共包含了 70000 个手写数字图像，每个图像的尺寸为 28×28 像素，每个像素点用灰度值表示，灰度值范围为 0 到 255，图像分为 10 种类别，分别代表 0-9。
- (2) FASHION-MNIST 数据集包含了 70000 个不同商品的正面灰度图像，与 MNIST 数据集一样，每个图像的尺寸为 28×28 像素，灰度值范围同样为 0 到 255。所有的图像分为 10 种类别，如：T 恤，牛仔裤，裙子等。虽然数据集格式与 MNIST 相同，但由于图像内容的差别，使得有些模型或者算法在 MNIST 和

FASHION-MNIST 的表现会有很大不同。因此对于分类任务，我们在这两个数据集上都进行了实验作为对比。

(3) CIFAR-10 数据集由 10 类 32x32 的彩色图片组成，一共包含 60000 张图片，每一类包含 6000 张图片。其中 50000 张图片作为训练集，10000 张图片作为测试集。CIFAR-10 数据集被划分成了 5 个训练的 batch 和 1 个测试的 batch，每个 batch 均包含 10000 张图片。测试集 batch 的图片是从每个类别中随机挑选的 1000 张图片组成的，训练集 batch 以随机的顺序包含剩下的 50000 张图片。不过一些训练集 batch 可能出现包含某一类图片比其他类的图片数量多的情况。训练集 batch 包含来自每一类的 5000 张图片，一共 50000 张训练图片。

5.2 实验环境与配置

本文中的所有的实验是在 Windows 10 系统下，使用 CPU Inter(R) Core i3-7100 @ 3.90GHz，GPU 的型号是 NVIDIA GeForce GTX1050，内存 8GB。在实验中使用了 Facebook 公司的 Pythorch 框架对神经网络模型进行编写，相比于 TensorFlow，PyTorch 网络定义方便，更有利于研究小规模项目快速做出原型。其对于并行化数据的支持更有利于分布式联邦系统的实验等）。在对样本数据预处理的部分，我们使用了 Pandas，Numpy 等第三方库。

5.3 实验设计

5.3.1 联邦学习模型

实验同样设置 30 名联邦学习的参与者，论文研究在分布式联邦系统中添加噪声达到差分隐私并使得整体模型的精度维持较优。首先考虑了如何设置超参数可以更好的让全局模型能够得到更好的训练。分布式联邦学习梯度选择的准则是选择差值变化最大的，调整梯度上传阈值，将上传比例 θ_u 设置为 0.1，将从参数服务器下载的全局参数的比例 θ_d 设置为 1。

接下来，在联邦系统中实施本文所提出隐私保护方案。实验在设置每个参与者在训练分布式联邦系统时每次迭代的总隐私预算为 ϵ ，将隐私预算分成 c 个部分，其中 c 是选择每次迭代满足层间前馈传播算法的梯度总数，即 $c = \theta_u |\Delta w|$ 。我们使用拉普拉斯机制根据分配的隐私预算在选择梯度过程中添加噪声。添加的噪声取决于隐私预算中所有参数的灵敏度 Δf 都相同，但具体情况下，不同的参数可能具有不同的灵敏度。

在分布式联邦学习模型中，实验评估了不同 $\frac{\theta_u}{\theta_c}$ 值的情况下 (θ_u 为选择梯度阈值的参数)，使用论文方案满足差分隐私的分布式联邦系统的全局模型准确率，并且将参数保护后系统精度与未保护的模型精度相比较。虽然与集中式深度学习有差距，由于参与者较多，而且当参与者共享很大一部分梯度时，模型的准确性要优于独立训练的准确性。但是，模型更好的准确性的效果是较低的隐私保护（即更大的 ϵ 值）带来的，更强的隐私保护效果（更小的 ϵ 值）会导致较低的模型精度。

5.3.2 神经网络模型

Shokri^[51] 在论文中公开提供了他们的源代码，实现了一个完整的分布式联邦学习系统。我们将攻击模型部署在该联邦系统中，并且使用其中的卷积神经网络 (CNN) 架构，如图5.1。在 CNN 架构中，网络的前端是卷积层和池化层，后端则是使用反向传播算法的全连接层。前端的网络结构是在一个 nn.SpatialConvolution 卷积层连接激活函数 TanH，后面再接一个 nn.SpatialMaxPooling 最大池化层。之后再连接卷积层、TanH 激活函数和池化层单元。后端的网络架构则是 nn.Linear 线性层加上 TanH 激活函数和分类输出层。CNN 网络结构中的参数个数计算如下：

$$32 \times 5 \times 5 + 32 + 64 \times 32 \times 5 \times 5 + 64 + 200 \times 256 + 200 + 10 \times 200 + 10 = 105506$$

CNN 网络中的损失函数为 nn.CrossEntropyLoss。该函数是将 nn.LogSoftmax 和 nn.NLLLoss 结合起来使用，使用 Softmax 函数和交叉熵损失函数，评估分类任务中的损失，同时可以更加方便地计算反向传播算法。在选择梯度上传的全连接层与传输协议中，部分超参数选择如下：选择参数比例 $\theta_u = 0.01$ ，全局参数 θ_d 下载比例

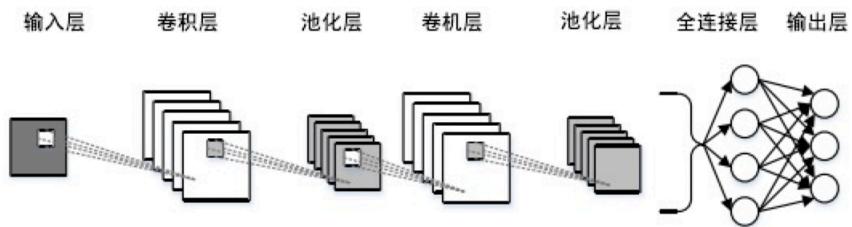


图 5.1: 卷积神经网络结构图

为 1。为了允许在学习中更多的随机性, 将学习率设置为 $\alpha = 1 \times 10^{-2}$, 学习速率衰减值为 1×10^{-7} 。参与者迭代过程使用表 CNN 网络训练本地数据集, 攻击者使用基于 CNN 网络的 DCGAN 算法与成员推理攻击的白盒算法。实验在这样的参数设置下搭建一个包含 29 个正常参与者和 1 个攻击者的分布式联邦学习系统, 30 个参与者 (包含攻击者) 都与中央参数服务器进行连接。

我们将与直接增加噪声的情况以及不加噪声的情况进行对比。实验中使用 20000 条数据作为训练数据集, 每一个客户端拥有 10 个样本的数据, 剩下的样本则作为测试数据集, 每种情况分别重复做 5 次并取平均值。Adult 实验参数为 $T = 200$, 步长 $\alpha = 1e - 4$, 衰减系数 $\gamma = 0.99$ 。

5.4 自适应扰动方案的实验评估

对于自适应扰动框架的实验, 评估指标主要有隐私预算参数 ϵ , 模型预测准确率。梯度自适应加噪的方法对于模型准确度的影响比传统的梯度固定加噪方法更小, 在相同的隐私预算约束下, 模型准确性有 3% 左右的提升。首先, 对于 MNIST 数据集, 在无差分隐私机制的原始模型上进行训练得到基准测试准确率约为 97%, 证明模型结果对于 MNIST 数据集是有效的。其次, 我们分别使用梯度固定加躁方法和梯度自适应加躁方法进行实验, 实验结果如下。

(1) 使用梯度固定加噪方法: 使用所有 D_{pub} 计算所得的平均梯度 0.001 作为固定的梯度裁剪阈值进行梯度裁剪, 每轮噪声添加的训练批次大小 L 为 600 个样本, 因此每个样本的采样率为 $q = \frac{L}{N} = \frac{600}{60000} = 0.01$, 噪声量采用中等噪声 $\sigma = 5$, 隐私

参数为 $\delta = 10^{-5}$ 。如图5.2所示，隐私预算参数 ϵ 为研究变量。随着隐私预算参数越大，差分隐私提供的隐私保护强度越小，噪声量越少，模型的准确率越高。

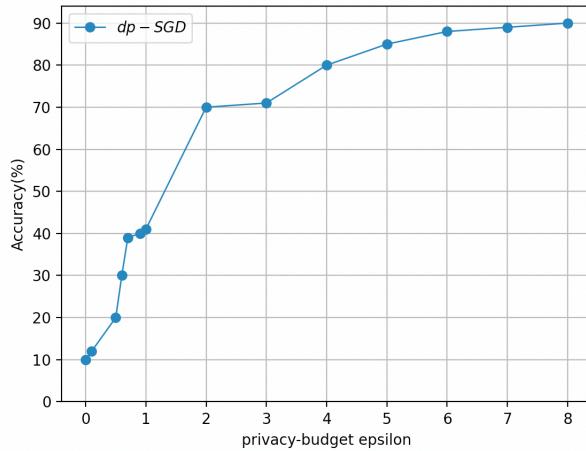


图 5.2: 梯度固定加噪方法下模型准确率随隐私预算变化情况

在不同的隐私预算下，随着训练轮数 epoch 的增加，模型的准确率对比如图??所示， ϵ 越大，最终模型预测准确率越高。这时候隐私预算参数越大，差分隐私提供的隐私保护强度越小，噪声量越少，符合理论原理。当隐私预算 $\frac{\epsilon}{c} \geq 5$ 后，隐私预算参数对于模型准确率影响趋于平稳，综合来看，当 $c \geq 5$ 后，部署了差分隐私机制的模型准确率可达 90% 左右，较原始模型存在约 7% 的准确率差距。

(2) 使用梯度自适应扰动方法：图5.3描绘了目标损失值和准确率随 γ 变化的趋势，当 γ 值越小，则越为接近平均分配时的情况。我们可以从图中可以看到，自适应权重在准确率上最高可以提高 6% 以上，损失值可以降低 0.15 以上。我们可以发现，当 γ 值变小的时候，得到的损失也会相应变大，准确率也会相应变小，整体趋势是随着接近平均的情况效果会下降，这是因为我们根据收敛规律合理分配隐私预算，结果与我们的上文分析所相吻合；另一方面，而当 γ 过于大的时候，损失很大，准确率很小，整体表现很差，这是因为前期分配的隐私预算过少，导致刚开始的迭代的噪声过大，很难通过后面少量的迭代来弥补。

接着，我们比较了在不同隐私预算下的自适应干扰模型的准确性，隐私预算分别为 ($\epsilon_1 = 0.1, \epsilon_2 = 0.5, \epsilon_3 = 2.0, \epsilon_4 = 8.0$)。隐私预算 ϵ 越小，噪音就越大。我们

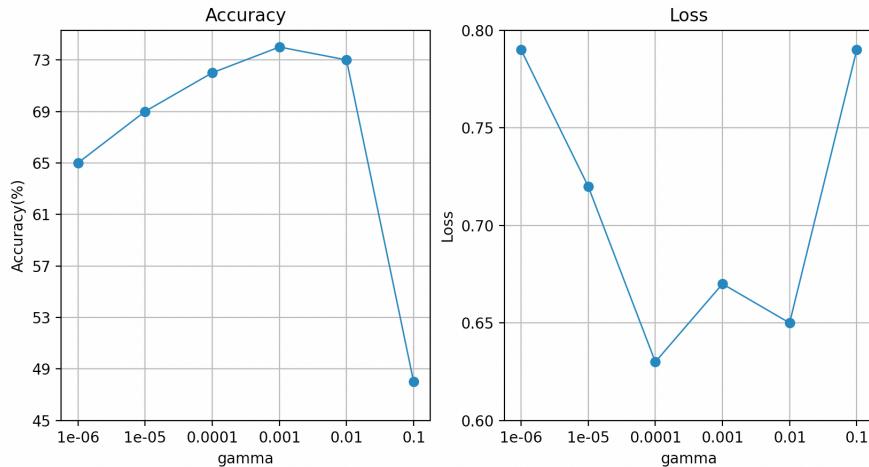


图 5.3: 梯度自适应扰动方法下模型精度、损失随隐私参数的变化趋势

还为每个隐私预算选择三种不同的参数取值 ((a): $f = 0.15, p = 0.85$, (b): $f = 0.10, p = 0.90$ (c): $f = 0.05, p = 0.95$)。可以肯定的是，设定的 ($f = 0.15, p = 0.85$) 可以保证系统的隐私水平。在实验中，隐私预算 ϵ 的值是 ϵ_c 、 ϵ_l 和 ϵ_c 的总和。我们将隐私预算的计算分为以下三个步骤：对于贡献的计算、线性转换中的计算和损失函数的计算，即： $\epsilon_c = \epsilon_l = \epsilon_f = \frac{\epsilon}{3}$ 。

正如图5.4所示，随着隐私预算 ϵ 的增加，我们系统的准确性保持稳定的增长趋势。随着调整因子范围的不断缩小，自适应干扰模型的准确率逐渐降低，但仍保持较高的水平。例如，当隐私预算 ϵ 设置为 8.0 时，在 $f=0.15$ 和 $p=0.85$ 的设置下，APFL 的准确率高达 97.34%，而在 $f=0.10$ 和 $p=0.90$ 的设置下，准确率为 96.57%，以及在 $f=0.05$ 和 $p=0.95$ 的设置下，准确率为 96.25%。

综上，自适应隐私预算分配可以根据一般问题的收敛规律，合理地分配隐私参数，从而提高模型表现，但参数 γ 需要小心选取，过大的 γ 值会导致训练的初始阶段噪声太大，从而影响模型的可用性。

我们还与近年来使用 DP 机制保护深度学习模型隐私的工作进行了比较，如 [1] 中的 DLPP 和 [18] 中的 DSSGD。在图5.5中，我们可以清楚地得到一个信息，即我们的工作即使在强隐私保证下 ($\epsilon=0.1$) 也表现良好。当调整因素设置为 $f = 0.15$ 和 $p = 0.85$ 时，模型的准确率在 200 个历时后达到 88.46%。此外，调整因素为 $f=0.05$ 和

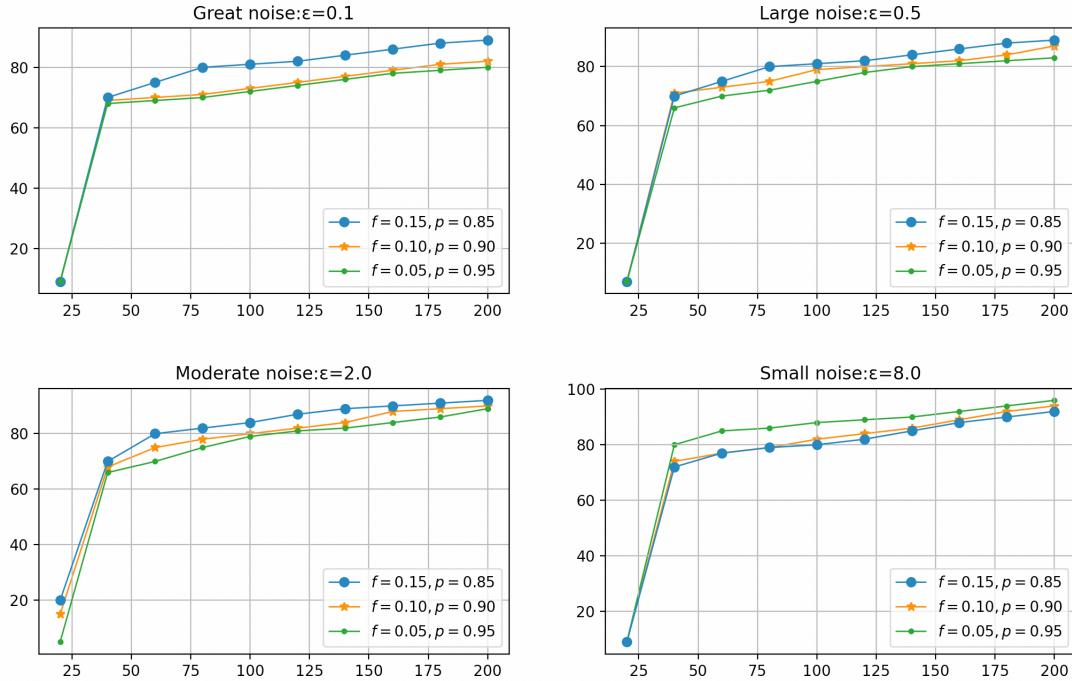


图 5.4: 不同隐私预算的自适应干扰模型的准确率

$p=0.95$, 自适应干扰模型的准确率为 86.79%。然而, 在相同的隐私预算下, DP-SGD 的准确性仅达到 79.63%, DLPP 模型的准确性低于 65.00%。

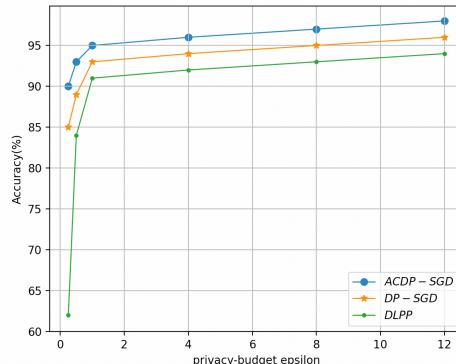


图 5.5: DP-SGD、DLPP、ACDP 在模型准确率和隐私预算上的对比

5.5 安全混洗算法的实验评估

我们在 MNIST、FMNIST 和 CIFAR 上评估所提出的安全聚合框架。为了评估参数: 模型迭代次数 f_r , 我们首先将客户端的数量固定为 500。

如图5.6所示，通过客户端采样机制和梯度的拆分混洗算法，我们的安全混洗模型(下文简称 SA-FL)能够以较低的隐私成本实现较高的准确性。在训练中增加客户数量 n 的同时，SA-FL 的表现与无噪声的联合学习一样接近。与 MNIST($n=100, \epsilon=1$)、FMNIST($n=200, \epsilon=5$) 相比，CIFAR-10($n=500, \epsilon=10$) 需要更多的客户端，这表明对于一个具有较大神经网络模型的更复杂的任务，它需要更多的本地数据和更多的客户端以获得对数据扰动的更好表现。

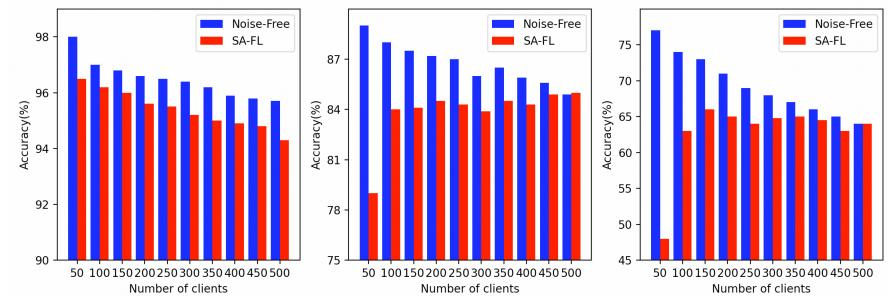


图 5.6: 安全混洗模型中参与混洗的本地客户端数量对联合模型精度的影响

由图5.7可以发现，当 f_r 太小的时候，并不影响在 MNIST 上的表现，但对 FASHION-MNIST 和 CIFAR-10 的表现影响很大。当 f_r 接近 1 时，安全聚合框架可以在 MNIST、FASHION-MNIST 和 CIFAR-10 上取得与无噪声结果几乎相同的性能。另一个重要的参数是中央参数聚合器和本地客户端之间的通信轮次。不难看出，随着通信次数的增加，我们可以通过所提出的模型在所有数据集上训练出更好的模型。然而，由于数据和任务的复杂性，CIFAR-10 需要更多的通信回合以获得更好的模型。

如图5.8(a-c) 中，SA-FL 在 $\epsilon=1$ 和 $n=100$ 的情况下可以达到 96.24% 的准确率，在 $\epsilon=4$ ， $n=200$ 的情况下可以达到 86.26% 的准确率，在 $\epsilon=10$ ， $n=500$ 的情况下，在 MNIST，FMNIST 和 CIFAR-10 上可以达到 61.4% 的准确率。我们的结果与之前的其他工作相比非常有竞争力。[Geyer 等人,2017] 首次将差分隐私应用于联邦学习，虽然他们使用了 100 个客户端，但在 MNIST 上，他们只能在 $(\epsilon, m) = (8, 11)$, $(8, 54)$ 和 $(8, 412)$ 的情况下达到 78%, 92% 和 96% 的准确率，保证了差异化的隐私，其中 (ϵ, m) 代表隐私预算和通信回合。[Bhowmick 等人, 2018] 首次在联合学习中

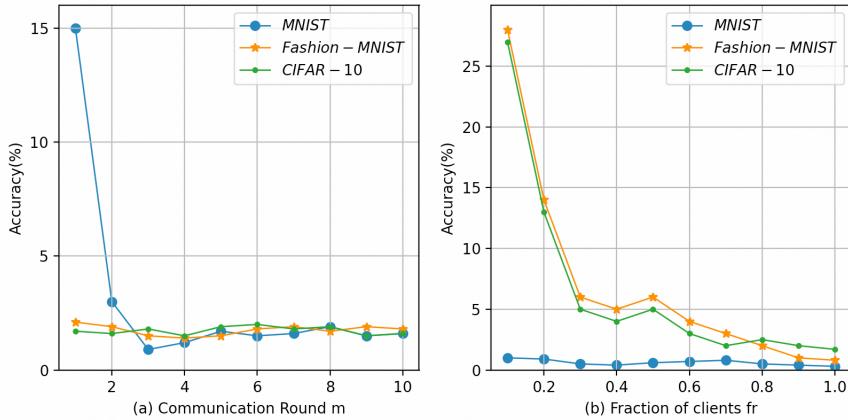


图 5.7: 安全混洗模型中通信轮数和客户端采样比对联合模型精度的影响

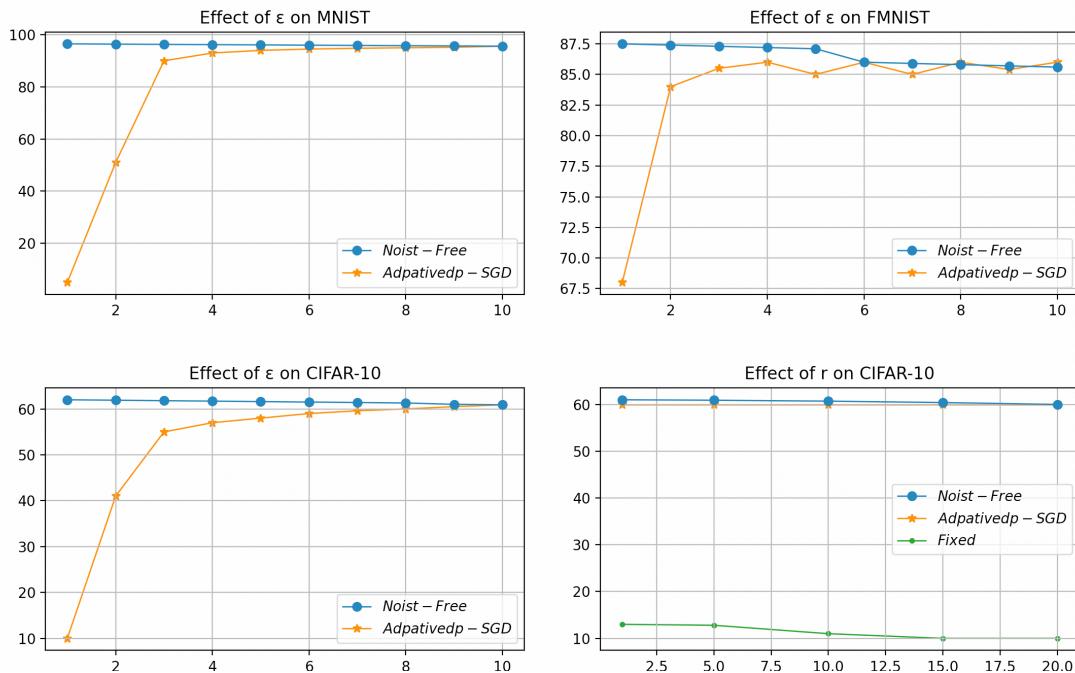


图 5.8: 训练参数对模型精度的影响

利用本地差分隐私。由于其机制的高变异性，它需要超过 200 轮的通信，并花费更多的隐私预算，即 MNIST ($\epsilon=500$) 和 CIFAR-10 ($\epsilon=5000$)。最近的工作 [Truex 等人, 2020] 将浓缩的局部差分隐私 (α -CLDP) 应用到联邦学习中，在 FMNIST 数据集上获得了 86.93% 的准确性。然而， α -CLDP 通过要求相对较大的隐私预算 $\epsilon = \alpha \cdot 2c \cdot 10\rho$ (例如， $\alpha = 1, c = 1, \rho = 10$) 来实现该性能，这导致了较弱的隐私保证。与以往的工作相比，我们的方法在客户端和云端之间需要更少的通信回合 (例如，MNIST 为 10，

FMNIST 和 CIFAR-10 为 15)，这使得整个解决方案在实际场景中更加实用。总的来说，SA-FL 在有效性、效率和维护成本方面都比之前的作品取得了更好的表现。

5.6 结果分析

为了验证自适应扰动算法对于模型训练的精度也能维持在较优的水平，我们进行了对比实验，使用自适应扰动算法和不使用这种方法在不同的隐私参数 ϵ 下的对比实验。由本章第三节对于自适应差分隐私方案的实验评估，我们可以看到自适应扰动算法基本上占有绝对的优势，尤其是在损失函数值，在隐私参数 $\epsilon=0.1$ 时，我们的方法不到 1，而传统的平均算法却在 100 左右。这么大的差距的原因在于，自适应扰动算法的权重分配使得聚合时个体信噪比不变，但整体的聚合结果的信噪比却提高了很多，因此当隐私参数 ϵ 很小，即噪声量很大的时候，表现越好。而当 ϵ 越大时，注入的噪声也就越小，自适应加躁方法的效果就没有噪声大的时候明显。

系统的额外开销主要来自服务器端的预训练过程，以及用户端在开始训练前对贡献的计算和扰动。我们使用 20 个历时来训练云服务器的初始化网络，这平均需要 68.22 秒。在独立和异步的训练过程之前，用户需要用层间依赖传播算法计算权重。这个过程只需要训练正向传播过程，而不需要计算反向传播过程中的梯度和损失惩罚。其平均时间为 4.35 毫秒。为了减轻隐私威胁，我们的解决方案是向权重、线性变换函数中的原始数据和损失函数的系数注入拉普拉斯噪声。向权重注入噪声的步骤可以与计算贡献同步进行，这需要额外的 2.67 毫秒时间。向线性变换中的原始数据和损失函数的系数注入自适应噪声的操作可以在训练前完成，每一个历时的计算都与扰动的权重相似。因此，在模型效率方面的提升是非常突出的。

从隐私成本和模型精度的总体上看，混淆差分隐私方法在各统计问题的结果可用性上都有着相比本地化差分隐私方法明显更优的结果。但从通信代价和计算代价的角度分析，安全混淆算法中混淆器的引入，一方面使得用户数据与用户所使用的编码器之间的关联性消失，使得分析器端的计算代价增大；另一方面促使

研究者们使用富含信息更多的多消息模式对数据编码，造成了分析器端的通信代价增大。如何兼顾数据的隐私性、可用性、算法的计算代价和通信代价是后续基于 SA-FL 框架构建的隐私保护方法需加以考量的部分。从各混淆差分隐私算法评估的结果看，随着的 $\frac{\epsilon}{c}$ 增大，各方法的数据可用性均会得到提高；而随着用户数据 n 的增加，基于本地化差分隐私方法设计的混淆差分隐私方法在计算误差上会有轻微的增加，其他大多数混淆差分隐私方法在计算误差上没有明显变化，甚至部分方法有着轻微的降低。总体上，基于多消息模式设计的混淆与用户数据相关的信息，有着相对较高的数据可用性，与前文的理论分析相一致。

5.7 本章小结

在本章中，我们选取了三个基准数据集对本文提出的自适应本地差分隐私和安全混淆框架进行了一系列的实验来测试其可行性，并且在联邦学习系统上也进行实验和研究。实验结果表明，我们的自适应本地差分隐私可以有效降低隐私预算，并且维持模型精度。安全混淆框架能通过客户端采样算法和梯度的拆分混淆算法，降低隐私保护预算，提高数据的可用性。

第六章 总结与展望

6.1 总结

随着深度学习的兴起，出现了越来越多新的模型和算法，能够更有效的解决各类问题。基于人工智能的产品也在各个领域迎来了一波新的发展热潮，给人民的生活带来了巨大的便利。然而用户在享受深度学习模型带来便利的同时，必须共享自己的数据，随着隐私泄露事件越来越多，数据的安全和隐私问题也逐步引起了人们的关注。

与此同时，各类智能设备也在不断发展，用户产生的数据也越来越多，智能设备的算力不断增强。用户不愿意向商业公司或商业机构提供个人隐私数据。分布式联邦学习系统解决分布式终端用户在本地更新模型的问题，联邦学习的目标是保障大数据共享信息时的数据安全、保护本地数据和个人隐私，在多计算节点之间高效的训练机器学习模型。

分布式联邦学习系统得到了广泛的研究和应用，成为传统集中式机器学习方法的一种改进方法。它不是将数据上传到中心服务器进行集中训练，而是参与者在本地进行模型训练并与参数服务器共享模型更新。参数服务器对来自多个参与者的权重进行聚合，并组合创建一个改进的全局模型，这有助于保障用户的数据隐私和降低通信成本。

本文主要研究针对分布式联邦学习系统的隐私安全问题。通过研究分布式联邦深度学习的系统漏洞，提出了一套分布式联邦系统中针对攻击的隐私安全方案对策。本文的主要工作和贡献如下：

- (1) 基于本地差分隐私的权重分配自适应干扰算法。在客户端本地训练的神经网

络模型中，通过分析前向传播算法，计算每个属性类对于模型输出的贡献比，然后，我们开发了一个自适应噪声添加的方案，根据贡献率注入不同隐私预算的噪声。与传统的注入噪声的方法相比，我们在相同的隐私保护程度下最大限度地提高了模型的准确性，减少噪声对模型输出结果的影响，提高模型精度。而且，本文也从本地差分隐私定义的角度，理论证明了提出的方法满足 \mathcal{E} -本地差分隐私。最后通过多组真实数据集以及合成数据集验证了本地自适应扰动机制的性能，证明了其在相同条件下要优于现有的同类方法。

- (2) 本文提出了 SA 安全混洗框架，混洗差分隐私摒弃了中心化差分隐私下对可信第三方的依赖，即无需任何可信第三方。对用户的原始数据进行统一的扰动处理，提高了隐私性；弥补了中心化差分隐私与本地化差分隐私在可用性上约 $O(n)$ 的间隙，在差分隐私的保证下实现了数据隐私度与可用性之间的更好平衡。

综上所述，本文的研究充分证明了所提出框架的有效性，可以极大的联邦学习模型的隐私性和可用性，从而进一步推进了联邦学习在安全领域的应用和发展。

6.2 展望

在可预见的未来，大规模、大数据、分布式的深度学习将得到快速发展。5G、边缘计算、物联网等技术也将迅速普及。人类将彻底步入人工智能时代。在此我将对我未来的研究做出几点展望：

- (1) 本文提出的自适应的差分隐私深度学习算法是一种基础算法，它可以令学习模型在训练过程中总体隐私不累加。因此后续可以研究其在大型数据集与复杂模型结构中的表现。
- (2) 现实中，分布式协作学习可能由极多的参与者组成，如百万部手机等。同时分布式协作学习中的每个设备可能计算、通信和存储能力等都有很大不同。因此有关实际应用中的通信、异构问题等也需要进行大量的研究。

(3) 分布式协作学习需要一个公平的平台和激励机制，可以在实际应用中明显体现出效果提升，并能够在永久数据记录机制（如区块链等）中留下记录。这样才能促进分布式协作学习的商业化与大规模应用。

参考文献

- [1] Garín-Mun T.Inbound international tourism to Canary Islands: a dynamic panel data model[J]. *Tourism management*, 2006, 27(2): 281-291.
- [2] Goddard M. The EU General Data Protection Regulation (GDPR): European regulation that has a global impact[J]. *International Journal of Market Research*, 2017, 59(6): 703-705.
- [3] Konečný J, McMahan H B, Yu F X, et al. Federated learning: Strategies for improving communication efficiency[J]. *arXiv preprint arXiv:1610.05492*, 2016.
- [4] Yang Q, Liu Y, Chen T, et al. Federated machine learning: Concept and applications[J]. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2019, 10(2): 1-19.
- [5] Zhang C, Xie Y, Bai H, et al. A survey on federated learning[J]. *Knowledge-Based Systems*, 2021, 216: 106775.
- [6] Xu G, Li H, Liu S, et al. Verifynet: Secure and verifiable federated learning[J]. *IEEE Transactions on Information Forensics and Security*, 2019, 15: 911-926.
- [7] 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*. 2015: 1322-1333.
- [8] Yan X, Cui B, Xu Y, et al. A method of information protection for collaborative deep

- learning under gan model attack[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics , 2019.
- [9] Sadhukhan P , Palit S. Reverse-nearest neighborhood based oversampling for imbalanced , multi-label datasets[J]. Pattern Recognition Letters , 2019 , 125: 813-820.
- [10] Sun Z, Kairouz P, Suresh A T , et al. Can you really backdoor federated learning?[J]. arXiv preprint arXiv:1911.07963 , 2019.
- [11] OMERAH YOUSUF, ROOHIE NAAZ MIR. A survey on the Internet of Things security: State-of-art, architecture, issues and countermeasures[J]. Information and Computer Security , 2019 , 27(2):292-323.
- [12] LITJENS G , KOOI T , BEJNORDI B E , et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis , 2017 , 42:60-88.
- [13] MARCO MEINARDI. In-Depth Assessment of Google Cloud Platform IaaS. Published: 10 August 2017. Google Inc Analyst(s). Gartner.com.
- [14] Truex S, Baracaldo N, Anwar A, et al. A hybrid approach to privacy-preserving federated learning[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. 2019: 1-11.
- [15] Xu R, Baracaldo N, Zhou Y , et al. Hybridalpha: An efficient approach for privacy-preserving federated learning[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. 2019: 13-23.
- [16] Liu X, Li H, Xu G , et al. Adaptive privacy-preserving federated learning[J]. Peer-to-Peer Networking and Applications , 2020 , 13(6): 2356-2366.
- [17] Li Y, Zhou Y, Jolfaei A , et al. Privacy-Preserving Federated Learning Framework

Based on Chained Secure Multiparty Computing[J]. IEEE Internet of Things Journal, 2020, 8(8): 6178-6186.

- [18] Wang Ning, Xiao Xiaokui, Yang Yin, et al. Collecting and analyzing multidimensional data with local differential privacy[C]//Proc of IEEE Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2019: 638-649.
- [19] Dwork C, Roth A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3/4):211–407.
- [20] Stanley L W. Randomized response: A survey technique for eliminating evasive answer bias[J]. Journal of the American Statistical Association, 1965, 60(309):63–69.
- [21] Duchi J C, Jordan M I, Wainwright M J. Privacy aware learning[J]. Journal of the Association for Computing Machinery, 2014, 61(6): 1–57.
- [22] Hamm J, Champion A C , Chen Guoxing, et al . Crowd-ML: A privacy-preserving learning framework for a crowd of smart devices[C]//Proc of IEEE ICDCS. Piscataway, NJ: IEEE, 2015:11–20.
- [23] Hamm J, Champion A C , Chen Guoxing, et al . Crowd-ML: A privacy-preserving learning framework for a crowd of smart devices[C]//Proc of IEEE ICDCS. Piscataway, NJ: IEEE, 2015:11–20.
- [24] Sun Lin, Ye Xiaojun, Zhao Jun, et al. BiSample: Bidirectional sampling for handling missing data with local differential privacy[C]//Proc of the 25th Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2020: 88-104.
- [25] Sicong Che, Hao Peng, Lichao Sun, Yong Chen, and Lifang He. Federated multi-view learning for private medical data integration and analysis. arXiv preprint arXiv:2105.01603, 2021.

- [26] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In Eurocrypt. Springer, 2019.
- [27] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. ASA, 2018.
- [28] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In SODA. SIAM, 2019.
- [29] Mohamed Seif, Ravi Tandon, and Ming Li. Wireless federated learning with local differential privacy. arXiv preprint arXiv:2002.05151, 2020.
- [30] Lichao Sun and Lingjuan Lyu. Federated model distillation with noise-free differential privacy. arXiv preprint arXiv:2009.05537, 2020.
- [31] Stacey Truex, Ling Liu, Ka-Ho Chow, Mehmet Emre Gursoy, and Wenqi Wei. Ldp-fed: federated learning with local differential privacy. In arXiv, 2020.
- [32] Xiaohang Xu, Hao Peng, Lichao Sun, Md Zakirul Alam Bhuiyan, Lianzhong Liu, and Lifang He. Fedmood: Federated learning on mobile health data for mood detection. arXiv preprint arXiv:2102.09342, 2021.
- [33] Lingjuan Lyu, Han Yu, Xingjun Ma, Lichao Sun, Jun Zhao, Qiang Yang, and Philip S Yu. Privacy and robustness in federated learning: Attacks and defenses. arXiv preprint arXiv:2012.06337, 2020.
- [34] Wang Ning, Xiao Xiaokui, Yang Yin, et al. Collecting and analyzing multidimensional data with local differential privacy[C]//Proc of IEEE Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2019: 638-649.

- [35] Xia Chang , Hua Jingyu , Tong Wei, et al. Distributed K-means clustering guaranteeing local differential privacy[J]. Journal of Computers Security , 2020 , 90:101699.
- [36] Athalye A , Carlini N , Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples [C]// Proc of the 35th International Conference on Machine Learning , Stockholm: ACM Press , 2018: 436–448.
- [37] Xu R , Baracaldo N , Zhou Y , et al. Hybridalpha: An efficient approach for privacy-preserving federated learning[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. 2019: 13-23.
- [38] Wang Ning , Xiao Xiaokui , Yang Yin , et al. Collecting and analyzing multidimensional data with local differential privacy[C]//Proc of IEEE Int Conf on Data Engineering (ICDE). Piscataway , NJ: IEEE , 2019: 638-649.
- [39] DWORK C , ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends® in Databases , 2014 , 9(3-4): 211-407.
- [40] BASSILY R , SMITH A , THAKURTA A. Private empirical risk minimization: efficient algorithms and tight error bounds[C]//Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science. Piscataway: IEEE Press , 2014: 464-473.
- [41] PAPERNOT N , SONG S , MIRONOV I , et al. Scalable private learning with pate[J]. arXiv preprint , 2018 , arXiv:1802. 08908.
- [42] WU X , LI F G , KUMAR A , et al. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics[C]// Proceedings of the 2017 ACM International Conference on Management of Data. New York: ACM Press , 2017: 1307-1322.

- [43] BUN M, STEINKE T. Concentrated differential privacy: simplifications, extensions, and lower bounds[C]//Proceedings of the Theory of Cryptography Conference. Berlin: Springer, 2016: 635-658.
- [44] TIANXX, SHACF, WANGXL, et al. Privacy preserving query processing on secret share based data storage[C]// Proceedings of the International Conference on Database Systems for Advanced Applications. Berlin: Springer, 2011: 108-122.
- [45] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for federated learning on user-held data[J]. arXiv preprint, 2016, arXiv:1611.04482.
- [46] PETTAI M, PEETER L. Combining differential privacy and secure multiparty computation[C]//Proceedings of the 31st Annual Computer Security Applications Conference. New York: ACM Press, 2015.
- [47] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2017: 1175-1191.
- [48] XU R H, BARACALDO N, ZHOU Y, et al. HybridAlpha: an efficient approach for privacy-preserving federated learning[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2019.
- [49] SLAWOMIR G, LI X. A comprehensive comparison of multiparty secure additions with differential privacy[J]. IEEE Transactions on Dependable and Secure Computing, 2015, 14(5): 463-477.
- [50] SADEGH RM, CHRISTIAN W, OLEKSANDR T, et al. Chameleon: a hybrid secure computation framework for machine learning applications[C]//Proceedings of the

2018 on Asia Conference on Computer and Communications Security. New York: ACM Press, 2018: 707-721.

- [51] Liu X, Li H, Xu G, et al. Adaptive privacy-preserving federated learning[J]. Peer-to-Peer Networking and Applications, 2020, 13(6): 2356-2366.
- [52] Li Y, Zhou Y, Jolfaei A, et al. Privacy-Preserving Federated Learning Framework Based on Chained Secure Multiparty Computing[J]. IEEE Internet of Things Journal, 2020, 8(8): 6178-6186.
- [53] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of cryptography conference. Springer, Berlin, Heidelberg, 2006: 265-284.

致 谢

研究生的学习过程是我人生中重要的一个阶段，期间个人的价值观发生了变化、学会了为人处世之道、专业知识有了更多积累。在毕业论文及各项实验室指标基本完成之时，感想颇多。借此向给予我帮助、理解和支持的你们致以真挚的感谢。

首先感谢我的母校——华东师范大学。在 2019 年的时候录取了我，当时的心情是那样的开心、激动，因为这给予了我肯定。学校给我们提供了优美的学习环境、丰富的教学资源和浓厚的学术氛围。因此就算在此期间遇到很多困难，也从不后悔选择华师大。

我的导师曹老师，是一个特别努力上进的人，对密码学与网络安全领域的研究有独到的见解。他一直是我学习的榜样，指引我前进的方向。每次遇到问题时，老师能够深入剖析，帮我们分析问题的解决思路。生活中的他也很亲切、和蔼。还感谢我们软件学院的所有老师，让我学到了丰富的计算机基础知识和前沿技术，辅导员老师等让我感受到华师大的温暖。

还要感谢研究生期间相处时间最多的实验室小伙伴们。我们一起学习、一起吃饭、一起加班、一起聊天、一起为论文奋斗，无比开心。之前我比较喜欢一个人学习，是你们教会了我团队协作。感谢一起进步的每一个日日夜夜！

最后感谢家人对我的理解和支持，你们浓浓的爱，是我前进的动力。

在即将说再见的时刻，心情错综复杂：有面对新环境的恐惧、朋友离别的伤心、顺利毕业的喜悦……感谢让我遇到你们，我想说你们辛苦了，愿你们家庭幸福、快快乐乐、心想事成、永生不忘！

何慧娴
二零二壹年九月

攻读硕士学位期间发表论文、参与科研和获得荣誉情况

■ 已完成学术论文

- [1] **Huixian He**, Zhenfu Cao. Adaptive Privacy-preserving and Shuffling Aggregation in Federated-learning[C]. 2021 The 11th International Workshop on Computer Science and Engineering, Shanghai, China.[第一作者]