

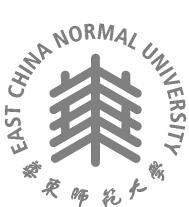
2022 届硕士专业学位研究生学位论文

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: 信息学部软件工程学院

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 20 日

Thesis (Professional) for Master's Degree in 2021

School Code: 10269

Student Number:51184501139

EAST CHINA NORMAL UNIVERSITY

TITLE: TECHNOLOGIES RESEARCH FOR PRIVACY PRESERVING BASED ON FEDERATED LEARNING

Department: Software Engineering Institute of

Information Department

Major: Software Engineering

Research Direction: Privacy Preserving

Supervisor: Associate 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 shuffle 框架，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. 为解决现有隐私算法通常需要牺牲模型精度来提高模型隐私性，从而使得模型可用性降低的挑战，我们对联邦学习场景下的本地差分模型进行了三方面的优化：第一，针对模型训练中的梯度加噪，提出了一个自适应加噪机制；第二，相对于传统的隐私预算平均分配，提出了一个更好的隐私预算分配策略；第三，根据所加的噪声量的大小分配权重，减小整体的噪声影响。

2. 在上述算法的基础上，进一步提出安全聚合模型，从而按需分配隐私模型，降低全局模型的噪声影响。此外，我们还分析了在差分隐私机制下联邦学习算法的收敛性，并根据训练中的两个误差项分别提了改进方法，即裁剪值学习方法和改进的组合方法。
3. 我们基于三个基准测试集评估了框架和算法的可行性和有效性，并且在实验中与其他方案对比，展示了模型精度的提升和隐私成本的节省。

关键词： 联邦学习，隐私保护，本地差分隐私，安全聚合

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. To address the challenge that existing privacy algorithms usually need to sacrifice model accuracy to improve model privacy, which makes the model less usable, we optimize the local difference model in the federal learning scenario in three ways: first, an adaptive noise addition mechanism is proposed for the gradient noise addition in model training; second, a better privacy budget is proposed compared to the traditional average privacy budget Second, a better privacy budget allocation strategy is proposed compared to the traditional equal allocation of privacy budgets; third, weights are assigned according to the magnitude of the noise added to reduce the overall noise impact.
2. Based on the above algorithm, a secure aggregation model is further proposed so that the privacy model can be assigned on demand and the noise impact of the global model can be reduced. In addition, we analyse the convergence of the federal learning algorithm under the differential privacy mechanism and propose improved methods based on the two error terms in training, namely the cropped value learning method and the improved combination method, respectively.
3. We evaluated the feasibility and effectiveness of the framework and algorithms based on three benchmark test sets, and demonstrated the improvement in model accuracy and privacy cost savings in experiments compared to other schemes.

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	3
1.2.1 数据异构	3
1.2.2 高昂的通信代价	4
1.2.3 安全性和隐私威胁	4
1.3 国内外研究现状	5
1.3.1 隐私威胁的研究现状	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	12
2.1 联邦学习	12
2.1.1 基本介绍	12
2.1.2 模型框架	14
2.1.3 安全和隐私威胁	15
2.2 差分隐私	17
2.2.1 基本定义	17
2.2.2 相关概念	18
2.2.3 实现机制	20

2.3	联邦学习中的差分隐私	21
2.4	神经网络	22
2.5	本章小结	23
第三章	联邦学习中的自适应本地差分机制	24
3.1	问题定义	25
3.1.1	攻击模型	25
3.2	模型概况	25
3.2.1	系统架构	25
3.2.2	本地训练	26
3.2.3	全局参数更新	28
3.3	方案设计	28
3.3.1	层间依赖传播算法	28
3.3.2	自适应噪声添加	30
3.3.3	隐私性证明	32
3.3.4	隐私预算分析	33
3.4	本章总结	35
第四章	联邦学习的安全聚合模型	36
4.1	问题定义	37
4.1.1	攻击模型	37
4.2	安全框架	37
4.2.1	Shuffle 模型	38
4.2.2	信任边界	39
4.3	方案设计	40
4.3.1	安全性假设	41
4.3.2	Shuffle 协议	41
4.3.3	隐私性证明	43
4.4	本章总结	44

第五章 实验与评估	45
5.1 基准数据集介绍	45
5.2 实验环境与配置	46
5.3 实验设计	46
5.3.1 联邦学习模型	46
5.3.2 神经网络模型	47
5.4 实验结果与分析	49
5.4.1 自适应扰动评估	49
5.4.2 安全聚合框架评估	53
5.4.3 对比实验	54
5.5 本章小结	55
第六章 总结与展望	56
6.1 总结	56
6.2 展望	57
参考文献	59
致谢	66
发表论文和科研情况	68

插图

1.1	联邦学习模型概况	3
2.1	联邦学习隐私攻击	15
2.2	差分隐私的相邻数据集示意图	18
2.3	深度神经网络结构图	23
3.1	联邦学习的系统架构	26
3.2	层间依赖传播算法	29
4.1	安全 shuffle 模型	38
5.1	卷积神经网络结构图	48
5.2	固定梯度剪裁方法下模型准确率随隐私预算变化情况	49
5.3	固定加噪方法不同隐私预算下模型训练和预测准确度变化情况	50
5.4	MINIST 数据集的精度、损失随隐私参数 c 的变化趋势	50
5.5	不同隐私预算的自适应干扰模型的准确率	52
5.6	CIFAR 数据集的精度、损失随隐私参数 c 的变化趋势	52
5.7	安全 shuffle 联邦学习模型准确率	54
5.8	EA 框架与其他联邦学习隐私保护框架在模型准确率和隐私预算的对比	55

List of Algorithms

1	联邦学习客户端本地训练算法	27
2	安全混洗算法	40
3	Shuffle 框架	41

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，或者面部识别技术帮助我们抓住了躲藏多年的逃犯，而大型工业企业也大力推动机器学习技术的应用。另一方面，我们也必须认识到，它的巨大潜力还有待实现，例如：构建基于大量病例的医疗救助诊断系统，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。传统的机器学习方法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。

大多数训练数据是由不同组织的个人或部门产生的，一个 AI 项目可能涉及多个领域，需要融合各个公司、各个部门的数据。（比如研究居民线上消费问题，需要各个消费平台的数据，可能还需要银行数据等等），但在现实中想要将分散在各地、各个机构的数据进行整合几乎是不可能的。传统的机器学习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来完成的。因此，这个中心位置不仅要有强大的计算机集群来训练和创建机器学习模型，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。然而，这用户完全信任服务器的假设已不再适用。在这种情况下，数据拥有者倾向于将数据掌握在自己手中，这就导致了孤立的数据孤岛，数据孤岛 [1] 使所有利益

相关者无法获得更多的数据。例如，每家医院的居民医疗记录的样本量完全不够，导致模型有偏差。在信贷领域，银行只能使用中央银行的信贷报告来建立风险控制模型。

然而，这些数据的采集可能涉及到用户的隐私，随着人们的隐私意识的普遍提高，相关的隐私法律法规的不断完善，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年欧洲联盟出台《通用数据保护条例》中强调保护用户的个人隐私和数据安全用户可以删除或撤回其个人数据。近年来，也有越来越多的涉及数据泄漏和隐私侵权的事情，用户们也越来越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。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，但我们被更多的小数据包围在孤岛中。大数据的基础就没有了，人工智能的基础也没有了。大数据的基础已经消失，人工智能的未来也岌岌可危。要解决大数据的困境，仅仅靠传统的方法已经出现瓶颈。两个公司简单的交换数据在很多法规包括《通用数据保护条例》是不允许的。用户是原始数据的拥有者，在用户没有批准的情况下，公司间不能交换数据。传统的机器学习和深度学习的方法本身已经成为解决大数据困境的绊脚石。简单地在两家公司之间交换数据，无论是《通用数据保护条例》还是 GDPR[2] 都是不允许的：用户是原始数据的所有者，未经其同意，数据不能在公司之间交换。

那如何创建一个机器学习框架，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、安全和监管要求，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机器学习框架，并且能够以分布式方式与成千上万的参与者协作，迭代训练一个特定的机器学习模型。由于训练数据在联合过程中保持在参与者的本地，这种机制允许参与者之间共享训练数据，同时确保每

个参与者的隐私。如图所示，联合学习的基本工作流程如下：(1) 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，指定相同的机器学习和模型训练目标。(2) 本地训练：在一个给定的通信回合中，联邦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，创建本地模型更新（即模型参数），并将这些更新发送到中央服务器。(3) 模型平均化：下一轮的全局模型是通过汇总所有通过训练不同的训练模式获得的模型更新并取其平均值来确定的。(4) 迭代地执行上述步骤以达到优化当前全局模型的目的，整个迭代过程将在全局模型参数满足收敛条件时停止。

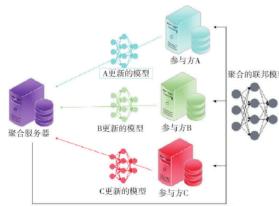


图 1.1: 联邦学习模型概况

联合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）中显示出巨大的前景，这是因为它具有独特的优势，能够从多个参与者的本地数据中训练出一个统一的机器学习模型，同时保护数据隐私 [4]。联合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，联邦学习显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开发准确的本地模型的用户（客户）来说，机器学习模型和算法的性能可以通过联合学习得到显著改善。

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存储、计算和通信能力都可能不同。由于网络和设备的限制，在任何时候都只有某些设备可以活动。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，目前的机器学习算法主要是基于对 IID 数据的假想假设。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。Federated Averageing (FedAvg) 方法来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。在训练期间，客户设备必须定期与中央服务器进行通信原始数据被储存在本地的远程客户端设备上，这些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

(1) 由于联合学习系统的云端服务器无法访问参与者的本地数据和他们的训练过程，恶意参与者可以发送无效的模型更新来达到并破坏全局模型。例如，内部攻

击者可以通过在修改后的训练数据上引起有毒的模型更新来有效地损害全局模型的准确性。内部攻击可以由联邦学习服务器发起，也可以由联邦学习参与方发起。外部攻击（包括偷听者）通过参与方与服务器之间的通信通道发起。外部攻击的发起者大部分为恶意的参与方，例如敌对的客户、敌对的分析者、破坏学习模型的敌对设备或者其组合。在联邦学习中，恶意设备可以通过白盒或者黑盒的方式访问最终模型，因此在防范来自系统外部的攻击时，需要考虑模型迭代过程中的参数是否存在泄露原始数据的风险，这对严格的隐私保护提出了新的挑战。

(2) 由于局部模型更新和全局模型参数的结合提供了关于训练数据的隐藏知识，用户的个人信息有可能泄露给不受信任的服务器或其他恶意用户。例如，即使是由其他用户的训练数据生成的样本原型也会被恶意用户隐蔽地窃取。在训练过程中，攻击方可以试图学习、影响或者破坏联邦学习模型。在联邦训练的过程中，攻击方可以通过数据中毒攻击的方式改变训练数据集合收集的完整性，或者通过模型中毒攻击改变学习过程的完整性。攻击方可以攻击一个参与方的参数更新过程，也可以攻击所有参与方的参数更新过程。若联邦学习的参与方想利用各方的数据集合训练一个模型，但是又不想让自己的数据集泄露给服务器，就需要约定联邦建模的模型算法（例如神经网络）和参数更新的机制（例如随机梯度下降（stochastic gradient descent, SGD））。那么在训练前，攻击方就可以获取联邦学习参数更新的机制，从而指定对应的推断攻击策略。

(3) 在不信任的云服务器和恶意参与者的勾结下，任何个人的确切私人信息都会被泄露。

1.3 国内外研究现状

尽管联邦学习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。本节将讨论关于联邦学习的攻击问题。从参与方的类型来看，可以将联邦学习的威胁模型细分为半诚实模型（semi-honest model）和恶意模型。对于联邦学习系统的攻击，本文按照不

同的维度进行不同层次的分类。从攻击方向角度来看，可以将联邦学习的攻击分为从内部发起和从外部发起两个方面。从攻击者的角色角度来看，可以将攻击分为参与方发起的攻击、中心服务器发起的攻击和第三方发起的攻击。从发动攻击的方式角度来看，可以将攻击分为中毒攻击和拜占庭攻击。从攻击发起的阶段角度来看，可以将攻击分为模型训练过程的攻击和模型推断过程的攻击。在密码学领域，基于模型安全的假设通常可以被分为半诚实但好奇 (onest but curious) 的攻击方假设以及恶意攻击方假设。

1.3.1 隐私威胁的研究现状

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 [6] 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。

模型反演攻击:[7] 利用这样的参数信息，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。

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。他们首先提出了一个由系统内的恶意用户发起的基于 GAN 的重建攻击。在训练阶段，攻击者可以冒充无害的用户，训练 GAN 来模拟由其他用户

的训练数据产生的原型样本。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，但其主要目标是重建被攻击用户的训练样本。

模型反演攻击：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翻转攻击 [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。

如上文所述，联邦学习机制要求所有参与者通过在本地数据集上训练全局模型来更新梯度。在这种情况下，如果联邦学习系统有一个不被信任的服务器，其知识不能被信任，那么用户的私人信息就不能得到保证。这个不受信任的服务器可以获得关于每个参与者的本地训练模型的大量额外信息（例如，模型结构、用户身份和梯度），并且能够充分损害用户的隐私信息。具体实现如下：攻击者首先在平均化后获得模型的全局参数，并在本地存储这些快照。然后，通过计算以下快照与进一步获取用户的隐私信息。

1.3.2 隐私保护的研究现状

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，重要的是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]。我们可以把它们分为两类：局部隐私和全局隐私。在本地隐私中，每个客户端发送一个不同的隐私值，该值是安全的加密的到服务器。在全局模型中，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。

安全多方计算模型涉及多方，并在一个定义明确的模拟框架中提供安全证明，

以保证完全的零知识，即每一方除了其输入和输出外一无所知。零知识是非常理想的，但这种理想的属性通常需要复杂的计算协议，而且可能无法有效实现。在某些情况下，如果提供安全保证，部分知识的披露可能被认为是可以接受的。有可能在较低的安全要求下建立一个具有 SMC 的安全模型，以换取效率 [14]。在 [15] 中，MPC 协议被用于模型训练和验证，而用户不会泄露敏感数据。最先进的 SMC 框架之一是 Sharemind[16]。[17] 的作者提出了一个具有诚实多数的 3PC 模型 [18]，并考虑了半诚实和恶意假设的安全性。这些作品要求参与者的数据在非共存的服务器之间秘密共享。

同态加密是一种加密形式，它允许人们对密文进行特定形式的代数运算得到仍然是加密的结果，将其解密所得到的结果与对明文进行同样的运算结果一样同态加密 [19]，明文通过同态加密方法得到密文后，可实现密文间的计算（密文计算后解密的结果等价于明文计算的结果）。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，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（多见为加法同态加密算法），通过加密机制下的参数交换来保护用户数据隐私 [20]，例如在 FATE 中使用的 Paillier 即为加法同态加密算法。

差分隐私方法涉及向数据添加噪音，或使用概括方法来掩盖某些敏感属性，直到第三方无法区分个人，从而使数据无法被还原以保护用户的隐私。利用差分隐私，可以在本地模型训练及全局模型整个过程中对相关参数进行扰动，从而令敌手无法获取真是模型参数，但是与密码学技术相比，差分隐私无法保证参数传递过程中的机密性，从而增加了模型遭受隐私攻击的可能性。例如刘俊旭等人 [21] 针对联邦学习下差分隐私中存在的攻击方法进行了详细的调研。在 [22] 中，作者为联合学习引入了一种差异化的隐私方法，以便通过在训练期间隐藏客户端的贡献来增加对客户端数据的保护。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

要通过使用噪声来扰乱少量步骤后的局部梯度，将差分隐私机制与 SGD 算法相结合。令人担忧的是，隐私保护预算的成本和联合学习的有效性之间的权衡是困难的，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。

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保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，通过改进层间依赖传播算法，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。
- (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淆实现客户端匿名，并在联邦学习中改进基于稀疏向量技术的差分隐私进行保护，并且证明了安全混淆模型的可行性。

- (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淆框架的结合，使得联邦学习的模型的精度和隐私预算达到平衡。

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和实际意义进行了阐述，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，根据神经网络层间依赖传播算法，分析属性值的贡献度，并分析了在差分隐私机制下的联邦学习算法的收敛性和隐私性。

第四章在上一章的基础之上，提出了一种联邦学习安全混淆模型，在联邦学习中改进基于稀疏向量技术的差分隐私进行保护，再将混淆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。并且证明了安全混淆模型的可行性。

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 联邦学习

2.1.1 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），删除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：**在两个数据集与用户重叠较多而与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。

2.1.2 模型框架

在很多横向联邦学习应用场景中，参与训练的参与方数据具有类似的数据结构(特征空间)，但是每个参与方拥有的用户是不相同的。有时参与方比较少，例如，银行系统在不同地区的两个分行需要实现联邦学习的联合模型训练；有时参与方会非常多，例如，做一个基于手机模型的智能系统，每一个手机的拥有者将会是一个独立的参与方。针对这类联合建模需求，可以通过一种基于服务器客户端的架构来满足很多横向联邦学习的需求。将每一个参与方看作一个客户端，然后引入一个大家都信任的服务器来帮助完成联邦学习的联合建模需求。在联合训练的过程中，被训练的数据将会被保存在每一个客户端本地，同时，所有的客户端可以一起参与训练一个共享的全局模型，最终所有的客户端可以一起享用联合训练完成的全局模型。

- 步骤 1: 中心服务器初始化联合训练模型，并且将初始参数传递给每一个客户端。
- 步骤 2: 客户端用本地数据和收到的初始化模型参数进行模型训练。具体步骤包括：计算训练梯度，使用加密、差分隐私等加密技术掩饰所选梯度，并将加密后的结果发送到服务器。

- 步骤 3: 服务器执行安全聚合。服务器只收到加密的模型参数，不会了解任何客户端的数据信息，实现隐私保护。服务器将安全聚合后的结果发送给客户端。
- 步骤 4: 参与方用解密的梯度信息更新各自的本地模型，具体方法重复步骤 2。

2.1.3 安全和隐私威胁

尽管联邦学习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。本节将讨论关于联邦学习的攻击问题。从参与方的类型来看，可以将联邦学习的威胁模型细分为半诚实模型 (semi-honest model) 和恶意模型。从攻击方向角度来看，可以将联邦学习的攻击分为从内部发起和从外部发起两个方面。从攻击者的角色角度来看，可以将攻击分为参与方发起的攻击、中心服务器发起的攻击和第三方发起的攻击。从发动攻击的方式角度来看，可以将攻击分为中毒攻击和拜占庭攻击。从攻击发起的阶段角度来看，可以将攻击分为模型训练过程的攻击和模型推断过程的攻击。

分类	威胁种类	发生阶段	攻击者能力	攻击者知识
针对数据	训练数据泄露	训练或预测	获得数据信息	有限知识
	模型反演攻击	预测	已知模型信息	黑盒或白盒
	成员推理攻击	预测	访问目标模型	黑盒
针对模型	模型窃取攻击	预测	访问目标模型	黑盒

图 2.1: 联邦学习隐私攻击

- **半诚实但好奇的攻击方：** 半诚实但好奇的攻击方假设也被称为被动攻击方假设。被动攻击方会在遵守联邦学习的密码安全协议的基础上，试图从协议执行过程中产生的中间结果推断或者提取出其他参与方的隐私数据。半诚实但好奇的供给方通常是客户端的角色，它们可以检测从服务器接收的所有消息，但是不能私自修改训练的过程。在一些情况下，安全包围或者可信执行环境 (trusted execution environment, TEE) 等安全计算技术的引入，可以在一定程

度上限制此类攻击者的影响或者信息的可见性。半诚实但好奇的参与方将很难从服务器传输回来的参数中推断出其他参与方的隐私信息，从而威胁程度被削弱。

- **恶意攻击方：**恶意攻击方也被称为主动攻击方。由于恶意攻击方不会遵守任何协议，为了达到获取隐私数据的目的，可以采取任何攻击手段，例如破坏协议的公平性、阻止协议的正常执行、拒绝参与协议、不按照协议恶意替换自己的输入、提前终止协议等方式，这些都会严重影响整个联邦学习协议的设计以及训练的完成情况。恶意的参与方可以是客户端，也可以是服务器，还可以是恶意的分析师或者恶意的模型工程师。恶意客户端可以获取联邦建模过程中所有参与方通信传输的模型参数，并且进行任意修改攻击。恶意服务器可以检测每次从客户端发送过来的更新模型参数，不按照协议，随意修改训练过程，从而发动攻击。恶意的分析师或者恶意的模型工程师可以访问联邦学习系统的输入和输出，并且进行各种恶意攻击。
- **成员推理攻击：**如上文所述，联邦学习机制要求所有参与者通过在本地数据集上训练全局模型来更新梯度。在这种情况下，如果联邦学习系统有一个不被信任的服务器，其知识不能被信任，那么用户的私人信息就不能得到保证。这个不受信任的服务器可以获得关于每个参与者的本地训练模型的大量额外信息（例如，模型结构、用户身份和梯度），并且能够充分损害用户的隐私信息。具体实现如下：攻击者首先在平均化后获得模型的全局参数，并在本地存储这些快照。然后，通过计算以下快照与进一步删除添加的更新，以获得其他用户的模型的汇总更新。通过这种方式，攻击者可以利用数据集的协助，得出所有其他参与者共同合作的数据样本。
- **GAN 攻击：**Hitaj 等人 [27] 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。他们首先提出了一个由系统内的恶意用户发起的基于GAN 的重建攻击。在训练阶段，攻击者可以冒充无害的用户，训练GAN 来模

拟由其他用户的训练数据产生的原型样本。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，但其主要目标是重建被攻击用户的训练样本。

2.2 差分隐私

差异化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，这就是差异性隐私 (DP) 概念的起源。差异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。

2.2.1 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)。

定义 2.2.1 (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]：

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

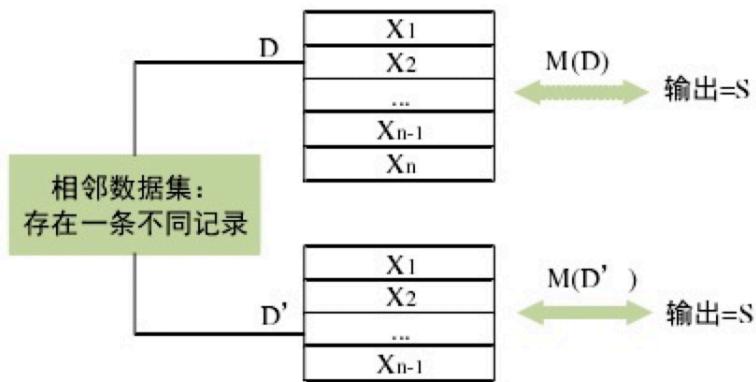


图 2.2: 差分隐私的相邻数据集示意图

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 ε -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 ε -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。如图 1 所示，算法 M 通过对输出结果的随机化来提供隐私保护，同时通过参数 ε 来保证在数据集中删除任一记录时，算法输出同一结果的概率不发生显著变化。

2.2.2 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 2.2.2 (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 2.2.3 (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 2.2.4 (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 2.2.5 (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 $(\max \{\varepsilon_1, \dots, \varepsilon_n\}) - DP$

2.2.3 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 (Exponential Mechanism) 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。

定义 2.2.6 (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 $Y \sim L(0, \frac{\Delta f}{\epsilon})$ 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。

定义 2.2.7 (指数机制). 指数机制满足差分隐私，如果：

$$M(D) = (\text{return } \varphi \propto \exp\left(\frac{\varepsilon q(D, \varphi)}{2\Delta q}\right))$$

评分函数 $q(D, \varphi)$ ，用于评估输出 φ 的质量。 Δq 代表了输出的敏感度。

ℓ_2 敏感度：对一个随机函数 $f : \mathbb{N}^{|k|} \rightarrow \mathbb{R}^k$ ，它的 ℓ_2 敏感度表示为：

$$\Delta_2 f = \max_{\substack{x, y \in \mathbb{N}^k \\ \|x - y\| = 1}} \|f(x) - f(y)\|_2$$

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 $N(0, \sigma^2)$ 。

定义 2.2.8 (高斯机制). 对于任意 $\varepsilon \in (0, 1)$ 与 $c^2 > 2 \ln(1.25/\delta)$ ，参数满足 $\sigma \geq c\Delta_2 f / \varepsilon$ 的高斯机制为 (ε, δ) -差分隐私。

2.3 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 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]。

2.4 神经网络

如图2.3所示, 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。

深度神经网络算法中, 为评估所提神经网络输出预测值与真实值之间的差异程

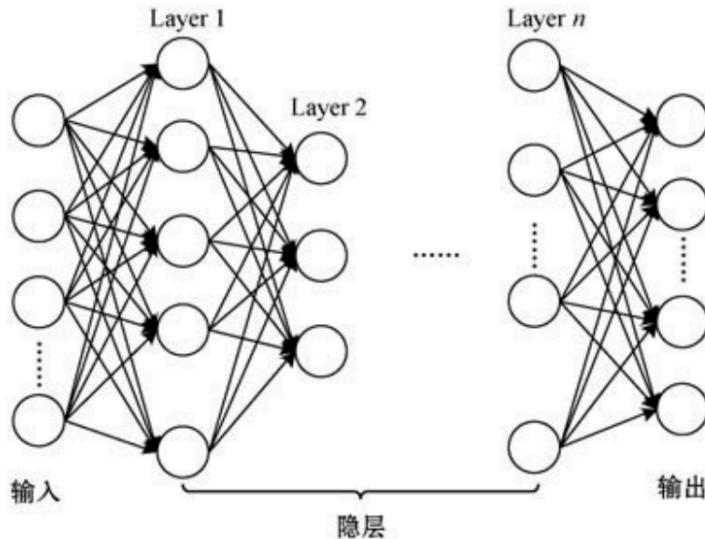


图 2.3: 深度神经网络结构图

度, 用损失函数 L 表示, 文中采用均方差损失函数, 表示为:

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

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, 下标 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 。

2.5 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

对于数据驱动的深度学习模型而言，模型参数学习数据特征的同时也记录下了敏感信息。在规定的差分隐私预算下，如何减少模型的可用性损失是关键。受传统的数据查询思想的影响，将差分隐私部署于深度学习模型中，首先可能会尝试通过仅处理由训练过程产生的最终参数（例如最终判别置信度等）来保护训练数据的私密性。这种方法将训练过程本身过程视为黑盒，但是深度学习的输出结果是一种高度抽象化的判别，细微的扰动就可能造成较大的差别。和数据隐私紧密相关的是训练过程中的梯度参数，因此针对梯度下降过程进行扰动，或将成为一种有效的保护方式。

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，方案目标为满足差分隐私条件下实现最优的模型可用性。本章节主要从以下三个方面展开研究：首先，在模型中的不同层次采用不同的梯度裁剪阈值。梯度裁剪阈值作为一个重要的参数，既影响了有效的梯度学习过程，又作为数据敏感度成为噪声添加的参数之一，区别化的梯度裁剪手段将更大限度上保留了模型的可用性。其次，分析了梯度裁剪带来的敏感度变化，结合 MA 机制进行隐私预算的跟踪，使得整体隐私损失计量更加精确化，为模型训练预留更多的空间。最后，为了从实验上证实该方法的有效性，结合成员推理攻击进行验证，进一步证实了梯度自适应加噪模型在理论约束和实际表现中具有双重隐私可靠性。

3.1 问题定义

3.1.1 攻击模型

最近研究表明深度神经网络容易受到对抗样本的攻击。为了解决这个问题，一些工作通过向图像中添加高斯噪声来训练网络，从而提高网络防御对抗样本的能力，但是该方法在添加噪声时并没有考虑到神经网络对图像中不同区域的敏感性是不同的。针对这一问题，提出了梯度指导噪声添加的对抗训练算法。该算法在训练网络时，根据图像中不同区域的敏感性向其添加自适应的噪声，在敏感性较大的区域上添加较大的噪声，抑制网络对图像变化的敏感程度，在敏感性较小的区域上添加较小的噪声，提高其分类精度。提出一种基于数据差分隐私保护的随机梯度下降算法。引入范数剪切与附加高斯噪声操作，对传统梯度更新策略进行改进。为衡量每次迭代过程中对数据隐私性的破坏，提出隐私损失累积函数在迭代过程中对数据隐私性的侵犯程度进行度量。

我们认为云服务器是一个“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 模型概况

3.2.1 系统架构

如图3.1所示，在我们的系统模型中，有两方，即云服务器和用户。

云服务器：云服务器事先与用户协商一个网络框架。然后，服务器通过公共数据训练一个初始模型，然后将初始模型的参数广播给用户。用户在本地训练各自的模型后，云服务器收集用户发送的模型梯度，并更新全球模型。

用户：用户下载由云服务器初始化的模型参数。然后，每个用户在本地数据集

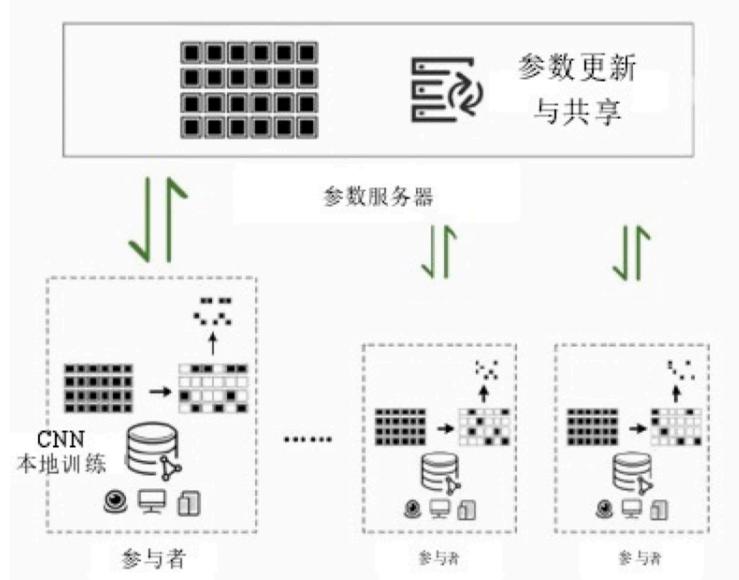


图 3.1: 联邦学习的系统架构

上训练私人模型。最后，用户将本地模型的扰动梯度发送到云服务器。

每个参与者在第一次进入联邦学习系统时，都会初始化参数。针对统一的学习目标，在本地训练集上进行模型的训练。联邦学习系统同样包括参数交换协议，在参数交换协议下，参与者将本地所得神经网络梯度的参数上传至参数服务器，同样通过参数服务器下载最新的全局参数值至本地继续训练。参与者可以在本地独立训练时，避免了使用局限训练集的单个本地模型的过拟合。模型经过训练之后，每个参与者都可以使用新的测试集独立且隐私的对其进行评估与测试，无需再进行交互操作。

3.2.2 本地训练

在分布式联邦学习中，第 i 个参与者在本地将会对全局神经网络参数的一个局部向量 w^i 进行维护、学习和更新称之为本地训练。参数服务器负责对全局参数向量 w^{global} 进行维护和更新。每个参与者在开始训练时可以随机初始化本地参数，也可以从参数服务器下载其参数最新值。

每个参与者都会使用统一标准的神经网络算法训练模型，使用的神经网络算法不局限于简单深度神经网络与卷积深度神经网络，但所有参与者需要进行统一，本

文使用的是采用选择性随机梯度下降算法全连接层的卷积神经网络 CNN。本地模型网络多次迭代训练其本地训练集。在本地训练期间,不同参与者之间不需要额外的共享样本和交互,他们通过参数服务器通过参数共享间接影响彼此的训练结果。

Algorithm 1 联邦学习客户端本地训练算法

- 1: **Input:** 全局模型参数 $\mathbf{w}^{\text{global}}$, 初始化参数 \mathbf{w}^i
 - 2: Enable users for training: initialize model//初始化模型
 - 3: for ($\text{epoch} = 1$ to n) do
 - 4: Download parameters θ_d from PS
 - 5: Run CNN on local dataset
 - 6: Update the \mathbf{w}^i according to (2 – 5)
 - 7: Compute $\Delta\mathbf{w}^i$
 - 8: Upload $\Delta\mathbf{w}_s^i$ to PS
 - 9: end
-

算法1描述了参与者在进行本地训练时具体步骤。每个参与者独立进行深度神学习训练,在每个训练阶段由五个步骤组成。在初始化之后,第 i 个参与者从参数服务器 (Parameter Server, PS) 中下载了最新参数的分量 θ_d , 将下载的值覆盖至其本地参数,之后会在本地训练数据集上训练神经网络。

在算法的第 6 步中, 参与者计算全连接层算法训练局部参数变化得到梯度向量 $\Delta\mathbf{w}^i$ 。参数 $\Delta\mathbf{w}^i$ 反映了对于第 i 个参与者, 每个神经元中的权重向量需要变化多少能够得到更精确的模型。 $\Delta\mathbf{w}^i$ 的参数信息正是其他参与者需要训练更好模型以及避免的本地数据过拟合的信息。 $\Delta\mathbf{w}_s^i$ 表示经过选择后上传的参数。在上传训练结果前, 选择一个大于阈值 T 的子集替代完整的参数向量, 参与者选择上传更有助于目标函数的梯度值, 可以使得训练迭代过程收敛更快, 模型精度更高, 以及陷入局部最优的可能性更小。

在本地训练时, 卷积神经网络的全连接层采用了选择性随机梯度下降算法。Shokri 在 [35] 中证明了其与传统的随机梯度下降算法有着几乎相同的准确性。原因是选择参数上传更新全局模型与传统随机梯度下降算法求最优化的原因相同, 选择的过程增加了最优化过程的随机性。

参与者单独训练模型时, 由于训练集的多次使用与缺少更新, 很容易陷入局部

最优。在训练本地模型时, 参与者使用梯度参数的子集对模型进行更新, 会增加模型优化过程中的随机性, 很大程度上避免了本地 SGD 过多使用相同的小样本集产生的模型过拟合。使用其他参与者用在不同数据集上训练学习的值覆盖本地学习的参数, 可以帮助每个参与者跳出局部最优, 从而得到更准确的模型。

3.2.3 全局参数更新

联邦学习通过协调深度学习任务, 建立统一的深度学习模型结构后, 参数服务器会初始化全局参数 w^{global} 。之后处理系统内参与者的上传和下载请求, 存储参与者的局部参数, 并计算更新全局参数 w^{global} 。当参与者上传参数时, 参数服务器会将上传的 $\Delta\mathbf{w}_s^i$ 的值添加至相应的全局参数中, 并为每个全局权重参数更新元数据和计数器 stat。具体更新规则如下:

对于所有的 $j \in S$:

$$w^{global} := w^{global} + \Delta\mathbf{w}_j^i \quad (3.1)$$

为了增加更新的参数的权重, 服务器可以周期性地将计数器乘以衰减因子 β , 即:

$$\text{stat} := \beta \cdot \text{stat} \quad (3.2)$$

当参与者从服务器获取具有最大统计值参数的最新值时, 将在下载期间使用这些统计信息。每个参与者都可以通过设置 θ_d 决定下载这些参数的某一部分。

3.3 方案设计

3.3.1 层间依赖传播算法

如图3.2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。

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

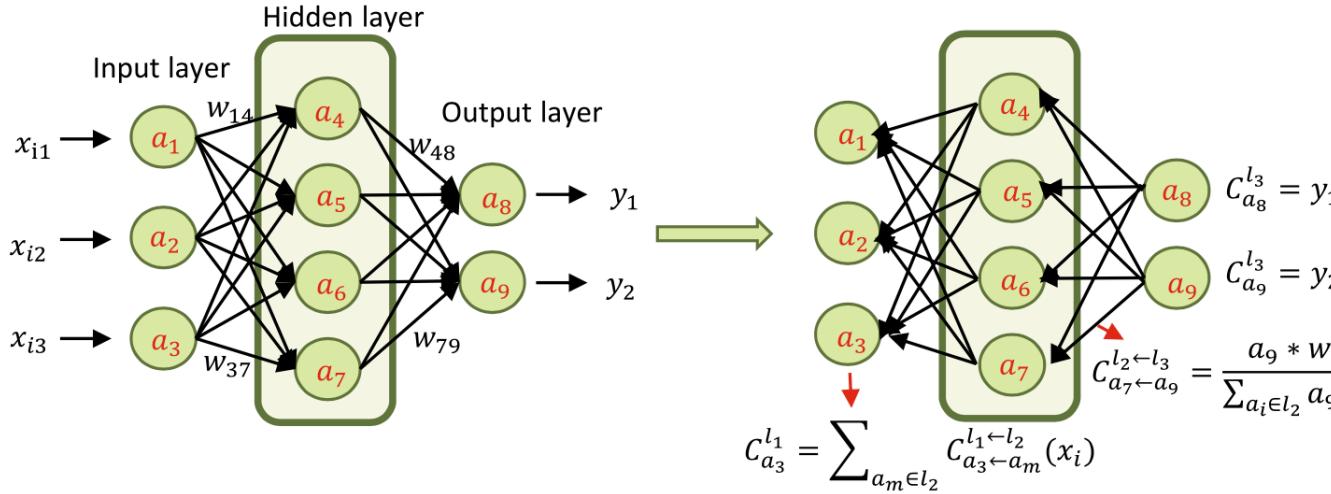


图 3.2: 层间依赖传播算法

到神经元 a_i 的相邻层的贡献之和:

$$C_{a_i}^{l_k}(x_i) = \sum_{a_j \in l_{k+1}} C_{a_i \leftarrow a_j}^{l_k \leftarrow l_{k+1}}(x_i) \quad (3.3)$$

比如，在图3.2中，存在：

$$C_{a_7}^{l_2}(x_i) = \sum_{a_j \in l_3} C_{a_7 \leftarrow a_j}^{l_2 \leftarrow l_3}(x_i) = C_{a_7 \leftarrow a_8}^{l_2 \leftarrow l_3}(x_i) + C_{a_7 \leftarrow a_9}^{l_2 \leftarrow l_3}(x_i) \quad (3.4)$$

其中，“ \leftarrow ”表示两部分之间的连接关系。具体来说，“ $l_2 \leftarrow l_3$ ”是指深度神经网络（DNN）中第二次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：

$$C_{a_i}^{l_k}(x_i) = f(x_i, \omega_i^r) \quad (3.5)$$

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：

$$C_{a_i \leftarrow a_j}^{l_{k-1} \leftarrow l_k}(x_i) = \begin{cases} \frac{a_i w_{i,j}}{\sum_{a_i \in l_{k-1}} a_i w_{i,j}} C_{a_j}^{l_k}(x_i) & \sum_{a_i \in l_{k-1}} a_i w_{i,j} \neq 0 \\ \mu & \sum_{a_i \in l_{k-1}} a_i w_{i,j} = 0 \end{cases} \quad (3.6)$$

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。

3.3.2 自适应噪声添加

该机制对连续数值型数据划分变换范围并进行分段，根据分段将其变换为1维二元分类数据。转换后使用随机响应机制进行扰动，再根据扰动后的数据标识的数值段从中随机均匀抽取数值作为扰动值。在真实数据和合成数据中的均值估计实验结果表明该机制极大地提高了准确性。除此之外，将分类变换扰动机制用于构建满足本地差分隐私的小批量梯度下降算法，并完成线性回归学习任务，实验结果证明该方法同样优于其他已有机制，可得到更小的均方误差。

在第二章中介绍了关于神经网络的结构，

$$y = a(\mathbf{x} * \omega + b) \quad (3.7)$$

公式3.7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，引入了两个调整因素。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：

$$\tilde{x}_{i,j} = \begin{cases} \ddot{x}_{i,j} & \beta \geq f \\ \bar{x}_{i,j} & \beta < f \end{cases} \quad (3.8)$$

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

$$\bar{x}_{i,j} = \begin{cases} \ddot{x}_{i,j} \text{ with probability } p \\ x_{i,j} \text{ with probability } 1 - p \end{cases} \quad (3.9)$$

f 和 p 是超参数，用户可以根据自己的情况来调整。

每个属性类的隐私预算比率 ϵ_j 由。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：

$$x'_{i,j} = x_{i,j} + \frac{1}{|D_i^t|} \text{Lap} \left(\frac{GS_l}{\epsilon_j} \right) \quad (3.10)$$

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。

我们用层间相关性传播（LRP）算法将输出分解到每一层。关于 LRP 算法的更多细节，我们将在以下部分进行介绍。每个用户都在本地对原始数据进行训练前馈操作，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 $k - th$ 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：

$$C_{a_i}^{l_k}(x_i) = \sum_{a_j \in l_{k+1}} C_{a_i \leftarrow a_j}^{l_k \leftarrow l_{k+1}}(x_i) \quad (3.11)$$

例如，如图 2 所示，我们有：

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

其中，“ \leftarrow ”表示两部分之间的连接关系。 l_{23} ”是指深度神经网络 (DNNs) 中 $2 - th$ 层和第 3 层之间相邻层的连接关系。当 $k - th$ 层为输出层时，我们有：

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

3.3.3 隐私性证明

随机隐私保护调整技术对线性变换函数进行了扰动，该函数满足 $\text{left}(\epsilon_c + \epsilon_l \text{right})$ 差分隐私。证明如下。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^{prime} 不同， $z(D_i^t)$ 和 $z(D_i^{t'})$ 分别为线性变换函数。RPAT 满足 $(\epsilon_c + \epsilon_l)$ 的差分隐私。

证明. 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：

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

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：

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

然后我们可以得到：

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

□

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，在下一节我们会给予隐私损失累积函数计算隐私成本。

3.3.4 隐私预算分析

对于所提差分隐私 SGD 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 $q = \frac{\mathcal{L}}{N}$ 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:

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

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:

$$\alpha_M(\lambda; \text{aux}, d, d') \triangleq \log \mathbb{E}_{o \sim M(\text{aux}, d)} [\exp(\lambda c(o); M(d, d'))] \quad (3.12)$$

进一步地, 映射机制 M 的损失边界值定义为:

$$\alpha_M(\lambda) \triangleq \max_{\text{aux}, d, d'} \alpha_M(\lambda; \text{aux}, d, d') \quad (3.13)$$

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:

$$\alpha_M(\lambda) \leq \sum_{i=1}^k \alpha_{M_i}(\lambda) \quad (3.14)$$

- 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:

$$\delta = \min_{\lambda} \exp(\alpha_M(\lambda) - \lambda\varepsilon) \quad (3.15)$$

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(5)–式(7)可推导得 $\alpha(\lambda) = \log \max(E_1, E_2)$, 其中:

$$E_1 = \mathbb{E}_{z \sim \mu_0} \left[\left(\frac{\mu_0(z)}{\mu(z)} \right)^\lambda \right] \quad (3.16)$$

$$E_2 = \mathbb{E}_{z \sim \mu_0} \left[\left(\frac{\mu_1(z)}{\mu(z)} \right)^\lambda \right] \quad (3.17)$$

3.4 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。通过将梯度裁剪的阈值根据网络层次进行适应性调整, 能够使得噪声的添加造成的模型可用性损失更小, 结合 MA 机制进行隐私预算的细化分配, 实验表现为相比较固定梯度裁剪阈值而言, 模型准确率更高。而且利用深度学习中的成员推理攻击证明梯度自适应加噪的方法在实验表现上也是有效的。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 联邦学习的安全聚合模型

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。

在梯度上传至参数服务器前，隐私保护方案会对梯度添加噪声，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的构成。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隐私预算为 $\epsilon_1+\epsilon_2$ 。使用联合学习训练 DNN 需要客户在多次迭代中向云上传梯度更新。如果在迭代训练过程中的每一次迭代都应用 LDP，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。这是因为局部噪声带来的误差会随着维度系数的增加而加剧 [38]。此外，由于参与一次迭代的客户端数量通常为几千人，会导致聚合升级为一个高维任务。

因此在这一小节中，论文提出一种信任域机制应用在分布式联邦学习协议中。我们将信任域与本地差分隐私相结合，用来提高在满足差分隐私保护后的联邦学习的模型精度。在本章节中我们提出了一个在联邦学习中的安全 SA 模型，本地数据使用本地差分隐私进行加密，然后所有人传到一个安全混洗器，安全混洗器打乱次序，再发给分析器（不包含任何标识信息）。安全混洗器可以作为一个可信第

三方，独立于服务器并专门用于安全混洗器。安全 SA 模型的精度增益来自于隐私放大效应 [39]，这表明本地随机机器的洗牌（即匿名）输出在差分隐私的中心视图中比没有洗牌器的输出提供更强的（放大的）隐私。因此，在洗牌模型中需要更少的本地噪音，以获得对不信任的分析器的相同水平的隐私。我们将会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。

4.1 问题定义

4.1.1 攻击模型

攻击模型以一种新的方式使用生成对抗网络。在分布式联邦学习系统框架中，我们将生成对抗网络用于从正常参与者的协作学习中提取隐私信息，生成对抗网络生成一些应该只在被攻击参与者本地训练的训练集样本。基于 GAN 的方法在联邦学习的训练阶段有效，攻击者以白盒方式访问其余参与者模型，可以得到模型的内部参数。因为联邦学习协议的目的就是共享参数，这让攻击者执行成员推理攻击时，不需要考虑只能查询每个特定输入与模型输出的黑盒模式。

在 GAN 攻击模型中，攻击方案并不是让攻击者自己完成 GAN 网络的对抗训练，而是利用生成对抗网络的思想，将本地的生成器网络和全局的神经网络进行对抗训练。在描述该部分攻击模型时，论文将分布式联邦学习系统简化为两个用户，一个为攻击者 A，一个为被攻击者 V，他们在系统中共享模型和参数。在简化模型中，被攻击者 V 有着一定量的训练集样本，对深度学习网络进行本地训练，上传模型参数。攻击者 A 在本地也有一些正常的训练样本，引导被攻击者 V 尽可能多次数的上传参数。

4.2 安全框架

该框架主要由编码器 (encoder)、混洗器 (shuffler) 和分析器 (analyzer) 3 部分组成：编码器运行在客户端，对用户数据进行本地化的编码、分割、扰动等处理；混洗器运行在一个半诚信的第三方，它可借助现有的安全混洗协议 [40][41][42][43][44][45][46]

在对数据一无所知的情况下完成安全的混淆操作；分析器运行在真正的数据收集者端，对收集的数据进行校正与分析。该框架中，混淆器完成了对用户数据完全匿名的操作，使得用户可以在尽可能对数据本身进行较小扰动的情况下，获得较多的隐私保护。

4.2.1 Shuffle 模型

SA 框架的协议由三个部分组成： $\mathcal{P} = \mathcal{A} \circ \mathcal{S} \circ \mathcal{R}^n$ ， \mathcal{A} 表示分析器， \mathcal{S} 表示 Shuffler， \mathcal{R}^n 表示 n 个用户的数据集。如图4.1所示。每个用户持有一维数据 $x \in \mathbb{X}$ ，我们将 n 个用户的数据表示为数据集 $X = (x_1, \dots, x_n) \in \mathbb{X}^n$ ，每个用户运行一个随机器 $\mathcal{R} : \mathbb{X} \rightarrow \mathbb{Y}^m$ ，将本地数据扰乱成满足 $\epsilon_{\text{local}}\text{-DP}$ 的信息。我们重点讨论 $m = 1$ 的单消息协议。Shuffler 执行 $\mathcal{S} : \mathbb{Y}^* \rightarrow \mathbb{Y}^*$ ，用均匀随机的扰动 p_i 对收到的消息进行处理。分析器函数 $\mathcal{A} : \mathbb{Y}^* \rightarrow \mathbb{Z}$ 将 shuffle 后的消息作为输入，并输出分析结果。

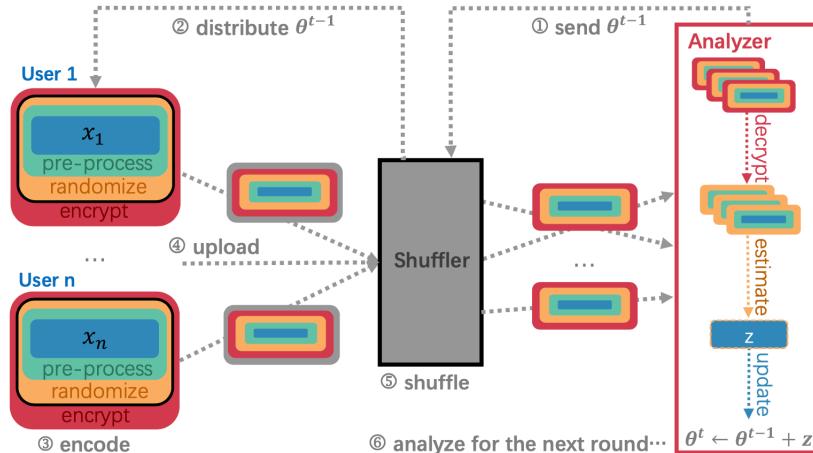


图 4.1: 安全 shuffle 模型

客户端参与方：假设有 n 个客户端参与房，每个参与房拥有一个 d -维的局部更新向量 x_i ，参照第三章的本地差分操作进行扰动，得到模型的输出 y_i 。

混淆器：混淆器是安全聚合的中心服务器，从各个客户端收到的模型输出参数进行混淆，然后发送给分析器。混淆器是独立的半诚信（semi-honest）服务器，可在对数据内容一无所知的情况下执行安全的混淆操作，是 ESA 框架的核心组件。

它的作用是接收用户编码后的数据，消除相应的元数据（包括接收时间、顺序、IP 地址等），并对接收数据进行混洗（即打乱顺序），以达到匿名目的。为保证足够的隐私保护效果，该混洗器需等待一段时间收集足够的用户数据进行混洗，并对数据量满足一定阈值的数据进行发布。当数据量为敏感信息时，可对该阈值添加满足差分隐私的噪声进行扰动，或者随机丢掉一些数据使数据量满足差分隐私 [47]，从而保护数据量隐私。

分析器 \mathcal{A} : 分析器 \mathcal{A} 从混洗器拿到安全聚合后的参数并对均方差进行估计，并更新全局模型。分析器由数据收集者运行，是不可信服务器。它的作用是接收混洗器发布的数据，依据相应的编码和混洗规则对数据进行分析与校正，并获取最终的统计结果。该框架中数据的隐私性主要是针对分析器而言的，即将分析器视为数据的窥探者。

SA 模型中的隐私目标是确保 $\mathcal{M} = \mathcal{S} \circ \mathcal{R}^n$ 满足 $(\epsilon_c, \delta_c) - \text{DP}$ ，因为 \mathcal{A} 是由一个不受信任的分析器执行，他没有义务保护用户的隐私。协议 \mathcal{P} 实现了与 \mathcal{M} 相同的隐私水平。因此，我们重点分析 $\mathcal{M}(X)$ 和 $\mathcal{M}(X')$ 的不可区分性。Erlingsson 等人 2019 年证明， \mathcal{M} 的隐私可以被“放大”。换句话说，当每个用户应用 \mathcal{R} 中的局部隐私预算 ϵ_l 时， \mathcal{M} 可以实现 $(\epsilon_c, \delta_c) - \text{DP-DP}$ 的更强隐私性，当 $\epsilon_c < \epsilon_l$ 时。与本地模型相比，洗牌模型需要更少的噪声来达到相同的隐私水平。

使用算法2中的局部随机器 $\mathcal{R}_{\gamma,b}$ ，其中 $\gamma = \frac{b}{e^{\epsilon_l} + b - 1}$ 表示从空白分布中输出一个元素的概率，输入值 x 被编码到一个离散域 $[b]$ ，然后进行随机化。在 \mathcal{S} 运行了一次置换后， \mathcal{A} 汇总洗牌结果 $\hat{z} \leftarrow \frac{1}{b} \sum_{i=1}^n y_i$ 并用以下方法进行去偏差：

$$z \leftarrow (\hat{z} - n\gamma/2)/(1 - \gamma) \quad (4.1)$$

4.2.2 信任边界

我们把观察者表示为 \mathcal{O} ，它可以是任何可以观察全局模型参数的好奇者。在分布式差分隐私的联邦学习模型中，隐私边界与 \mathcal{O} 和 \mathcal{A} 相关。而在本地差分隐私的联邦学习模型中，隐私边界在每个个体用户和其他成员之间。通过引入 Shuffler

Algorithm 2 安全混淆算法

```

1: Input: scalar  $x \in [0, 1]$ 
2: Output: perturbed value  $y \in [b]$ 
3:  $\bar{x} \leftarrow \lfloor xb \rfloor + \text{Ber}(xb - \lfloor xb \rfloor)$ 
4: Sample  $r \leftarrow \text{Ber}(\gamma)$ 
5:  $y = \begin{cases} \bar{x} & \text{if } r = 0 \\ \text{Unif}(\{1, \dots, b\}) & \text{else.} \end{cases}$ 
6: end

```

\mathcal{S} , 避免了像 DP-FL 那样将全部信任放在任何一方, 同时能够实现比 LDP-FL 更好的模型效用。

我们设计了一个细粒度的隐私分离方案, 并与表格中的 DP-FL 和 LDP-FL 进行比较。具体来说, 我们将每个本地更新的信息分为: 索引、相应的值和用户身份 (即图 2 中的 ID)。应该注意的是, 当索引被选择并以取决于值的方式发送到 shuffler \mathcal{S} 时, 它们可能是敏感的。因此, 我们的隐私目标是以数据无关的方式选择索引, 梯度的真实值对 \mathcal{S} 来说是不可见的。但 \mathcal{S} 应该知道用户的隐性身份, 以便分发全局模型参数和接收本地上传的信息。对于观察者 \mathcal{O} 来说, $(\epsilon_c, \delta_c) - \text{DP}$ 通过后处理属性而成立 [47], Shuffler 处理后的信息不会暴露用户身份, 并且满足针对 A 的 $(\epsilon_c, \delta_c) - \text{DP}$ 。在 LDP-FL 中, 每个用户都需要一个 $(\epsilon_c, \delta_c) - \text{LDP} \mathcal{R}$ 来实现这一目标。

4.3 方案设计

联邦学习的安全 shuffle 有三个构建过程: 编码 ϵ_c , 混洗 \mathcal{S} 和中央服务器 \mathcal{Z} 。在第 8 行, \mathcal{C} 是矢量的剪切阈值。我们用 ϵ_l 表示每个本地向量的本地隐私预算。 pk_a 和 sk_a 分别代表中央服务器产生的公钥和秘钥。下面的不同协议是通过实现第 10 行的 Randomize (\cdot) 和第 15 行的 Shuffle (\cdot) 函数, 以不同的策略设计的。需要注意的是, 算法 1 中的 $\mathcal{R}_{\gamma,b}$ 可以作为基本随机器应用于 Randomize (\cdot) 中, 这与行 (18) 的方程 (1) 的估计相吻合。也可以应用通用的随机器 (例如拉普拉斯机制), 这并不影响我们后面的推论。

4.3.1 安全性假设

我们假设洗牌者和分析者没有勾结（否则，模型会简化为 LDP-FL）。我们还假设加密原语是安全的，并且对手在计算上很难从密码文本中获得任何信息。

4.3.2 Shuffle 协议

Algorithm 3 Shuffle 框架

```

1: Input:  $\mathcal{A}, \mathcal{S}, \mathcal{E}, n, T, \epsilon_l, pk_a, sk_a$ 
2: Output:  $\theta^T$ 
3: 中央服务器  $\mathcal{Z}$  中央服务器公开公钥  $pk_a$ 
4: for  $t = 1, \dots, T$  do
5:    $x_i \leftarrow \text{LocalUpdate}(\theta^{t-1})$ 
6:   每个本地用户通过编码器  $\epsilon_c$  编码
7:    $\bar{x}_i \leftarrow \text{Clip}(x_i, -C, C)$ 
8:    $\tilde{x}_i \leftarrow (\bar{x}_i + C) / (2C)$ 
9:    $\langle idx_i, y_i \rangle \leftarrow \text{Randomize}(\tilde{x}_i, \epsilon_l)$ 
10:   $c_i \leftarrow \text{Enc}_{pk_a}(y_i)$ 
11:  本地用户将  $m_i = \langle idx_i, c_i \rangle$  发送给 Shuffler
12: end for
13: Shuffle
14: Shuffle 将混洗后的信息  $\text{Shuffle}(m_{i \in [n]})$  发送给中央服务器
15: 中央服务器
16: 解密消息  $y_{\pi(i) \in [n]} \leftarrow \text{Dec}_{sk_a}(c_{\pi(i) \in [n]})$ 
17: 计算均方差  $\bar{z} \leftarrow \frac{1}{n} \sum_{i \in [n]} \langle idx_i, y_i \rangle$ 
18: 求均值  $z \leftarrow C \cdot (2\bar{z} - 1)$ 
19: 更新全局模型  $\theta^t \leftarrow \theta^{t-1} + z$ 

```

我们首先提出 Shuffle 协议： $P = A \circ S \circ R_n$ ，用于 Shuffle 框架下的 d 维聚合。简而言之，我们扩展了一维协议，对每个维度进行其随机化和聚集。根据差分隐私的组成属性， R 应该满足 ϵ_{ld} -LDP，其中 $\epsilon_{ld} = \epsilon_l/d$ 。我们在 3 中声明一个 Randomize (\cdot) ，当 $idx_i \leftarrow \{1, \dots, d\}$ ，并且 $y_i \leftarrow \{\mathcal{R}_{\epsilon_{ld}}(x_{i,1}), \dots, \mathcal{R}_{\epsilon_{ld}}(x_{i,d})\}$ 。函数 Shuffle (\cdot) 简单地生成一个排列组合 π ，并输出 $mm_{\pi(i) \in [n]}$ 。然后我们针对中央服务器的 DP 后的信息进行核算。对 $\mathcal{R}_{\gamma,b}$ 或其他通用 \mathcal{R} 的数值评估采取 ϵ_{ld} 应用到 Lemma 1，可以得出一个放大的中央隐私 $(\epsilon_{cd}, \delta_{cd})$ 。有了 Lemma 3 中的构成，我们可以很容易地推导出定理 1 中的向量级构成。推论 2 提炼了从 ϵ_l 到 ϵ_c 的放大作用。

因此，相对于求和操作，缩放后的更新的灵敏度由 S 限制。GM 现在将噪声（缩放至灵敏度 S ）添加到所有缩放后的更新之和。将 GM 的输出除以 mt 可以得出所有客户更新的真实平均值的近似值，同时可以防止泄露有关个人的重要信息。

通过将此近似值添加到当前的中央模型 w_t 中，可以分配新的中央模型 w_{t+1} ：

$$w_{t+1} = w_t + \frac{1}{m_t} \left(\underbrace{\sum_{k=0}^{mt} \Delta w^k / \max \left(1, \frac{\|\Delta w^k\|_2}{S} \right)}_{\text{Gaussian mechanism approximating sum of updates}} \right)^{\text{Noise scaled to } S} \quad (4.2)$$

当将 $1 / mt$ 分解为高斯机制时，我们注意到平均值的失真由噪声方差 $S^2\sigma^2 / m$ 决定。但是，这种失真不应超过某个限制。否则，来自二次采样平均值的太多信息会被添加的噪声破坏，并且不会有任何学习进度。GM 和随机子采样都是随机机制。（实际上，[48] 正是在 dp-SGD 中使用了这种平均逼近。但是，它用于梯度平均，在每次迭代时都隐藏单个数据点的梯度）。因此， σ 和 m 还定义了随机机制提供平均近似值时所引起的隐私损失。

为了跟踪这种隐私损失，我们利用了 Abadi 等人提出的时刻会计 [49]。与标准组成定理相比，这种会计方法对发生的隐私损失提供了更严格的限制。策展人每次分配新模型时，会计都会根据 q , σ 和 m 评估 δ 。一旦 δ 达到某个阈值，即客户贡献被透露的可能性变得过高，培训应停止。 δ 阈值的选择取决于客户 K 的总数。要确定不保留许多人的隐私，而要以泄露一些人的全部信息为代价，我们必须确保 $\delta \ll K$ 。

选择 S : 削减贡献时，需要权衡取舍。一方面，应选择较小的 S ，以使噪声方差保持较小。另一方面，一个人想保持尽可能多的原始贡献。按照 [50] 提出的程序，在每个通信回合中，我们计算所有未裁剪贡献的中位数范数，并将其用作裁剪界限 $S = \text{median} \{\Delta w_k\}_{k \in Z_t}$ 。我们不使用随机机制来计算中位数，严格来说，这是对隐私的侵犯。但是，通过中位数的信息泄漏很小（未来的工作将包含这种隐私措施）。

选择 σ 和 m : 对于固定的 S ，比率 $r = \sigma^2 / m$ 控制失真和隐私损失。因此， σ 越

高, m 越小, 隐私损失就越大。隐私权会计师告诉我们, 对于固定的 $r = \sigma^2/m$, 即, 对于相同的失真水平, 对于 σ 和 m 都较小的隐私权损失较小。因此, 失真率 r 的上限和子采样客户数 m 的下限将导致 σ 的选择。但是, 很难估计 m 的下限。也就是说, 因为联合设置中的数据是非 IID, 并且来自客户端的贡献可能非常不同。因此, 我们将客户之间的差异 V_c 定义为客户更新之间相似度的量度。

4.3.3 隐私性证明

因为稀疏向量技术在隐私预算方面的潜在节省, 本文选择将应用稀疏向量技术的差分隐私与应用选择梯度上传的分布式联邦学习协议相结合, 达到更好的隐私保护效果。在应用稀疏向量技术时, 我们对算法中有关输出向量阈值的选择与分布式联邦学习协议中的选择梯度上传阈值优化为同一步骤。为了保证经过这一修改后的算法仍然满足差分隐私, 将在这一节对论文使用的方案是否满足差分隐私进行进一步的数学推导证明。

证明. 证明: 基于安全 shuffle 的联邦学习随机梯度上传满足差分隐私。

考虑任意输出向量 $a \in \{T, 1\}^\ell$, a_i 表示 a 的第 i 个组成部分。 $I_T = \{i : a_i = T\}$ 表示查询的结果高于阈值的个数, $I_\perp = \{i : a_i = 1\}$, 表示查询的结果低于阈值的个数, 参数的敏感度不超过 Δ ; 隐私参数稀疏最大值 c , 差分隐私保护隐私预算 ϵ 。

可以得到梯度的 $\Pr[A(W) = a]$ 概率:

$$\Pr[A(W) = a] = \int_{-\infty}^{+\infty} \Pr[\rho = z] \prod_{i \in I_T} \Pr[q_i(W) + v_i \geq \mathbf{T}_i + z] dz \times \prod_{i \in I_\perp} \Pr[q_i(W) + v_i < \mathbf{T}_i + z] dz \quad (4.3)$$

对于梯度 W 和保护后的梯度 W' , 可以得到:

$$\frac{\Pr[A(W) = a]}{\Pr[A(W') = a]} = \frac{\int_{-\infty}^{+\infty} \Pr[\rho = z] \prod_{i \in I_T} \Pr[q_i(W) + v_i \geq \mathbf{T}_i + z] dz}{\int_{-\infty}^{+\infty} \Pr[\rho = z] \prod_{i \in I_T} \Pr[q_i(W') + v_i \geq \mathbf{T}_i + z] dz} \times \frac{\int_{-\infty}^{+\infty} \Pr[\rho = z] \prod_{i \in I_\perp} \Pr[q_i(W) + v_i < \mathbf{T}_i + z] dz}{\int_{-\infty}^{+\infty} \Pr[\rho = z] \prod_{i \in I_\perp} \Pr[q_i(W') + v_i < \mathbf{T}_i + z] dz} \quad (4.4)$$

设

$$\begin{aligned} f_i(W, z) &= \Pr[q_i(W)v_i < \mathbf{T}_i + z] \\ g_i(W, z) &= \Pr[q_i(W)v_i \geq \mathbf{T}_i + z] \end{aligned} \tag{4.5}$$

可以得到:

$$\begin{aligned} &\frac{\Pr[A(W) = a]}{\Pr[A(W') = a]} \\ &= \frac{\int_{-\infty}^{+\infty} \Pr[\rho = z] \prod_{i \in I_T} g_i(W, z) \prod_{i \in I_\perp} f_i(W, z)] dz}{\int_{-\infty}^{+\infty} \Pr[\rho = z] \prod_{i \in I_T} g_i(W', z) \prod_{i \in I_\perp} f_i(W', z)] dz} \\ &= \frac{\int_{-\infty}^{+\infty} \Pr[\rho = z - \Delta] \prod_{i \in I_T} g_i(W, z - \Delta) \prod_{i \in I_\perp} f_i(W, z - \Delta)] dz}{\int_{-\infty}^{+\infty} \Pr[\rho = z] \prod_{i \in I_T} g_i(W', z) \prod_{i \in I_\perp} f_i(W', z)] dz} \\ &\leq \frac{\int_{-\infty}^{+\infty} e^{\epsilon_1} \Pr[\rho = z] \prod_{i \in I_\perp} f_i(W', z) \prod_{i \in I_T} g_i(W, z - \Delta)] dz}{\int_{-\infty}^{+\infty} \Pr[\rho = z] \prod_{i \in I_\perp} f_i(W', z) \prod_{i \in I_T} g_i(W', z)] dz} \\ &\leq \frac{\int_{-\infty}^{+\infty} e^{\epsilon_1} \Pr[\rho = z] \prod_{i \in I_\perp} f_i(W', z) \prod_{i \in I_T} e^{\frac{\epsilon_\Omega}{c}} g_i(W', z)] dz}{\int_{-\infty}^{+\infty} \Pr[\rho = z] \prod_{i \in I_\perp} f_i(W', z) \prod_{i \in I_T} g_i(W', z)] dz} \\ &\leq e^{\epsilon_1} \left(e^{\frac{\epsilon_\Omega}{c}} \right)^c = e^{\epsilon_1 + \epsilon_1} = e^\epsilon \end{aligned}$$

□

根据上述推导证明可知, 基于安全 shuffle 的联邦学习随机梯度上传满足差分隐私, 所上传的梯度是满足 $(\epsilon_1 + \epsilon_2)$ -差分隐私的。

4.4 本章总结

本章针对第三章所提的联邦学习自适应本地差分机制为提出了基于信任域的优化方案。结合联邦系统参与者共享参数时选择梯度上传的规则, 优化参数上传的阈值选择, 在分布式联邦学习系统中设计更优的差分隐私保护方案。通过合理的参数选择, 降低方案复杂度, 对每部分上传的梯度进行差分保护, 使分布式联邦学习系统达到隐私保护的效果。在此基础上, 将信任域群保护和差分隐私保护方法结合在分布式系统中, 在隐私保护协议下提高系统的整体性能。实验评估了差分隐私的保护效果, 合理的参数选择才能够平衡数据隐私性和模型效率。

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。本实验是关于分布式联邦深度学习系统的保护方案。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树模型鲁棒性验证框架：

- (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 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对整个系统全局模型精度的影响之前，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，如下图所示，实验首先研究学习率 α 的不同

对 MNIST 数据集, 如图 4.3(a), 和 CIFAR 数据集, 如图 4.3(b), 的影响。同时, 将 30 名参与者组成的联邦学习系统与集中式深度学习系统进行比较。

更好的学习率设置能明显的提升系统的准确率。结果表明, 较高的学习率确实能够更快地收敛到最大的精度。在超参数选择合理的情况下, 分布式联邦学习中地选择梯度上传算法并没有改变梯度神经网络的整体训练收敛行为。

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 是每次迭代满足选择梯度稀疏向量算法的梯度总数, 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验所设置所有参数的灵敏度 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那个, 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：

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

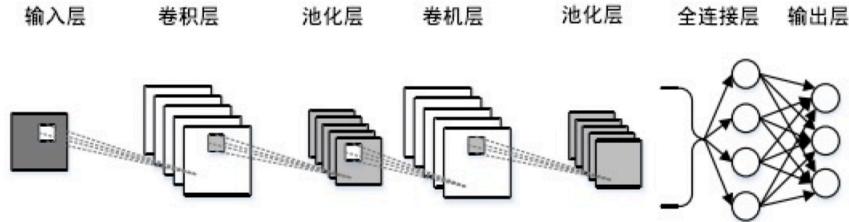


图 5.1: 卷积神经网络结构图

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 $\gamma = 0.99$ ；Adult 实验参数为 $T = 250$ ，步长 $\alpha = 6e - 5$ ，衰减系数 $\gamma = 0.99$ 。

5.4 实验结果与分析

5.4.1 自适应扰动评估

对于实验（1），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。

(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 $\delta = 10^{-5}$ 。隐私预算参数 ϵ 为研究变量。在不同的隐私预算变量下，模型的准确度变化如图5.3所示， ϵ 越大，最终模型预测准确率越高。这时候由隐私预算参数越大，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 $\epsilon \geq 5$ 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。

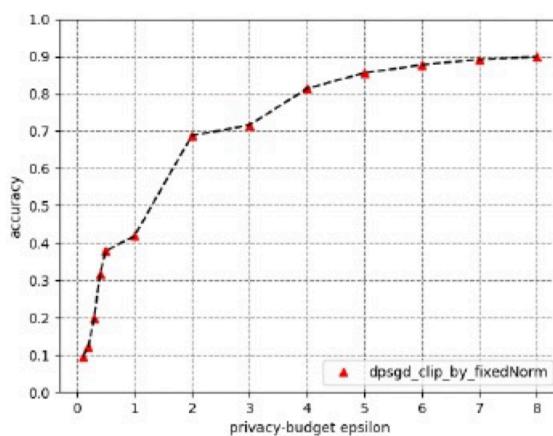


图 5.2: 固定梯度剪裁方法下模型准确率随隐私预算变化情况

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下：

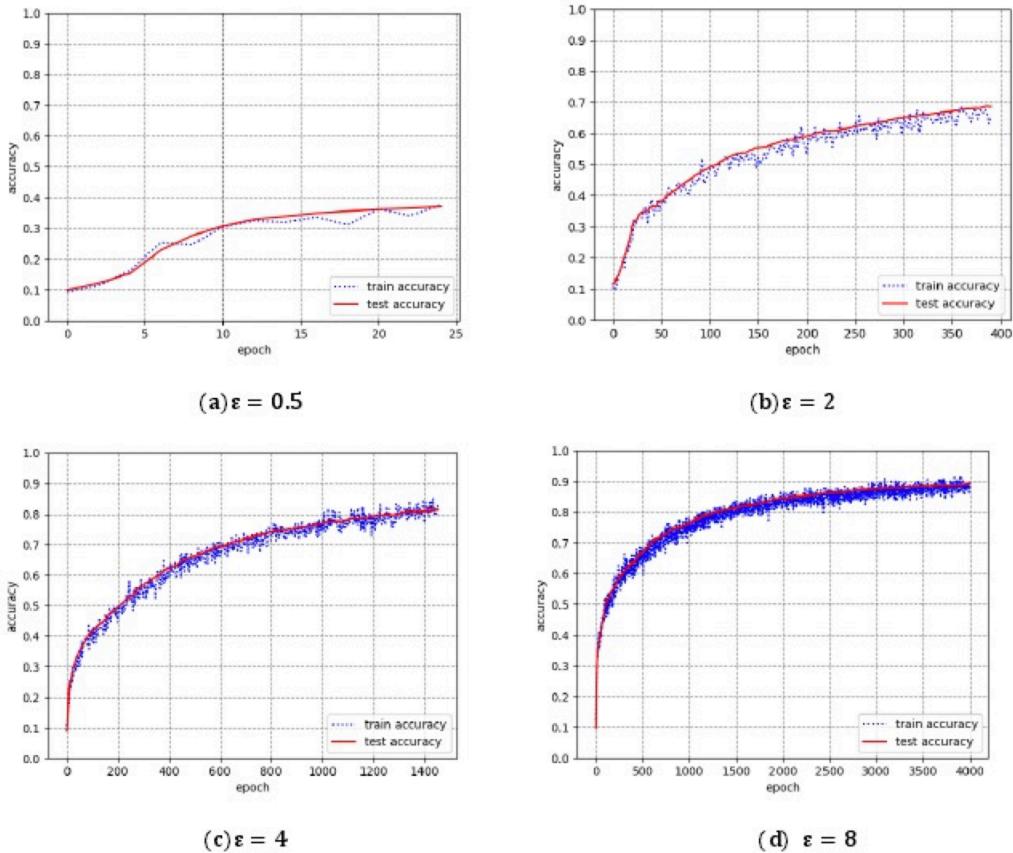


图 5.3: 固定加噪方法不同隐私预算下模型训练和预测准确度变化情况

为了证明自适应隐私预算分配的有效性，我们对 γ 参数的取值进行分析。实验中，我们实验设置与先前的保持一致，隐私参数 $\epsilon = 0.1$ 。

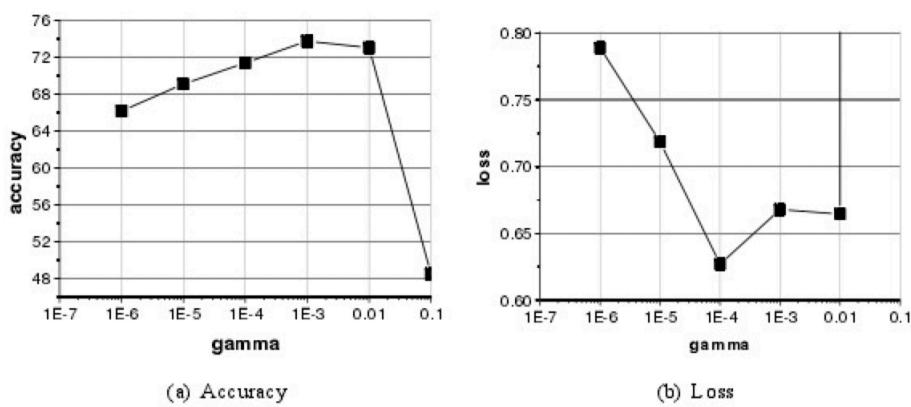


图 5.4: MINIST 数据集的精度、损失随隐私参数 c 的变化趋势

图??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。

我们比较了不同隐私预算 ($\epsilon_1 = 0$) 下的自适应干扰模型的准确性。隐私预算 ($c_1 = 0.1, c_2 = 0.5, c_3 = 2.0, c_4 = 8.0$)。隐私预算 c 越小，噪音就越大。我们还为每个隐私预算选择三个不同的调整因素预算 ((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。系统的隐私水平。此外，值得注意的是，实验中的隐私预算 c 的值是实验中的隐私预算 c 是 $cc, \epsilon l$ 和 ϵf 。我们将 ϵ 平均分为以下三个步骤：贡献计算、线性转换和贡献的计算，线性转换和即： $cc = cl = cf = 3$ 。如5.5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，APFL 的精度逐渐降低，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。

为了验证自适应权重的有效性，我们进行了使用自适应权重和不使用这方法在不同的隐私参数 c 下的对比实验。图 5.6描绘了两种方法在不同隐私参数 c 下的趋势情况。我们可以看到自适应权重基本上占有绝对的优势，尤其是在损失函数值，在隐私参数 $c=0.1$ 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，自适应权重的分配使得聚合时个体信噪比不变，但整体的

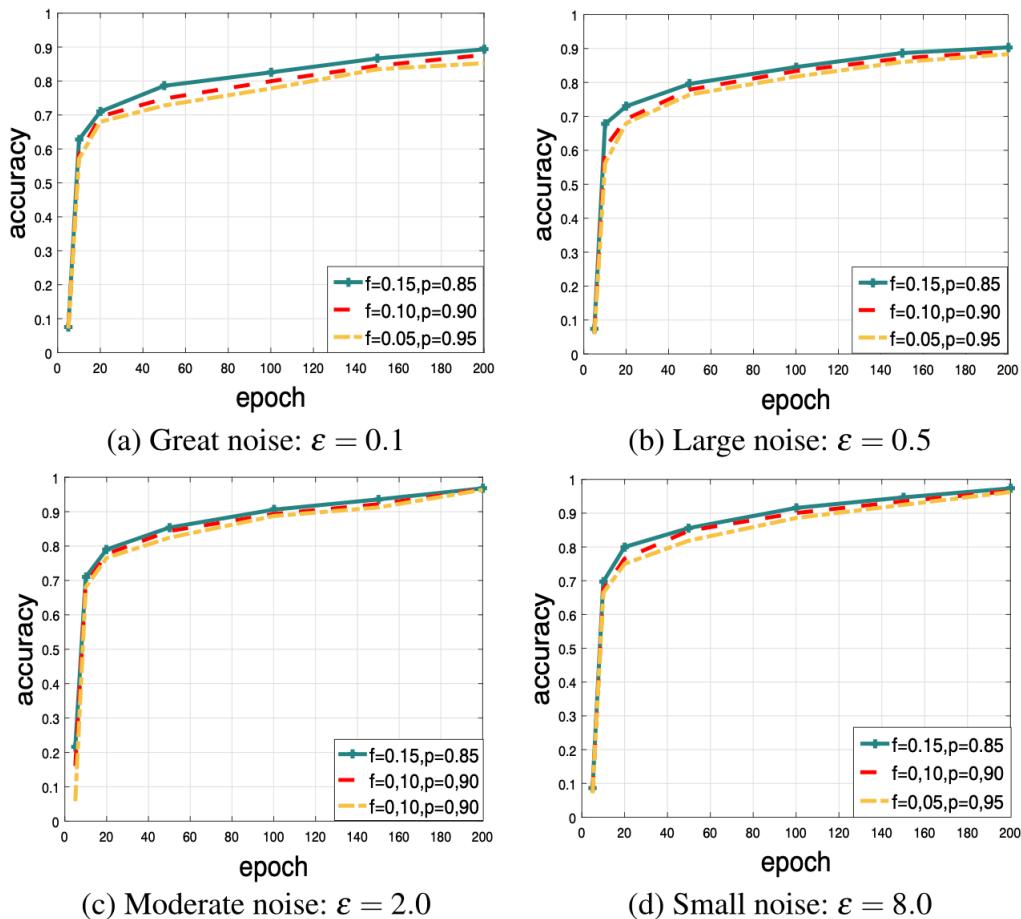
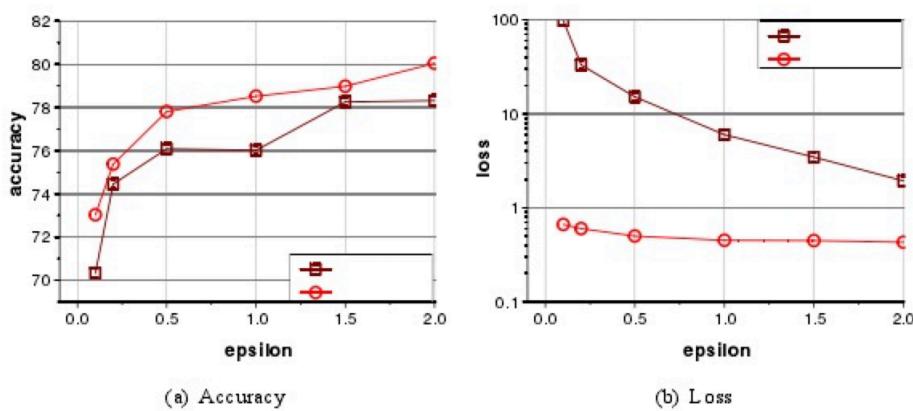


图 5.5: 不同隐私预算的自适应干扰模型的准确率

图 5.6: CIFAR 数据集的精度、损失随隐私参数 c 的变化趋势

聚合结果的信噪比却提高了很多，因此当隐私参数 c 很小，即噪声量很大的时候，表现越好。而当 c 越大时，注入的噪声也就越小，自适应权重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，用户需要用层间相关性传播算法计算权重。这个过程只需要训练中的正向传播过程，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。

5.4.2 安全聚合框架评估

为了选择适合模型训练的超参数，在保证总隐私预算不变的情况下，实验分别将分布式联邦学习系统分为 $kk = 3, kk = 6, kk = 10$ 个信任域进行对比，并且加入集中式深度学习和无信任域联邦学习（即 $kk = 30$ ）对照组与信任域实验组进行比较。

通过上面的实验，如图5.7(a) 所示，可以发现每次上传使用的隐私预算不变，相比于直接在每个参与者共享参数时添加噪声满足差分隐私，将参与者分至不同的信任域内，参与者先在信任域内训练的方式，可以得到更高精度的模型，这证明论文基于信任域对差分隐私联邦学习优化的方案是正确的。对比分成不同数量的信任域的情况可以发现，当分组的 k 值不是很大时，基于信任域方案的准确率变化并不是很大。这是因为在信任域间传输参数的量减少，进而降低了参数的扰动情况，使联邦学习的全局训练模型可以更接近于未添加噪声模型的准确度。为了使系统得到更充足的保护，我们改变每次上传的隐私预算，添加更多噪声。如图5.7(b)，将隐私预算设为 $c=0.1$ ，进行直接差分保护的模型精度会变得很低，而基于信任域方案的准确率则能够得到较好的模型。当实验进一步改变隐私预算，增加隐私保护

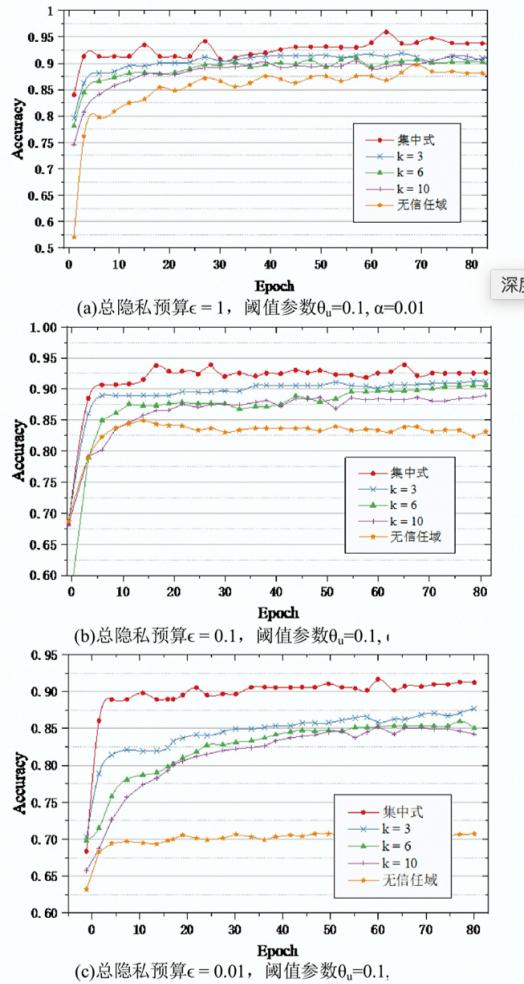


图 5.7: 安全 shuffle 联邦学习模型准确率

效果, 当隐私预算设为 $\frac{\epsilon}{c}=0.01$, 基于稀疏向量的差分隐私方案会因为添加过多的噪声, 导致整个联邦系统无法很好的收敛, 使用信任域优化方案则可以改善模型糟糕的准确率。

5.4.3 对比实验

从隐私成本和模型精度的总体上看, 混洗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, ESA 框架中混洗器的引入, 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 EA 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 ϵ_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。

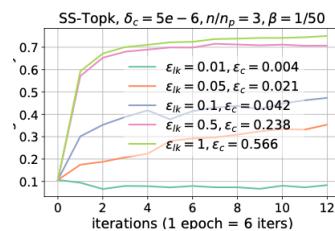


图 5.8: EA 框架与其他联邦学习隐私保护框架在模型准确率和隐私预算的对比

5.5 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。安全混淆框架能在信任域环境下通过混淆差分隐私提高数据的可用性。

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，提出了一套针对分布式数据的攻击模型，同时研究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) 本文提出了一个满足本地差分隐私的分类变换扰动机制。该机制将数值型数

据的扰动与分类型数据的扰动进行结合，提高了均值估计的准确性。同时，将该机制用于梯度下降中的每次迭代的梯度扰动，保护了训练过程中用户隐私的同时得到了 1 个较为准确的模型。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分类变换扰动机制的性能，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 [9-30]，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) 本文提出的基于解析高斯机制与函数机制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,etal. 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.

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。黄老师有着女生的特质，很细心、考虑面面俱到、管理井井有条，让我很钦佩。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。感谢一直陪伴着我的女朋友，无论是我开心，还是伤心。我们同甘共苦，一起走过多个春夏秋冬。

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]