

西安交通大学

硕士学位论文

基于卡尔曼滤波的联邦学习优化研究

学位申请人：王炜飞

指导教师：任雪斌副教授

学科名称：计算机科学与技术

2024 年 05 月

Research On Federated Learning Optimization Based On Kalman Filter

A thesis submitted to
Xi'an Jiaotong University
in partial fulfillment of the requirements
for the degree of
Master of Engineering

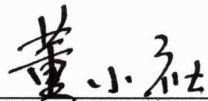
By
Weifei Wang
Supervisor: (Associate) Prof. Xuebin Ren
Computer Science and Technology
May 2024

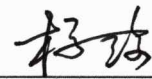
硕士学位论文答辩委员会

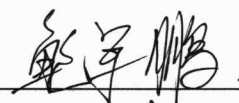
基于卡尔曼滤波的联邦学习优化研究

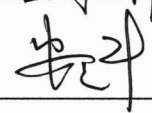
答辩人：王炜飞

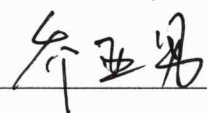
答辩委员会委员：

西安交通大学教授：董小社  (注：主席)

西安交通大学研究员：杨琦 

西安交通大学副教授：鲍军鹏 

西安交通大学高级工程师：安健 

西安交通大学教授：乔亚男 

答辩时间：2024 年 05 月 14 日

答辩地点：西安交通大学创新港 4-7216

摘要

在物联网时代，数据通常是从分布式系统的边缘位置捕获的，其他组织往往难以收集到所需数据进而造成了数据孤岛现象。联邦学习通过共享模型参数的方式协同训练模型，有效打破了数据孤岛。然而，当前的联邦学习方法无法有效抵御推理攻击，差分隐私能有效解决该问题并给联邦学习提供更强的隐私保证，但会额外引发隐私效用权衡问题。由于联邦学习是由多个相对独立的机器学习训练过程所组成，所以开展机器学习场景下的隐私效用权衡问题研究对解决该问题有较强意义。除此之外，去中心化联邦学习可以有效解决联邦学习中的信任依赖等问题，但去中心化架构下的共识问题会给训练过程带来较大影响，故开展去中心化联邦学习的共识优化研究也具备一定意义。综上，本文针对差分隐私造成的隐私效用权衡问题和去中心化联邦架构中的共识问题，开展基于卡尔曼滤波的联邦学习优化研究。具体贡献如下：

针对差分隐私造成的隐私效用权衡问题，开展基于卡尔曼滤波的隐私保护机器学习优化方法研究，提出两个基于卡尔曼滤波的隐私保护机器学习优化算法，通过利用卡尔曼滤波器能过滤高斯噪声的性质，加快模型收敛速率的同时提升了模型的收敛精度，实现了隐私和效用之间更好的权衡。具体地，首先通过卡尔曼滤波器过滤模型训练过程中添加的高斯噪声，以降低噪声对模型精度影响，基于梯度线性变化假设提出了基于卡尔曼滤波的梯度过滤方法；其次引入动量更新方法来加快收敛速率，同时针对其带来的超调问题，基于超调噪声线性变化假设进一步提出基于卡尔曼滤波的超调过滤方法，以确保模型能在训练后期稳定地收敛，提升收敛精度。本文经过大量的实验证明，该方法可以在相同隐私预算的前提下使得模型最终能够达到更好的效果。

针对去中心化联邦学习架构中的系统共识问题，开展基于卡尔曼滤波器的隐私保护联邦学习算法优化研究，提出了基于共识卡尔曼滤波器的去中心化联邦学习优化方法，通过利用共识卡尔曼滤波器能过滤高斯噪声以及在去中心化系统中加速达成共识的性质，加快去中心化架构中各参与方达成共识的速率，有效解决了共识问题对模型训练过程的影响，进而提升系统最终的收敛精度。具体地，首先将创新点一提出的两个基于卡尔曼滤波器的优化方法扩展至联邦学习架构下，然后通过采用共识卡尔曼滤波器，将传统的联邦学习架构替换为去中心化的分散系统架构。本文经过大量的实验证明，该方法能够有效解决系统共识问题对最终模型精度的影响。

* 本研究得到国家自然科学基金（编号：62172329，U21A6005）的资助

关 键 词： 联邦学习；差分隐私；卡尔曼滤波器；去中心化联邦学习；超调问题

论文类型： 应用研究

ABSTRACT

In the IoT era, data is often captured from the edge of distributed systems, making it difficult for other organizations to collect the required data, creating data silos. Federated learning collaboratively trains models by sharing model parameters, effectively breaking data silos. However, current federated learning methods cannot effectively resist inference attacks. Differential privacy can effectively solve this problem and provide stronger privacy guarantees for federated learning, but it will additionally cause privacy utility trade-off issues. Since federated learning is composed of multiple relatively independent machine learning training processes, it is of great significance to study the privacy utility trade-off problem in machine learning scenarios to solve this problem. In addition, decentralized federated learning can effectively solve problems such as trust dependence in federated learning. However, the consensus issue under the decentralized architecture will have a greater impact on the training process. Therefore, the consensus of developing decentralized federated learning is Optimization research also has certain significance. In summary, this paper conducts research on federated learning optimization based on Kalman filtering to address the privacy utility trade-off problem caused by differential privacy and the consensus problem in decentralized federated architecture. The specific contributions are as follows:

Aiming at the privacy utility trade-off problem caused by differential privacy, we conducted research on privacy-preserving machine learning optimization methods based on Kalman filtering, and proposed two privacy-preserving machine learning optimization algorithms based on Kalman filtering. By using the Kalman filter, we can filter Gaussian noise. It speeds up the convergence rate of the model while improving the convergence accuracy of the model, achieving a better trade-off between privacy and utility. Specifically, the Gaussian noise added during the model training process is first filtered through the Kalman filter to reduce the impact of noise on the model accuracy. Based on the assumption of linear gradient changes, a gradient filtering method based on Kalman filter is proposed; secondly, the momentum update method is introduced to speed up the process. Convergence rate, and in view of the overshoot problem caused by it, based on the linear change assumption of overshoot noise, an overshoot filtering method based on Kalman filter is further proposed to ensure that the model can converge stably in the late training period and improve the convergence accuracy. Through a large number of experiments, this article proves that this method can enable the model to achieve better results

* The work was supported by the Foundation (No.62172329, No.U21A6005)

under the premise of the same privacy budget.

Aiming at the system consensus problem in the decentralized federated learning architecture, we conducted research on the optimization of privacy-preserving federated learning algorithms based on the Kalman filter, and proposed a decentralized federated learning optimization method based on the consensus Kalman filter. By using the consensus Kalman filter can filter Gaussian noise and accelerate consensus in decentralized systems, speeding up the rate at which all participants in the decentralized architecture reach consensus, effectively solving the impact of consensus issues on the model training process, thereby improving the final performance of the system. Convergence accuracy. Specifically, the two Kalman filter-based optimization methods proposed in Innovation Point 1 are first extended to the federated learning architecture, and then the traditional federated learning architecture is replaced with a decentralized distributed system by using the consensus Kalman filter. architecture. Through a large number of experiments, this paper proves that this method can effectively solve the impact of system consensus problems on the accuracy of the final model.

KEY WORDS: Federated learning; Differential privacy; Kalman filter; Decentralized federated learning; Overshoot problem

TYPE OF THESIS: Application Research

目 录

| | |
|---------------------------------|-----|
| 摘 要 | I |
| ABSTRACT | III |
| 1 绪论 | 1 |
| 1.1 研究背景和意义 | 1 |
| 1.2 相关研究现状 | 2 |
| 1.2.1 隐私保护机器学习 | 2 |
| 1.2.2 隐私增强的联邦学习 | 3 |
| 1.2.3 去中心化联邦学习 | 4 |
| 1.3 研究内容 | 5 |
| 1.4 论文结构组织 | 5 |
| 2 相关理论知识和技术 | 7 |
| 2.1 差分隐私理论基础 | 7 |
| 2.1.1 差分隐私定义 | 7 |
| 2.1.2 差分隐私变种 | 9 |
| 2.1.3 基于差分隐私的随机梯度下降方法 | 11 |
| 2.2 卡尔曼滤波理论基础 | 12 |
| 2.2.1 卡尔曼滤波器 | 12 |
| 2.2.2 卡尔曼信息共识滤波器 | 14 |
| 2.3 动量优化方法和超调现象 | 15 |
| 2.3.1 动量优化方法 | 15 |
| 2.3.2 动量超调问题及解决方案 | 16 |
| 2.4 本章小结 | 17 |
| 3 基于卡尔曼滤波器的隐私保护机器学习优化方法研究 | 18 |
| 3.1 系统模型及问题定义 | 18 |
| 3.1.1 系统模型 | 18 |
| 3.1.2 问题定义 | 19 |
| 3.2 基于卡尔曼滤波器的隐私保护机器学习优化方法 | 22 |
| 3.2.1 梯度过滤方法 | 22 |
| 3.2.2 超调过滤方法 | 24 |
| 3.2.3 隐私保护机器学习优化方法 | 27 |
| 3.3 实验结果与分析 | 28 |
| 3.3.1 实验设置 | 29 |
| 3.3.2 梯度过滤方法实验 | 30 |
| 3.3.3 超调过滤方法实验 | 34 |
| 3.3.4 隐私保护机器学习优化方法实验 | 38 |
| 3.4 本章小结 | 40 |

| | |
|------------------------------------|----|
| 4 基于卡尔曼滤波器的隐私保护联邦学习算法研究 | 41 |
| 4.1 系统模型及问题定义 | 41 |
| 4.1.1 系统模型 | 41 |
| 4.1.2 问题定义 | 42 |
| 4.2 基于卡尔曼滤波器的隐私保护联邦学习算法 | 43 |
| 4.2.1 联邦架构下基于卡尔曼滤波器的联邦学习优化方法 | 43 |
| 4.2.2 去中心化联邦架构下的梯度过滤方法 | 45 |
| 4.2.3 去中心化联邦架构下的超调过滤方法 | 47 |
| 4.2.4 去中心化联邦架构下的联邦学习优化方法 | 49 |
| 4.3 实验结果与分析 | 51 |
| 4.3.1 实验设置 | 51 |
| 4.3.2 基于卡尔曼滤波器的联邦学习优化方法实验 | 51 |
| 4.3.3 去中心化联邦架构下的梯度过滤方法实验 | 53 |
| 4.3.4 去中心化联邦架构下的超调过滤方法实验 | 54 |
| 4.3.5 去中心化联邦架构下的联邦学习优化方法实验 | 54 |
| 4.4 本章小结 | 56 |
| 5 结论和展望 | 57 |
| 5.1 研究结论 | 57 |
| 5.2 未来展望 | 58 |
| 致谢 | 59 |
| 参考文献 | 60 |
| 攻读学位期间取得的研究成果 | 64 |
| 答辩委员会会议决议 | 65 |
| 常规评阅人名单 | 66 |
| 声明 | |

CONTENTS

| | |
|---|-----|
| ABSTRACT (Chinese) | I |
| ABSTRACT (English) | III |
| 1 Introduction | 1 |
| 1.1 Background and Significance | 1 |
| 1.2 Related work | 2 |
| 1.2.1 Privacy-preserving Machine Learning | 2 |
| 1.2.2 Privacy-enhanced Federated Learning | 3 |
| 1.2.3 Decentralized Federated Learning | 4 |
| 1.3 Main Content | 5 |
| 1.4 Organization of the Thesis | 5 |
| 2 Related Theoretical Knowledge and Technology | 7 |
| 2.1 Differential Privacy Theoretical | 7 |
| 2.1.1 Differential Privacy Definition | 7 |
| 2.1.2 Differential Privacy Variants | 9 |
| 2.1.3 Differential Privacy Stochastic Gradient Descent | 11 |
| 2.2 Kalman Filter Theoretical | 12 |
| 2.2.1 Kalman Filter | 12 |
| 2.2.2 Kalman Information Consensus Filter | 14 |
| 2.3 Momentum Optimization Method and OverShoot | 15 |
| 2.3.1 Momentum Optimization Method | 15 |
| 2.3.2 Overshoot and Solutions | 16 |
| 2.4 Brief Summary | 17 |
| 3 Privacy-preserving Machine Learning Optimization Method Based on Kalman Filter .. | 18 |
| 3.1 System Model and Problem Definition | 18 |
| 3.1.1 System Model | 18 |
| 3.1.2 Problem Definition | 19 |
| 3.2 Privacy-preserving Machine Learning Optimization Method Based on Kalman Filter | 22 |
| 3.2.1 Gradient Filtering Method | 22 |
| 3.2.2 Overshoot Filtering Method | 24 |
| 3.2.3 Privacy-preserving Machine Learning Optimization Method | 27 |
| 3.3 Experiment and Analysis | 28 |
| 3.3.1 Experiment Settings | 29 |
| 3.3.2 Experiment of Gradient Filtering Algorithm | 30 |
| 3.3.3 Experiment of Overshoot Filtering Algorithm | 34 |
| 3.3.4 Experiment of Privacy-preserving Machine Learning Optimization Method Al- | |
| gorithm | 38 |
| 3.4 Brief Summary | 40 |

| | |
|--|----|
| 4 Privacy-preserving Federated Learning Algorithm Based on Kalman Filter | 41 |
| 4.1 System Model and Problem Definition | 41 |
| 4.1.1 System Model | 41 |
| 4.1.2 Problem Definition | 42 |
| 4.2 Privacy-preserving Federated Learning Algorithm Based on Kalman Filter | 43 |
| 4.2.1 Privacy-preserving Federated Learning Optimization Method Based on Kalman Filter | 43 |
| 4.2.2 Gradient Filtering Method Under Decentralized Federated Architecture | 45 |
| 4.2.3 Overshoot Filtering Method Under Decentralized Federated Architecture | 47 |
| 4.2.4 Federated Learning Optimization Method Under Decentralized Federated Ar- chitecture | 49 |
| 4.3 Experiment and Analysis | 51 |
| 4.3.1 Experiment Settings | 51 |
| 4.3.2 Experiment of Privacy-preserving Federated Learning Algorithm Based on Kalman Filter | 51 |
| 4.3.3 Experiment of Gradient Filtering Method Under Decentralized Federated Ar- chitecture | 53 |
| 4.3.4 Experiment of Overshoot Filtering Method Under Decentralized Federated Ar- chitecture | 54 |
| 4.3.5 Experiment of Federated Learning Optimization Method Under Decentralized Federated Architecture | 54 |
| 4.4 Brief Summary | 56 |
| 5 Conclusion and Prospect | 57 |
| 5.1 Conclusion | 57 |
| 5.2 Prospect | 58 |
| Acknowledgements | 59 |
| References | 60 |
| Achievements | 64 |
| Decision of Defense Committee | 65 |
| General Reviewers List | 66 |
| Declarations | |

1 绪论

1.1 研究背景和意义

随着计算能力的提升和数据的海量增长,人工智能在人们生活各个领域的应用越来越普及,影响到包括医疗、交通等多个领域的发展^[1]。传统的机器学习和深度学习模型主要依赖于集中式的数据中心,需要大量高质量的数据来训练。然而,在物联网时代,数据通常是从多个拥有不同所有权的分布式边缘位置生成和捕获的。在生产生活中,其他公司和组织需要高额成本才能收集大量的数据。并且除了成本之外,保护用户隐私也是限制数据获取的重要因素^[2-3]。针对上述问题,谷歌提出了联邦学习^[4],引发了学术界和工业界的广泛关注。在联邦学习的应用场景中,参与方不需要提供自己的本地数据。它们只需要在自己的私有本地数据上执行本地训练算法,通过与可信第三方共享模型参数,从而和其他参与方一起协同训练模型。联邦学习能够保证数据不出本地,进而能够在一定程度上保护用户的数据隐私^[5]。目前,联邦学习被广泛应用于各类场景下,如谷歌的预测键盘、语言助手、目标检测^[6]、医疗保健^[7]等。

然而,联邦学习并不能使用户的隐私完全无懈可击,它在保护底层训练数据的隐私信息免受推理攻击这方面仍然是不够的^[8-9],训练过程中发送的模型参数(或者梯度)以及训练模型的输出仍是隐私泄露的攻击面。攻击者可以通过分析模型参数或梯度更新等信息来推断出参与方的隐私信息,同时恶意参与方也可以通过提供错误的模型更新或篡改梯度等方式,对其他参与方的模型质量进行破坏,这些问题给联邦学习的大规模应用带来了严重安全威胁。差分隐私技术(Differential Privacy, DP)^[10]可以有效解决联邦学习中的安全问题^[11],其通过向隐私数据中添加噪声,进而可以防止攻击者通过分析模型参数或梯度等信息来推断出隐私数据的行为,减少隐私泄露的风险。并且差分隐私的隐私保护机制,还能检测和限制恶意行为,避免其他的恶意参与方篡改梯度或提供错误更新。除此之外,差分隐私技术还可以减少对可信第三方的信任依赖,因为它会提前在本地给共享的信息添加噪声,不需要将原始数据共享给第三方,因此可以减少因第三方的不可信行为或者由于安全漏洞而导致的隐私泄露风险。

尽管差分隐私在保护隐私方面有很多优势,但它也会给联邦学习带来一些额外的问题^[12]。差分隐私技术通常需要在数据中添加噪声或采用其他隐私保护机制用来保护数据的隐私信息,但这种方法往往会大大降低模型的收敛速率,并且降低模型的效用。在应用差分隐私的过程中,方差较高的噪声会提供更强的隐私保障,但同时也会导致模型的收敛速度变慢以及收敛精度降低的问题,甚至会造成模型不收敛的情况。而方差较低的噪声虽然能确保模型的收敛速度和收敛精度在一个合理的范围内,但是会消耗更多的隐私预算,无法对用户的隐私信息提供充足的保障。在实际场景中,差分隐私

往往很难在隐私保护和模型效用之间取得平衡^[13]。卡尔曼滤波器^[14]可以过滤线性动态系统中的高斯噪声，进而实现对系统状态的最优估计。在基于差分隐私的数据发布场景中^[15]，常常使用卡尔曼滤波器来过滤所添加的差分隐私噪声，基于差分隐私的后处理性质^[10]，该方法可以在确保隐私保护程度不变的前提下提升数据的效用，实现了更好的隐私效用权衡。然而尽管卡尔曼滤波器对差分隐私有很好的优化作用，但针对卡尔曼滤波在梯度下降过程中的相关研究却很少。由于联邦学习是由多个相对独立的机器学习协同训练组成的，因此开展基于卡尔曼滤波器的隐私保护机器学习的优化研究具有较强的现实意义。除此之外，由于联邦学习需要依赖可信第三方用于模型的聚合分发，这便带来了信任依赖问题^[16]、单点故障^[17]和通信瓶颈问题^[18]，去中心化的联邦学习架构可以很好的解决上述问题，与集中式联邦学习相比，去中心化联邦学习架构改善了单点故障、信任依赖以及服务器节点瓶颈的限制^[18]。然而在去中心化联邦系统中由于没有可信第三方进行统一的模型聚合分发，所以导致在训练过程中，各个参与方难以达成共识，进而给系统最终的收敛精度带来了较大的影响。因此开展基于去中心化联邦学习架构下的共识优化研究具有较强的现实意义。

综合上述背景，本文将进行基于卡尔曼滤波的隐私保护联邦学习优化方法研究，致力于在差分隐私场景下更好地实现隐私效用权衡的同时，解决传统联邦学习架构中存在的信任依赖、通信瓶颈和单点故障问题。由于联邦学习是由多个相对独立的机器学习所组成的，故为了简化问题，本文首先在集中式场景下对上述问题进行研究，然后进一步将其迁移至联邦学习场景下，具体地，该方案主要由两部分组成：一是基于卡尔曼滤波的隐私保护机器学习优化方法研究，使用基于卡尔曼滤波的梯度过滤方法过滤梯度噪声，以加快收敛速率并提高全局模型的准确性；同时引入动量优化方法加速模型的收敛，针对动量优化方法带来的超调问题，进一步引入基于卡尔曼滤波的超调过滤方法，过滤参数中的超调噪声，在模型收敛后期加快模型的收敛速率。二是基于卡尔曼滤波器的隐私保护联邦学习算法优化研究，利用共识卡尔曼滤波器用来解决去中心化联邦系统的共识问题，将提出的基于卡尔曼滤波的梯度过滤方法和基于卡尔曼滤波的超调过滤方法中的卡尔曼滤波器扩展至共识卡尔曼滤波器，利用分散系统中相邻节点的信息做模型的聚合，以解决去中心化架构下的共识问题。

1.2 相关研究现状

1.2.1 隐私保护机器学习

Abadi 等人^[19]提出了一种将机器学习或深度学习方法与差分隐私相结合的算法，名为基于差分隐私的随机梯度下降方法 (Differential Privacy Stochastic Gradient Descent, DPSGD)，并提出了一种新的计算理论，使用矩量统计算法将隐私损失视为随机变量并

估计其尾部界限。Mironov 等人^[20]提出了瑞丽差分隐私的概念，以信息论中的瑞丽熵作为基础来衡量两个分布之间的差异，提供了更紧的隐私界限。鉴于复杂的数据和任务场景，当前有许多研究工作去拓展 DPSGD 算法在各类场景下应用。Choudhury 等人^[21]提出了一种基于样本聚合框架的方法，针对由于重尾数据不规则性导致无法提供 DP 保证的问题，基于梯度平滑和修剪的方案，实现强凸损失函数下的 DP 保证。Phan 等人^[22]为了抵御隐私推理攻击和对抗性示例攻击，通过利用 DP 中的顺序组合理论来随机化输入空间和潜在空间，加强了稳健性界限。同时基于 DP 的后处理特性设计了一个原始的对抗性目标函数，以解决模型实用性、隐私损失和鲁棒性之间的权衡。Kairouz 等人^[23]考虑了在线学习，即流式数据场景，在无法进行采样和扰乱样本顺序的前提下，基于树的聚合方法，在没有采用任何形式的隐私放大方法的前提下，获得了较好准确性和隐私的权衡。Nasr 等人^[24]研究了差分隐私在机器学习中，抵御推理攻击中所起到的作用。在 DPSGD 算法中添加噪声过大会影响模型效用而过小又会影响隐私保护性能，这个问题在很大程度上限制了 DPSGD 的应用范围，因此有大量学者针对差分隐私的隐私效用权衡问题进行研究。Gong 等人^[25]提出了基于相关性分析的通用差分隐私深度学习框架，根据不同层中的神经元与模型输出之间的相关性来扰动梯度。具体地，在反向传播的过程中，与模型输出相关性较小的神经元的梯度会被添加更多的噪声，反之亦然。相较于在每个步骤中保持相同的梯度裁剪阈值和噪声方差来控制隐私成本，Du 等人^[26]则通过在每个步骤中动态调整剪裁阈值和噪声功率来减小性能损失差距，在强隐私保护区域显著提高了模型精度。Xu 等人^[27]则通过自适应学习速率提高收敛速度，显著降低了隐私开销；并通过引入自适应噪声，缓解了差分隐私对模型精度的负面影响。针对梯度裁剪和噪声添加不成比例的问题，Xu 等人^[28]根据裁剪偏差自适应地调整组中样本贡献，进而解决差分隐私造成的效用损失不等性问题。上述这些方法的整体优化思路基本上是一致的，都是通过某些策略来给有效的梯度添加更少的噪声，降低噪声对模型梯度的影响，进而提升模型的效果。但这些研究仍存在着复杂性较高、未充分利用历史梯度信息等问题，并且这些研究之间也存在一定的互斥性，难以互相结合使用。

1.2.2 隐私增强的联邦学习

联邦学习改善了隐私保护问题，因为用户的本地数据不会发送到中央服务器。然而恶意用户仍可以仅使用发送到服务器的本地梯度来重建用户的本地数据集^[8,29-30]，差分隐私可以解决该问题，但又会造成隐私效用权衡的问题。Gong 等人^[31]提出了一个基于差分隐私和同态加密的框架来保护参与 FL 模型的客户端的隐私，通过添加拉普拉斯噪声来保护客户端的梯度免受中央服务器的影响，但由于同态加密算法的计算复杂性过高，导致该框架效率过低。Kim 等人^[32]通过研究差分隐私与联邦学习结合过程中的隐私预算、效用以及通信之间的权衡问题，分析得出了客户端和服务器之间保证目

标隐私预算所需的最小高斯噪声方差, 但此方法依赖于过多假设, 在复杂的现实应用场景中往往难以直接使用。针对添加噪声与参与方数量成正比的问题, Liu 等人^[33]提出 FedSel 算法, 根据训练过程中的贡献选取 Top-k 维度, 并用梯度累积技术来稳定带有噪声的训练过程, 有效缓解了该问题。然而由于深度学习的不可解释性, 此方法难以估计合适的 k 值。Shi 等人^[34]针对现有的 DP 联邦方法对权重扰动鲁棒性较差并且使损失函数空间更难优化的问题, 提出了 DP-FedSAM 算法, 通过集成 SAM 优化器, 生成具有更好稳定性和权重扰动鲁棒性的局部平坦度模型, 从而提高了性能。Cheng 等人^[35]提出了“有界局部更新正则化”和“局部更新稀疏化”技术用来缓解由于限制本地更新的范数而造成联邦学习性能下降的问题, 在不牺牲隐私的情况下提高模型质量。然而这两个方法大大增加了算法的复杂度, 对联邦学习系统中节点有较高的要求。

1.2.3 去中心化联邦学习

针对传统中心化联邦学习中所存在的信任依赖、通信瓶颈和单点故障等问题, 去中心化联邦学习于 2018 年被提出, 其取消了可信第三方, 通过与相邻参与者之间进行模型参数的聚合和分发^[36], 进而有效解决了传统联邦学习中的信任依赖和通信问题。该方法在车联网^[37]、无线通信^[38]或无人机设备^[39]等场景中都具有较强的适用性。去中心化联邦学习方法的重点是如何传输由每个参与方本地的更新到其余的联邦节点。Chen 等人^[40]分析了联邦学习的进展, 包括去中心化联邦学习中的通信、隐私和安全要求等方面, 并且还提供了使用更通用算法的协作训练解决方案。Witt 等人^[41]提出了一套解决去中心化联邦学习中的隐私和安全问题的方案, 基于智能合约的奖励系统, 激励诚实的客户参与联邦训练流程。除了隐私方面, 王恺祺等人^[16]提出了一种去中心化联邦学习的可行方案, 能够同时解决联邦学习参与方的数据机密性问题和学习公平性问题。该方法提出一种基于区块链和联邦学习的生产-消费模型, 用来在模型安全聚合过程中审查参与者的本地行为, 并且在此基础上提出 APoS 共识机制, 提供一种激励与审查机制, 使参与者在训练过程中倾向于选择诚实的参与方进行协作。但这两个方法均基于分布式账本技术, 过高的计算成本和通信压力导致其很难大规模推广使用。Cyffers 等人^[42]提出了一种新的本地差分隐私松弛方法, 该方法是一种完全去中心化的协议, 训练过程受到特定的令牌控制, 令牌由接收它的设备顺序更新。在实现令牌之前, 每个节点都会在贡献中添加随机噪声以确保差分隐私。在获得最终模型之前, 此过程会重复 K (预定义值) 次, 该方法在实用性和隐私性之间实现良好的权衡。但该研究的一个大问题是它容易受到标签翻转和数据中毒攻击的影响。攻击者很容易渗透到网络中并破坏学习过程。Tran 等人^[43]提出了一个安全去中心化训练框架, 以保护参与训练节点的隐私信息。在训练过程中的每个时期, 参与方都会从参与方中选取一个主节点来计算全局梯度并将其发送到所有节点, 依此类推, 直到算法收敛, 该算法与其他大多数侧重于

分析分布式账本技术 (例如区块链技术)^[44]的去中心化联邦学习方法一样, 都大大增加了通信成本开销和计算成本开销。当前基于去中心化的联邦学习架构相关研究除了上述不足之外, 还都忽视了共识问题对去中心化联邦学习架构的影响。

1.3 研究内容

本文主要针对基于差分隐私的机器学习过程中所存在的隐私效用权衡问题, 以及去中心化联邦学习架构下的系统共识问题, 开展了基于卡尔曼滤波器的联邦学习优化研究。本文的主要研究内容从整体上看可分为如下两点:

1) 基于卡尔曼滤波的隐私保护机器学习优化方法研究。基于差分隐私的后处理性质, 将常用于差分隐私数据发布优化的卡尔曼滤波器, 迁移至机器学习场景下, 提出了基于卡尔曼滤波的梯度过滤方法用来过滤差分隐私梯度下降训练过程中的噪声, 以提升最终模型的精度。并且为了加快模型的收敛速率, 又引入了动量优化方法, 针对模型收敛后期的超调现象, 进一步提出了基于卡尔曼滤波的超调过滤方法, 在模型收敛的后期, 能够确保模型能够稳定的收敛。通过将这两种方法结合, 不仅能够大大加快模型的收敛速率, 同时能够提升模型最终收敛的精度。即在使用相同的隐私预算下, 得到更好的模型效果, 更好地实现了隐私和效用权衡。

2) 基于卡尔曼滤波器的隐私保护联邦学习算法优化研究。针对去中心化联邦学习架构下的系统共识问题, 本文提出了基于共识卡尔曼滤波器的去中心化联邦学习优化方法。具体地, 将研究点一所提出的基于卡尔曼滤波的梯度过滤方法和基于卡尔曼滤波的超调过滤方法中的卡尔曼滤波器扩展为共识卡尔曼滤波方法。该方法可以在去中心化的联邦学习架构下, 通过利用自身以及其相邻节点的数据信息用来进行模型的训练。两种共识卡尔曼滤波方法通过加快分散系统达成共识的速率, 有效解决了共识问题对训练过程的影响, 进而提升了系统的最终精度。

1.4 论文结构组织

本文主要开展基于差分隐私机器学习方法的隐私效用权衡研究和联邦学习架构优化研究, 共分为五章。具体的组织结构如图1-1所示。

第1章是绪论部分, 阐述了针对隐私效用权衡问题和联邦架构问题的研究背景和意义; 之后总结了近几年国内外关于隐私保护机器学习算法、安全增强联邦学习算法以及去中心化联邦学习架构的相关研究现状并分析当前研究中所存在的优势和不足之处; 最后阐述了本文的主要研究方向和重点研究内容。

第2章是相关理论与技术, 主要介绍了差分隐私、卡尔曼滤波器以及动量更新方法和超调的相关理论和技术。首先介绍了差分隐私的基本理论定义和性质, 重点介绍

了基于差分隐私的机器学习优化方法。其次，体系化地介绍了卡尔曼滤波器的数学原理及其工作机制，并介绍了适用于分散系统的卡尔曼共识滤波器的数学原理及应用场景。最后介绍了动量更新方法以及其所造成的超调问题，并给出了常见解决方案。

第3章是对基于卡尔曼滤波的隐私保护机器学习优化方法的详细阐述，首先给出问题定义并引出了在差分隐私优化场景下的梯度波动问题以及动量更新方法所引发的超调问题，然后针对这两个问题分别提出了基于卡尔曼滤波器的梯度过滤方法以及基于卡尔曼滤波器的超调过滤方法两个算法，并且将上述两种方法进行了有机结合，最后基于本章节提出的三个方法做了大量的实验分析。

第4章是对基于卡尔曼滤波器的隐私保护联邦学习优化方法的详细阐述，首先将第三章所提出的三个基于卡尔曼滤波器的优化方法扩展至联邦学习的场景下；其次针对传统联邦学习的问题进一步提出了去中心化的联邦学习架构，针对去中心化架构下的系统共识问题，本文提出了去中心化联邦学习架构下基于共识卡尔曼滤波器的隐私保护联邦学习优化方法，并进行了大量的分布式实验进行验证分析。

第5章主要是总结了本文所提出的适用于不同联邦学习架构下的两大方法，分析了当前工作所存在的不足之处，并指出了两个未来可继续优化的方向。

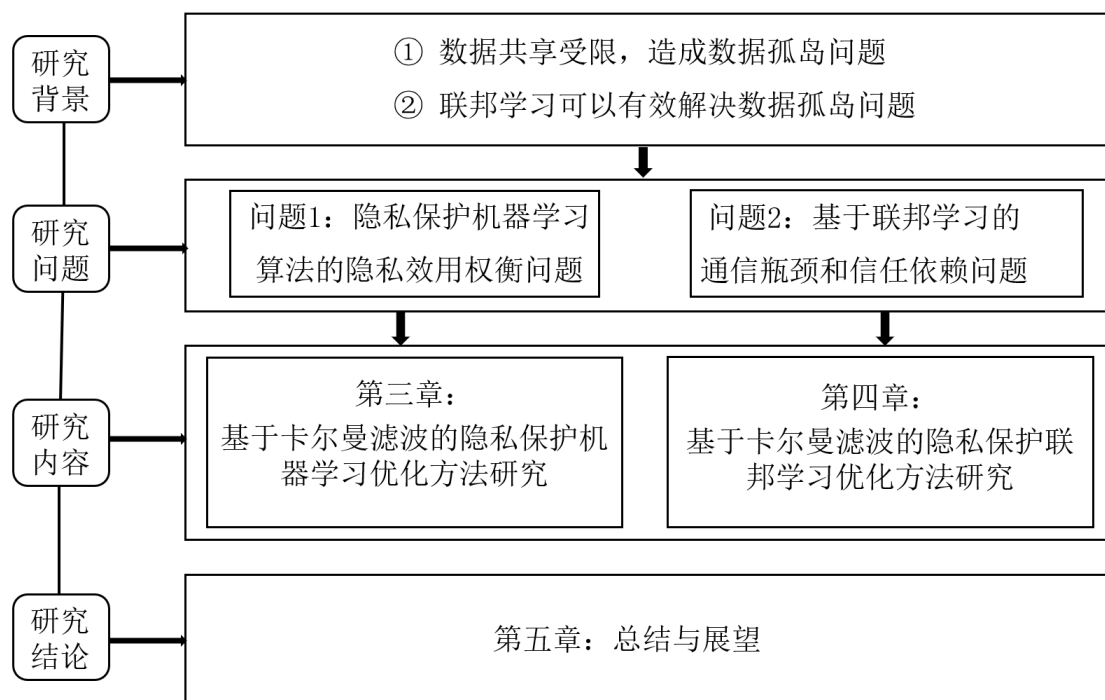


图 1-1 论文组织结构图

2 相关理论知识和技术

本章主要介绍与本文研究内容相关的理论知识和技术细节, 将用三个小节, 分别介绍差分隐私的相关理论基础、卡尔曼滤波器的理论基础以及动量优化方法和超调现象。

2.1 差分隐私理论基础

差分隐私 (Differential Privacy, DP)^[10]作为一种隐私保护技术, 提供了一种严格的数学隐私保证, 确保个体的隐私信息不会因为他们的数据参与分析而被泄露。差分隐私的核心思想是通过在原始数据中引入一定的随机扰动 (噪声) 来保护隐私。这种随机扰动使得分析结果对于单个个体的贡献变得不明显, 但是在整体上对统计结果的影响较小, 从而保护隐私的同时在一定程度上也确保了数据的效用性。差分隐私在数据发布以及机器学习训练的场景中被广泛地应用。从添加噪声位置的角度来讲, 常常将差分隐私分为本地差分隐私和全局差分隐私两类^[45], 具体区别如下:

1) 全局差分隐私^[46](Global Differential Privacy, GDP) 是一种在中央服务器进行隐私保护的方法, 其中数据处理方在对数据进行各类分析之前需要对其进行噪声处理。全局差分隐私的优势在于能够在整个数据集上提供更为一致的隐私保护, 但全局差分隐私需要信任数据处理方的合规性以及数据使用目的。

2) 本地差分隐私^[47](Local Differential Privacy, LDP) 是一种保护个体隐私的方法, 其中每个个体在本地对其自己的数据进行噪声处理, 然后再将处理后的数据提供给数据收集者或分析者。本地差分隐私的优势是个体对自己的数据有更多的控制权, 但也可能导致噪声的累积, 降低了分析结果的准确性。

差分隐私衡量了当某个个体的数据参与分析时, 对于其他个体的隐私泄露风险。通过控制噪声的强度, 可以进行数据隐私和效用的平衡。所添加的噪声越大, 则提供的隐私保护效果也就越好, 但数据效用性会越差; 但当添加噪声过小时, 由于隐私保护能力太差, 攻击者可能通过一定的方法从加噪后的数据中分析出有关个人的敏感信息。最近, Ren 等人^[48]便在梯度中添加小噪声的情况下, 成功恢复了原始数据集, 然而当噪声添加过大时又会影响数据的效用。因此, 在实践中如何确保隐私信息不被泄露的前提下, 尽可能地提升聚合后数据的效用性是差分隐私应用的一大难题。

2.1.1 差分隐私定义

将满足 DP 的机制 (或算法) 称为 ϵ -差分隐私^[10], 其中 ϵ 代表着隐私预算, 即代表着隐私保护程度的强弱, 隐私预算越大, 则隐私泄露的风险就越大。在提供 ϵ -差分隐私机制的定义之前, 先给出相邻数据集以及查询函数 f 敏感度的定义。

定义 2.1 (相邻数据集^[47]) 设 D^n 为所有数据集的域, 若 D, D' 只有一项不同且 $D, D' \in D^n$, 即从 D 中移除或者添加一个数据便得到的 D' , 则称 D, D' 为相邻数据集。

如定义2.1所示, 两个至多只差一条记录的数据集 D 和 D' 被称为相邻的数据集, 需要特别强调的是, D 与自身也相邻。接下来将给出敏感度的定义。

定义 2.2 (查询函数 f 的敏感度^[47]) 给定两个相邻数据集 D, D' , 以及一个将数据库映射到实数的查询函数 $f: D^n \rightarrow R^d$ 。则查询函数敏感度的数学形式为 $\Delta_f = \max_{D, D'} \|f(D) - f(D')\|_1$, 即针对相邻数据集的某个查询函数 f , 其最终的查询结果最多相差 Δ_f 。

基于相邻数据集和敏感度的定义, 具体的中心差分隐私的定义如下:

定义 2.3 (ϵ -差分隐私^[47]) 如果对于所有相邻数据集 $D, D' \in D^n$, 以及 $S \subseteq Y$, 其中 Y 是有可能输出的集合。若算法 M 满足 $Pr[M(D) \in S] \leq e^\epsilon Pr[M(D') \in S]$, 则称算法 M 满足 ϵ -差分隐私, 其中 ϵ 为隐私预算。

根据敏感度定义2.2以及中心化差分隐私的定义2.3, 通过简答的数学推导便可以分析得出, 当 $Pr[M(D) \in S]$ 与 $Pr[M(D') \in S]$ 二者均大于 0 时, 有如(2-1)等式成立:

$$e^{-\epsilon} \leq \frac{Pr[M(D) \in S]}{Pr[M(D') \in S]} \leq e^\epsilon \quad (2-1)$$

根据式(2-1)分析得知, 当隐私预算 ϵ 越低, 则在相邻两个数据集上查询到相同结果的概率就会越高, 该算法所提供的隐私保护程度也就越高。

根据差分隐私的定义2.3可以得知单次使用差分隐私机制的时候满足 ϵ -差分隐私, 然而现实应用时, 针对数据的处理往往十分复杂, 需要涉及多个环节步骤, 接下来将介绍一下差分隐私拥有三个良好特性, 即顺序组合性、并行组合性以及后处理性质。

定理 2.4 (顺序组合^[10]) 若 M_1 是一个 ϵ_1 -差分隐私机制, 而 M_2 是一个 ϵ_2 -差分隐私机制。则 M_1 和 M_2 二者的组合 $M_{1,2}$ 是一个 $(\epsilon_1 + \epsilon_2)$ -差分隐私机制。

基于顺序组合定理 2.4 可以在实际使用场景中更好地分析差分隐私预算的消耗情况。例如在机器学习场景下, 若在每个训练周期都实现了 ϵ_1 -差分隐私, 那么经过 k 个周期后根据组合定理, 该模型至少满足 $k\epsilon$ -差分隐私。

定理 2.5 (并行组合^[10]) 将一个数据集 D 分成 k 个集合, 分别为 D_1, D_2, \dots, D_k , 令 A_1, A_2, \dots, A_k 是 k 个分别满足 $\epsilon_1, \epsilon_2, \dots, \epsilon_k$ 的差分隐私算法, 则 $A_1(D_1), A_2(D_2), \dots, A_k(D_k)$ 的结果满足 $\max(\epsilon_1, \epsilon_2, \dots, \epsilon_k)$ 。

并行组合定理2.5的性质说明了, 当有多个差分隐私算法序列分别作用在一个数据集上多个不同子集上时, 最终的差分隐私预算消耗等价于算法序列中所有隐私预算的最大值。当需要进行多个数据处理任务时, 每个任务都可以采用各自单独的机制来实现差分隐私, 然后通过并行组合定理将它们组合在一起。这样可以在保证整个数据处

理流程的隐私保护性质的同时，又能够保持数据的可用性。

定理 2.6 (后处理性质^[10]) 给定任意一个算法 A_1 满足 ϵ_1 -差分隐私，对于任意算法 A_2 (其中 A_2 不一定是满足差分隐私的算法)，则有 $A(D) = A_2(A_1(D))$ 满足 ϵ_1 -差分隐私。

后处理定理2.6表明一旦数据得到了差分隐私的保护，便可以安全地对这些数据进行进一步的处理和分析，而不会破坏差分隐私的保护效果。后处理定理的作用是提供了一种灵活性和可扩展性，使得差分隐私可以与其他数据处理技术相结合。

2.1.2 差分隐私变种

本节将介绍差分隐私的两个变种 (ϵ, δ) -差分隐私^[49-50]定义以及 (α, δ) -差分隐私^[20]定义，最后介绍了几个常用的满足差分隐私的概率分布及实现机制。

1) (ϵ, δ) -差分隐私

尽管 ϵ -差分隐私最初的定义对数据隐私提供了强有力的隐私保护，但仍然没有严格处理由于组合而导致的隐私泄露。并且在训练联邦学习模型时也会出现组合问题，因为随着训练周期数的增加，隐私泄露也会增加。例如，如果我们应用在每个训练 epoch 都会有 ϵ 的隐私损失，那么在训练结束时，我们将导致 $k\epsilon$ 的隐私损失，其中 k 是训练期间的总时期数。为了减轻组合下的隐私泄露，Dwork^[49-50]提出了一个更为松弛的差分隐私定义： (ϵ, δ) -差分隐私。其目的是为了减轻组合下的隐私泄露。相较于 ϵ -差分隐私， (ϵ, δ) -差分隐私在组合的情况下提供较小的累计损失，其定义如下。

定义 2.7 ((ϵ, δ) -差分隐私^[49-50]) 如果对于所有相邻数据集 $D, D' \in D^n$ ，以及 $S \subseteq Y$ ，其中 Y 是所有可能输出的集合，若算法 M 满足式(2-2)：

$$Pr[M(D) \in S] \leq e^\epsilon Pr[M(D') \in S] + \delta \quad (2-2)$$

算法 M 有 $1 - \delta$ 的概率满足 ϵ -差分隐私，而有 $1 - \epsilon$ 的概率不满足。

(ϵ, δ) -差分隐私不适用 S 是单例集合的情况。与集合 S 的大小相比， δ 的大小应该可以忽略不计 (即 $\delta \ll 1/|S|$)，以避免总是侵犯数据集 δ 部分隐私的最坏情况。

2) (α, δ) -差分隐私

瑞丽差分隐私 (Rényi Differential Privacy)^[20]是差分隐私的另一种变种，Mironov^[51]以信息论中的瑞丽熵作为基础来衡量两个分布之间的差异，进而提供了更紧的隐私界限。 (α, ϵ) -RDP 的这种新变体可以更为准确地跟踪由于合成而导致的隐私泄露。除此之外， (α, ϵ) -RDP 提供了一种定量准确的方法来跟踪组合下的累积隐私泄漏。与传统的差分隐私相比，瑞丽差分隐私引入了一个平滑性参数 α 来控制隐私保护和数据可用性之间的

平衡。通过调整 α 的值，可以灵活地调整隐私保护级别和数据准确性之间的权衡。在定义 (α, ϵ) -RDP 之前，先给出瑞丽散度^[20]的定义：

定义 2.8 (瑞丽散度^[20]) 在 R 上定义的两个概率分布 P 和 Q ，阶数 $\alpha > 1$ 的瑞丽散度为：

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log E_{x \sim Q} \left(\frac{P(x)}{Q(x)} \right)^\alpha \quad (2-3)$$

其中 $P(x)$ 是 P 在 x 处的密度。这里的对数是自然对数， $x \sim Q$ 表示 x 服从分布 Q 。

如式(2-3)所示，瑞丽散度 (Rényi divergence)，也称为瑞丽熵 (Rényi entropy)，是信息论中的一个度量两个概率分布之间的差异的指标。其中， α 是一个大于 0 且不等于 1 的参数。当 α 趋近于 1 时，瑞丽散度趋近于 Kullback-Leibler 散度 (KL 散度)。

定义 2.9 ((α, δ) -差分隐私^[20]) 对于所有相邻数据集 $D, D' \in D^n$ ，以及 $S \subseteq Y$ ， Y 是所有可能输出的集合，若算法 M 满足下式(2-4)要求，则算法 M 满足 α 阶的 (α, δ) -差分隐私。

$$D_\alpha(M(D)\|M(D')) \leq \epsilon \quad (2-4)$$

据此可进一步得出，对于 (α, δ) -差分隐私有如(2-5)的不等式成立：

$$Pr[M(D) \in S] \leq e^\epsilon (Pr[M(D') \in S])^{\frac{\alpha-1}{\alpha}} \quad (2-5)$$

与先前定义的两种差分隐私所确定隐私界限的方法相比， (α, δ) -差分隐私针对组合定理而导致的隐私泄露问题，通过从瑞丽散度的角度分析两个分布之间的差距进而得到了更为严格的隐私界限。当阶数 α 接近于 1 时，瑞丽差分隐私接近于 ϵ -差分隐私；而当 α 取非常大的值时，瑞丽差分隐私则会提供更强的隐私保护功能。

3) 高斯机制

为了实现差分隐私保护，需要向数据中添加特定的噪声，以满足不同变种差分隐私的定义，高斯机制 (Gaussian mechanism) 是差分隐私保护中常用的一种噪音添加机制。它通过在计算结果上添加服从高斯分布的噪音来保护个体隐私。在基于差分隐私的机器学习优化方法场景中，一般都采用高斯噪声作为差分隐私噪声。

定义 2.10 (高斯机制) 给定一个查询函数 $f: D^n \rightarrow R^d$ ，其中 Y 是所有可能输出的集合且 $\epsilon > 0$ 。高斯分布为：

$$M(D) = f(D) + \mathcal{N}(0, \frac{\Delta_f^2}{\epsilon^2}) \quad (2-6)$$

如式(2-6)所示，其中 $\mathcal{N}(0, \frac{\Delta_f^2}{\epsilon^2})$ 是高斯分布，高斯分布 (也称为正态分布) 是一个连续概率分布，其具体的概率密度函数定义为：

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} M(D) = f(D) + \mathcal{N}(0, \frac{\Delta_f^2}{\epsilon^2}) \quad (2-7)$$

在式 (2-7) 中, μ 是分布的均值, σ 是分布的标准差。高斯分布的特点是在均值附近有较高的概率密度, 而在标准差越大, 分布越平缓。高斯机制在保护个体隐私的同时, 仍然能够提供合理的数据查询结果。它在差分隐私领域得到了广泛的应用。需要注意的是, 为了保护隐私, 需要根据敏感性和隐私预算来选择适当的噪音标准差。

2.1.3 基于差分隐私的随机梯度下降方法

在机器学习场景下为了防止个体数据隐私信息的泄露, 常见的方法便是 DPSGD 算法^[19]。在 DPSGD 算法训练的过程中, 通过往模型的梯度中添加高斯噪声, 进而对模型提供差分隐私保护, 以抵御各类推理攻击^[52-53]。常见的 DPSGD 算法一般都分为四步, 即计算梯度、梯度裁剪、添加噪声以及进行梯度下降。梯度裁剪的目的是为了防止某些样本的梯度过大进而增大隐私泄露的风险, 一般来讲, 某个样本的梯度越大, 则说明该样本此步训练中的作用越大, 也更容易泄露自身的隐私信息。通过设置裁减系数, 对所有梯度进行 l_2 裁剪, 以限制梯度下降过程中的敏感度。添加高斯噪声的目的是提供差分隐私保护, 通常添加的噪声为高斯噪声。其中最终加入到梯度中的高斯噪声的方差大小和裁减系数以及噪声方差两个相关。具体的 DPSGD 算法如下:

算法 2-1 DPSGD 算法

输入: 样本 x_1, x_2, \dots, x_n , 损失函数 $\mathcal{L}(\theta) = \sum_{i=1}^n \mathcal{L}(\theta, x_i)$, 学习率 η , 噪声方差 σ , 批大小 L , 梯度裁剪系数 C

输出: θ_T , 以及隐私损耗 (ϵ, δ)

- 1 随机初始化模型参数 θ_0 ;
- 2 **while** $t < T$ **do**
- 3 以 L/N 的概率从全体样本中随机抽取样本集合 L_t ;
- 4 计算梯度;
- 5 针对 L_t 中的每个样本, 计算其梯度 $g_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$;
- 6 梯度裁剪;
- 7 $\bar{g}_t(x_i) \leftarrow g_t(x_i) / \max(1, \frac{\|g_t(x_i)\|_2}{C})$;
- 8 添加噪声;
- 9 $\tilde{g}(t) \leftarrow \frac{1}{L} (\sum_i \bar{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I))$;
- 10 梯度下降;
- 11 $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{g}(t)$;
- 12 **end**

如算法 2-1 所示, 首先进行的样本采样, 以实现隐私放大的效果, 进而节省隐私预算; 第二步便是基于采样的样本单独计算每个样本的梯度信息; 第三步对每个样本的梯度信息根据全局敏感度的大小进行梯度裁剪, 对整体的梯度进行放缩; 之后便是根据相应的隐私预算大小, 给梯度信息中添加合理的噪声, 并进行平均, 需要注意的是, 先给每个样本的梯度添加相应的高斯噪声之后然后对所有的梯度进行平均, 这一步平

均会大大降低真实梯度中所包含的高斯噪声方差。最后便是基于平均后的包含高斯噪声的梯度信息，进行梯度下降。按照此种流程，不断循环，直至模型最终收敛。

上节提到，传统的 ϵ -差分隐私受其组合定理的影响，进行 k 次梯度下降，则最终的隐私预算消耗是 $k\epsilon$ 。然而一般的机器学习模型，尤其是复杂的深度模型的收敛需要较高的迭代次数。若添加较小的噪声，这就需要更高的隐私预算消耗，大大削弱了隐私保护强度；若添加的噪声过大又会很大程度上影响模型最终的效果，甚至会造成模型不收敛的情况。不过，幸运的是传统的组合定理的隐私边界是十分宽松的。Martín 等人^[19]基于矩设计了矩量统计方法，通过跟踪隐私损失的详细信息 (更高矩)，进而从渐进和经验两个维度获得对整体隐私损失更为严格的估计，并且进行了大量的实验，其实验结果证明，在隐私预算合理的情况下 (隐私预算控制在 10 以内)，模型能达到和 SGD 近似相同的效果，使得 DPSGD 具有较高实践实用价值。除此之外，Martín^[51]等人基于瑞丽散度，提出了 (ϵ, δ) -RDP 瑞丽差分隐私的概念。相较于矩量统计方法，瑞丽差分隐私对隐私预算损失进行了更为严格的估计。文章并给出了 (ϵ, δ) -RDP 和 (α, ϵ) -DP 之间相互转换的方法。本文采用的隐私预算计算方式均为瑞丽差分隐私。

2.2 卡尔曼滤波理论基础

2.2.1 卡尔曼滤波器

卡尔曼滤波器 (Kalman Filter)^[54]是一种用于估计包含统计噪声的动态过程的有效算法。一个带有噪声的底层动态过程可以通过线性时变模型 (又名过程模型) 来表示:

$$r(t+1) = A(t)r(t) + \omega(t) \quad (2-8)$$

如2-8所示，其中 $r(t)$ 是在时间戳为 t 时的状态 ($r(0)$ 是满足正态分布 $\mathcal{N}(\bar{r}(0), P_0)$ 的一个初始化状态), $\omega(t)$ 是满足正态分布 $\mathcal{N}(0, Q_t)$ 的噪声，其中 Q_t 为系统方差，即系统自身所存在的波动，而 $A(t)$ 是描述过程转换的转换矩阵。

在一个分布式网络中，每个节点 i 都可以通过以下线性传感器 (又名测量模型) 对动态过程进行观测，得到观测值 x_i ，具体观测方程如下：

$$x_i(t+1) = H_i(t)r(t) + v_i(t) \quad (2-9)$$

在式(3-8)中 $H_i(t)$ 是观测矩阵， $v_i(t)$ 是观测噪声，且观测噪声满足正态分布 $\mathcal{N}(0, R_t)$ 的假设。卡尔曼滤波器可以通过计算得到系统状态 $r(t)$ 的最优估计。假设 $\hat{r}_i(t)$ 和 $\bar{r}_i(t)$ 为节点 i 的后验估计和先验估计，则系统状态 $r(t)$ 的最优估计 $\hat{r}_i(t)$ 可以由先验估计 $\bar{r}_i(t)$ 和观测值 $x_i(t)$ 的线性组合得到，具体过程如下：

$$\hat{x}_i(t) = \bar{x}_i(t) + K_i(t)(x_i(t) - H_i(t)\bar{x}_i(t)) \quad (2-10)$$

式(2-10)中 $K_i(t)$ 被成为卡尔曼增益, 其用于最小化每个时间戳的后验误差的协方差。在实践中, 可以根据过程模型和测量模型来预测先验估计 $x_i(t)$ 。

标准卡尔曼滤波器仅适用于单独生成每个节点的真实状态 $r(t)$ 的估计。尽管如此, 所有 m 个节点都测量2-8中描述的相同动态过程。一旦它们的测量在网络之间共享, 它们的估计就可以更好地校准。具体的卡尔曼滤波算法如下:

算法 2-2 卡尔曼滤波器

输入: 观测数据 $x_i(t)$, 观测矩阵 $H_i(t)$, 初始化方差 $P_i(0) = P_0$, 测量方差 $R_i(t)$, 总过程次数 T
输出: $\hat{x}_i(T)$

- 1 初始化设置;
- 2 **while** $t < T$ **do**
- 3 计算卡尔曼系数;
- 4 $K(t) = P(t)H_i^T(R + H_i(t)P(t)H_i(t)^T)^{-1}$;
- 5 更新状态估计;
- 6 $\hat{x}_i(t) = \bar{x}_i(t) + K_i(t)(x_i(t) - H_i(t)\bar{x}_i(t))$;
- 7 更新状态协方差;
- 8 $M(t) = P(t) - P(t)H_i^T(R + H_iP(t)H_i^T)^{-1}H_i(t)P(t)$;
- 9 预测状态的协方差;
- 10 $P(t+1) = AM(t)A + Q(t)$;
- 11 预测系统的状态;
- 12 $\bar{x}_i(t+1) = A(t)\hat{x}_i(t)$;
- 13 **end**

如算法 2-2 所示, 卡尔曼滤波器的过程整体上可以分为两个主要步骤: 更新和预测。更新部分主要由三部分组成, 主要是计算卡尔曼系数、更新状态估计以及更新状态协方差。其中卡尔曼系数是用于融合测量值和预测状态的权重, 它基于预测状态的协方差和测量方程 (通常是线性的) 的协方差来确定。然后使用卡尔曼增益和测量值来计算更新后的状态估计值, 这是预测状态和测量值的加权平均。最后使用卡尔曼增益来计算更新后的状态协方差矩阵, 它反映了更新后状态估计的不确定性。通过不断地进行预测和更新步骤, 卡尔曼滤波器可以根据测量值和先前的状态估计, 递归地提供对系统当前状态的最优估计。预测部分主要分为两步, 其分别为: 预测系统的状态和预测状态的协方差。具体地, 预测系统的状态主要是通过使用系统状态转移方程在没有测量的情况下的预测系统的下一个状态。而预测状态的协方差则是使用系统的状态转移方程和先前状态的协方差估计来计算预测状态的协方差矩阵, 它反映了状态估计的不确定性。需要注意的是, 卡尔曼滤波器的有效性要求系统满足线性动态和高斯噪声假设, 并且对系统的建模和参数估计需要准确。在实际应用中, 卡尔曼滤波器常用于估计具有线性动态的系统, 例如目标跟踪、导航和姿态估计等领域。

2.2.2 卡尔曼信息共识滤波器

随着物联网和分布式传感器网络的发展，越来越多的系统由多个分布式节点组成，这些节点数量众多、且通信频繁。由于这些分布式传感器的数据往往存在一定噪声，因此为了更好地分析这些数据，需要对这些数据进行过滤处理，以增加数据的效用性。传统的中心化滤波方法往往是将所有数据传输到中心节点进行统一处理，然而这种统一处理的方法可能会导致分布式系统出现高延迟、高能耗和通信负载过大的问题。除此之外，在一些应用场景中，节点的观测数据可能包含敏感信息。传统的中心化滤波方法要求所有节点将其观测数据传输到中心节点，可能会引发隐私泄露的风险。而分布式的卡尔曼滤波可以在保护节点观测数据隐私的同时，实现对整个系统状态的估计，为隐私保护提供了技术支持。因此，研究分布式的卡尔曼滤波可以更好地适应分布式系统的需求。卡尔曼共识信息过滤器 (Kalman-Consensus Information Filter, KCIF)^[55]是卡尔曼滤波器的分布式形式，可以在分布式系统中更好地完成对系统的最优估计 $r(t)$ 。通过卡尔曼共识滤波器，多个智能体可以通过信息交换和融合，实现对系统状态的一致估计，提高整个系统的状态估计精度和鲁棒性。这在分布式感知、分布式控制和协同定位等领域具有广泛的应用。并且，除了标准卡尔曼估计器操作之外，针对分散系统中存在的系统共识问题，通过给本地先验估计时添加一个共识项，进而有效加快了所有节点之间达成共识的速率。卡尔曼共识信息过滤器的具体聚合过程为：

$$\hat{x}_i(t) = \bar{x}_i(t) + M_i(t)(y_i(t) - Y_i(t)\bar{x}_i(t)) + C_i(t) \sum_{j \in N_i} (\bar{x}_j(t) - \bar{x}_i(t)) \quad (2-11)$$

如式(2-11)所示，其中 $y_i(t)$ 和 Y_i 分别是 i 的邻居节点的加权测量和信息矩阵， N_i 是指节点 i 的一跳邻居集合， $M_i(t)$ 是后验估计协方差， $C_i(t)$ 是共识增益，它保持分布式卡尔曼估计量的共识和稳定性之间的平衡。由式(2-11)可得，相较于传统的卡尔曼滤波器，共识卡尔曼滤波器多了一个共识项 $C_i(t)$ ，以确保在分散系统中，除了过滤高斯噪声外，还可以加快系统达成共识的速率。

如算法2-3所示，与传统卡尔曼滤波器不同的点在于，卡尔曼共识滤波器额外增添了信息交换部分和共识增益计算部分。其中信息交换部分是卡尔曼滤波器将自己的信息打包然后发送给所有与自己相邻的节点，同时也接受所有自己邻居节点的信息数据，以此种方式来进行系统的聚合。除此之外，除了包含用于过滤高斯噪声的过滤部分外，共识卡尔曼滤波器还额外增添了共识部分，该部分的目的是加快分散系统中各个参与方在分散系统中达成共识的速率。当若共识增益系数为 0 时，共识卡尔曼滤波器退化为传统的卡尔曼滤波器。与传统的集中式卡尔曼滤波相比，卡尔曼共识信息滤波器在分散的系统中，和其邻居节点互相交换信息，这种分布式计算可以减少通信开销，提高系统的估计效率。针对真实场景中节点之间的通信可能会受到延迟、丢包或节点故障

等不稳定因素的问题，卡尔曼共识信息滤波器通过节点之间的信息交换和协作，实现了对信息的共识处理，从而提高了系统的鲁棒性和容错性。即使某个节点出现故障或通信中断，其他节点仍然可以通过共识机制来保持一致性。

算法 2-3 卡尔曼共识滤波器

输入：观测数据 $x_i(t)$, 观测矩阵 $H_i(t)$, 初始化方差 $P_i(0) = P_0$, 测量方差 $R_i(t)$, 总过程次数 T

输出：后验聚合估计 $\hat{x}_i(T)$

```

1 initialization;
2 while  $t < T$  do
3   计算聚合信息:  $z_i(t) = f(D_{i,t})$ ;
4   广播信息并接受信息:  $u_i(t) = \frac{H_i(t)z_i(t)}{R_i(t)}$ ,  $U_i = \frac{(H_i(t))^2}{R_i(t)}$ ;
5   聚合信息:  $msg_i(t) = (u_i(t), U_i, \bar{x}_i(t))$ ;
6   计算后验估计方差:  $M_i(t) = (P_i(t)^{-1} + H_i(t)R(t)^{-1}H_i(t)^T)^{-1}$ ;
7   计算共识增益系数:  $C_i(t) = \gamma P_i(t)$ ;
8   计算后验估计:
       $\hat{x}_i(t) = \bar{x}_i(t) + M_i(t)(y_i(t) - H_i(t)R(t)^{-1}H_i(t)^T\bar{x}_i(t)) + C_i(\sum_{j \in N_i}(\bar{x}_j(t) - \bar{x}_i(t)))$ ;
9   更新先验估计:  $\bar{x}_i(t+1) = A(t)\hat{x}_i(t)$ ;
10  更新先验估计方差:  $P_i(t)^+ = AM_i(t)A^T + BQB^T$ ;
11 end

```

2.3 动量优化方法和超调现象

2.3.1 动量优化方法

在深度学习模型的训练过程中，还有一种常见的技术能大大提升模型的训练速度，它就是动量优化方法^[56]。动量优化方法常被作为高阶优化器方法的子方法 (例如 Adam 和 RMSprop^[56]等)，被广泛应用于深度学习模型的训练中。在传统的随机梯度下降算法中，参数的更新是根据当前的梯度直接进行的。然而这种方法存在的问题是，当梯度的方向变化较快时，会导致参数在优化过程中来回震荡，不仅会降低模型收敛速率的同时，还会使得模型难以得到全局最优解。动量优化方法通过引入动量来缓解这个问题。动量是基于历史梯度的指数加权平均，它使得参数更新不仅依赖于当前梯度，还受到之前梯度的影响。因此动量方法可以在梯度变化剧烈时减缓更新的速度，而在梯度方向一致时加快更新的速度，进而加速模型的收敛过程，同时还可以帮助参数跳出局部最优点，动量优化方法的参数更新规则如下：

$$V_{t+1} = \alpha V_t - \gamma \partial L_t / \partial \theta_t \quad (2-12)$$

如式(2-12)所示，其中 V_t 是历史梯度的累积， $\alpha \in (0, 1)$ 是动量系数。通过一定的数学转换，可以将2-12中的 V_t 移除，故此时模型训练过程中的更新规则如下：

$$\theta_{t+1} = \theta_t - \eta \sum_{i=0}^{t-1} \partial L_i / \partial \theta_i \alpha^{t-i} \quad (2-13)$$

由式 (2-13) 可以得出, 参数的更新依赖于过去梯度的积分 ($\sum_{i=0}^{t-1} \partial L_i / \partial \theta_i \alpha^{t-i}$) 和当前梯度 $\partial L_t / \partial \theta_t$ 。随着训练的进行, 历史梯度的误差也会随之增大, 两个梯度相隔越远, 则其偏差越大。因此通过引入动量系数 α , 来增加最新梯度的权重占比, 减少动量中所包含的误差。动量优化方法的效果主要体现在两个方面。首先, 它可以加速训练的速度, 特别是在参数空间中存在平坦区域时, 可以更快地跳出局部最小值, 找到全局最优解。其次, 它可以增加训练的稳定性, 减少参数更新的震荡, 避免陷入局部最小值或鞍点。它在深度学习中被广泛使用, 对于解决复杂问题和提高模型性能具有重要作用。

2.3.2 动量超调问题及解决方案

尽管动量优化方法通过累积历史梯度可以加速训练过程, 但如果权重改变其下降方向, 历史梯度将滞后于权重的更新。这种由历史梯度引起的现象称为超调 (其英文名称为 *OverShoot*)。超调问题的出现是动量积累的结果。当模型在参数空间中接近最优解时, 由于历史梯度的影响, 动量会继续积累, 并且可能增大参数更新的步伐。这可能导致模型跳过最优解并继续在参数空间中震荡, 无法稳定地收敛。解决超调问题的一种直观的方法是调整动量因子 α 的值。较小的 α 值可以减小动量的积累效果, 从而减小参数更新的步伐; 而较大的 α 值则可以增加动量的积累效果, 加快参数更新的速度。通过合理选择 α 的值, 可以平衡动量的作用, 在一定程度上缓解超调问题的发生。另外, 超调问题也可能与学习率的选择有关。较大的学习率可能会导致参数更新的步伐过大, 进而加剧超调现象的抖动。因此, 合理选择学习率并结合动量因子进行调整, 可以在一定程度上有效地缓解超调问题。然而上述两种方法均属于经验值, 基于不同的数据以及不同的模型其适用的范围都有所不同, 需要大量实验进行验证。然而 DPSGD 重复训练的成本是高昂的, 故在差分隐私的场景下并不适用。针对动量优化方法中的超调问题, Wang 等人^[57]中给出了系统化的解决方案。文中基于 PID 反馈控制器对动量方法进行了优化。PID 反馈控制器的工作原理是通过不断调整控制器输出来使系统的输出值与期望值尽可能接近。具体来说, 控制器根据系统的实际输出值和期望值计算出误差, 然后根据比例、积分和微分的原则调整输出。这个过程不断迭代, 直到系统的输出值稳定在期望值附近。其具体的数学形式为:

$$\mu(t) = K_p e(t) + K_i \int_0^t e(t) dx + K_d \frac{de(t)}{dt} \quad (2-14)$$

如式(2-14)中所示, 其中 K_p 为比例项, 其根据系统误差的大小, 以 K_p 的倍数调整控制器的输出, 这部分控制器的作用是使系统对误差的响应更快, 但可能会导致超调

和震荡； K_i 为积分项，其根据系统误差的累计值，以积分系数 K_i 的倍数调整控制器的输出，这部分控制器的作用是消除系统的稳态误差，使得系统的输出更接近期望值；而 K_d 为微分项，其根据系统误差的变化率，以微分系数 K_d 的倍数调整控制器输出。这部分控制器的作用是抑制系统对误差变化的敏感度，减小系统的超调和震荡。PID 反馈控制器的作用是实现对系统的稳定控制和优化。它可以通过自动调节控制器输出来使系统的输出值尽可能接近期望值，An 等人^[57]将随机梯度下降等价于 P ，而将动量优化方法等价于 PI ；通过引入未来梯度的噪声，采用 PID 对动量方法进行了优化。随着训练的进行，若梯度的值在变小但仍未反向，此时说明在此方向模型已经逐渐逼近至局部最优点，然而动量方法仍然会继续朝着该梯度方向累计，直至梯度反向，如此反复直至最终收敛，这便是导致超调问题出现的主要原因。针对该问题，PID 方法通过引入相邻两次梯度的差值 $D_{t+1} = \alpha D_t + (1 - \alpha)(\frac{dL_t}{d\theta_t} - \frac{dL_{t-1}}{d\theta_{t-1}})$ 作为微分项，以提前感知梯度变化的趋势，减轻其在局部最优点附近的波动，缓解超调问题。An 等人^[57]所提出的 PID 优化器方法具体的更新过程如下：

$$\theta_{t+1} = \theta_t + V_{t+1} + K_d D_{t+1} \quad (2-15)$$

从等式(2-15)可以得出，与动量优化方法相比，PID 优化器方法额外引入了一个超参数 K_d ，用两次梯度的差值等价于系统的微分项，进而在动量累计放缓时加速放缓的过程，在动量累计加速的时候加快加速的过程。同时文章通过采用拉普拉斯变换理论和齐格勒-尼科尔斯方法^[58]进而可以很好地初始化超参数 K_d 。

2.4 本章小结

本章首先介绍了差分隐私的相关理论基础，给出了三种不同的差分隐私定义方式，并重点介绍了基于差分隐私的随机梯度下降方法。其次本章介绍了卡尔曼滤波器以及共识卡尔曼滤波器的相关理论基础，并阐述了其具体的工作流程。最后介绍了动量优化方法以及其造成的超调问题，并介绍了超调问题常见解决方案。

3 基于卡尔曼滤波器的隐私保护机器学习优化方法研究

基于差分隐私的联邦学习训练过程由多个差分隐私机器学习优化过程组成，故本章主要针对差分隐私机器学习优化过程进行优化。DPSGD 方法^[19]通过在训练过程中添加噪声，进而对隐私数据提供更强的保护，以抵御推理攻击等各种攻击方式^[52]。然而噪声会造成模型效用的损失，引发了隐私和效用权衡的问题。为了优化隐私和效用之间的权衡问题，本章将针对基于差分隐私的机器学习优化方法中存在的“梯度波动”和“动量超调”两个问题，提出了基于卡尔曼滤波器的梯度过滤算法和基于卡尔曼滤波器的超调过滤算法，并且二者可以结合形成基于卡尔曼滤波器的隐私保护机器学习优化方法，这三个方法均可以在隐私保护程度不变的前提下，加快 DPSGD 算法的收敛速率并使模型最终收敛到更高的精度，更好地实现了隐私效用之间的权衡。

3.1 系统模型及问题定义

3.1.1 系统模型

本节将对基于差分隐私的机器学习优化过程进行建模。设参与训练的样本为 $(x_j, y_j) \in \mathcal{N}$ ，其中 $j \in (1, n)$ ，即数据集 \mathcal{N} 中共包含有 n 个样本， x_j 为高维的特征信息，而 y_j 为标签信息，每个样本中的特征信息和标签信息均属于隐私信息。模型的参数为 θ ，机器学习任务的损失函数为 \mathcal{L} ，模型训练的目标如下：

$$\theta^* \triangleq \operatorname{argmin} \mathcal{L}(\mathcal{N}; \theta) \quad (3-1)$$

如式(3-1)所示，模型训练的目标是寻找使得损失函数在全量数据集上平均值最小的最优模型参数 θ^* 。在本文 2.1.3 节已经对 DPSGD 算法的训练过程进行了详细的介绍，在基于 DPSGD 的模型训练过程中，每一步计算的梯度除了随机梯度外还额外包含了高斯噪声 σ_{dp} ，设每一步随机抽取的样本数量为 B ，即样本批大小为 B ，则训练过程中每一步的梯度为 $g_{dpsgd}(t) = \frac{1}{|B|} \sum_{x \in B} \nabla_{\theta} \mathcal{L}(\theta, x) + \sigma_{dp}$ ，其中 $\frac{1}{|B|} \sum_{x \in B} \nabla_{\theta} \mathcal{L}(\theta, x)$ 为基于此批样本所计算出的随机梯度信息，而 σ_{dp} 则是为满足差分隐私而添加的高斯噪声，基于差分隐私的机器学习优化过程便是不断沿着 $g_{dpsgd}(t)$ 方向向最优点不断进行靠近的过程。

为了加快模型的训练速率，可以将动量优化方法^[56]和 DPSGD 算法结合，来协同进行模型的训练。将二者结合后的动量累计过程如下：

$$V_{t+1} = \alpha V_t - \gamma g_{dpsgd}(t) \quad (3-2)$$

如式(3-2)所示, 动量 V_t 是基于历史梯度和差分隐私噪声的指数加权平均, 参数的更新依赖于过去梯度的累计 ($\sum_{i=0}^{t-1} g_{dpsgd}(i)\alpha^{t-i}$) 和当前梯度 $g_{dpsgd}(t)$ 。

3.1.2 问题定义

本节将对基于差分隐私的机器学习优化方法中的两个问题进行系统化定义。一是基于差分隐私机器学习优化方法的梯度波动问题, 即差分隐私噪声的引入导致梯度波动过大进而影响模型收敛精度; 二是基于动量更新和差分隐私的超调问题, 即差分隐私方法的引入加剧了训练过程超调现象, 进而影响模型的收敛速度和精度。

1) 基于差分隐私机器学习优化方法的梯度波动问题

在基于差分隐私的机器学习优化过程中, 与机器学习一样, 其优化目标 \mathcal{L} 都是使平均损失函数值在全部样本空间 \mathcal{N} 上达到最小, 损失函数的具体数学形式如下:

$$\mathcal{L} = \frac{1}{|\mathcal{N}|} \sum_{x \in \mathcal{N}} \nabla_{\theta} \mathcal{L}(\theta, x) \quad (3-3)$$

如式(3-3)所示, 损失函数 \mathcal{L} 为在全体样本上损失的平均值, 具体的函数形式根据任务的不同而有所区别。模型训练的过程便是利用随机梯度下降等一系列方法来不断降低损失函数值, 当损失函数达到最小值时, 模型便完成了收敛。随机梯度下降及其变种是当前最常用的模型训练方法, 其主要思想是每次采样一部分样本的平均梯度作为全局梯度的无偏估计并基于此更新。但基于差分隐私的机器学习优化方法由于在每一步都添加了高斯噪声, 这会影响每一步的梯度的准确性, 进而影响优化过程。

$$g_{dpsgd}(t) = g_{gd} + \sigma_{dp} + \sigma_{sgd} \quad (3-4)$$

如式(3-4)所示, 在基于差分隐私的机器学习模型训练过程中所使用的梯度 g_{dpsgd} , 除了全局梯度 g_{gd} 外, 还包含了 σ_{dp} 和 σ_{sgd} 两种噪声。其中 σ_{sgd} 是由于使用的是随机梯度下降技术, 梯度中会包含随机采样所带来的噪声, 该噪声的均值为 0, 方差大小和采样率有关, 采样率为 1 时, $\sigma_{sgd} = 0$ 。而噪声 σ_{dp} 则是使用了差分隐私技术所主动添加的高斯噪声, 该噪声均值为 0, 方差大小和隐私预算相关, 隐私预算越小, 则噪声方差越大。这两种噪声大大影响了梯度的准确性, 使得模型训练过程中无法向着最优方向前进。

为了更为直观地观测模型训练过程中梯度的变化曲线, 本文随机选取两个模型训练过程中的参数梯度变化曲线并绘制出来 (其中一个为 CNN 模型, 使用 CIFAR10 数据集进行训练, 一个为两层 MLP 模型, 使用 MNIST 数据集进行训练)。随机选取模型中某个参数的梯度, 然后绘制出其更新过程中每一步的全局梯度的变化曲线以及基于 DPSGD 算法所求得的梯度的变化曲线。如图3-1所示, 其中 DPSGD 为添加了基于差分

隐私的随机梯度下降过程中的梯度变化曲线，而 BGD 为使用全局梯度下降的梯度变化曲线。根据图中曲线可知，由于 BGD 是使用全局样本计算出的梯度，故相较于随机梯度其更为稳定，其变化曲线平稳而光滑，并且随着迭代次数的不断增加，梯度的值也逐渐趋于 0，也意味着随着模型的不训练，模型逐渐收敛。DPSGD 由于其中包含了随机采样的梯度噪声以及额外的高斯噪声，导致梯度在全局梯度上下一直波动，但是由于添加的两个噪声均是偏差为 0 的高斯噪声，故梯度在整体上仍然是无偏的。且随着添加高斯噪声的方差增大，梯度的波动也会逐渐变大。

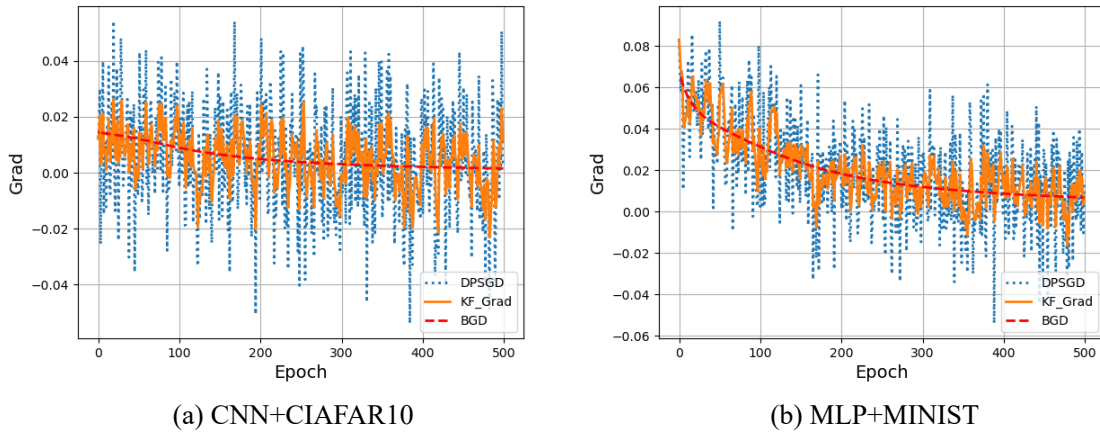


图 3-1 梯度变化图像

2) 基于差分隐私和动量更新的超调问题

在本文的 2.3.2 节中，已经对超调问题进行了系统化的定义并解释了其出现的原因。当机器学习优化方法中通过引入动量优化方法以加快模型训练过程时，随着动量的累计在模型收敛后期会引发超调现象，超调现象大大降低了模型的收敛速率以及最终的收敛精度。在离散时间控制系统^[59]中将由动量更新所引发的超调过程定义为“从系统期望响应测得的响应曲线的最大峰值”。在数学上，它被定义为：

$$overshoot = \frac{\theta_{max} - \theta^*}{\theta^*} \quad (3-5)$$

式(3-5)中的 θ_{max} 和 θ^* 分别是训练过程中参数的最大值和最优值。

在基于差分隐私的机器学习优化过程中，每一步随机梯度都被额外的添加了高斯噪声，并且动量方法是通过不断累计过去的梯度用来形成动量，所以差分隐私噪声也会随着一起不断地进行累计。如果 DPSGD 训练过程中每一步添加的噪声方差为 σ ，且动量系数 $\alpha \in (0, 1)$ ，则 $t+1$ 轮后累计高斯噪声 \mathcal{G}_{t+1} 为：

$$\mathcal{G}_{t+1} = \alpha * \mathcal{G}_t + \sigma \quad (3-6)$$

公式(3-6)通过一定数学变换，便可以得到在动量更新场景下的噪声累计通式：

$$\mathcal{G}_t = (\alpha^{t-1} - 1) \frac{\sigma}{\alpha - 1} \quad (3-7)$$

由等式 (3-7) 可以得出，当使用动量优化进行动量更新时，噪声会在一定程度上不断的出现累计 ($\mathcal{G}_t \geq \sigma$)，不过噪声的方差并不会出现无限扩大的情况，其噪声最高累计值为 $\frac{\sigma}{\alpha-1}$ 。动量优化方法在 DPSGD 的场景下由于引入了更多额外的高斯噪声，往往会使得超调现象加重。并且在差分隐私的场景下，传统的用于解决超调问题的 PID 方法也可能不再适用。这是因为 PID 方法的主要思想是利用积分项 $D_t = \frac{d\mathcal{L}_t}{d\theta_t} - \frac{d\mathcal{L}_{t-1}}{d\theta_{t-1}}$ 这个相邻两次之间的差值来代表梯度变化的趋势，进而缓解超调现象，但由于每次梯度都包含了高斯噪声 σ ，故此时 PID 方法的积分项 D_t 会额外引入两倍的高斯噪声 2σ 。额外噪声的引入不仅可能无法缓解超调现象甚至还会加剧超调现象。

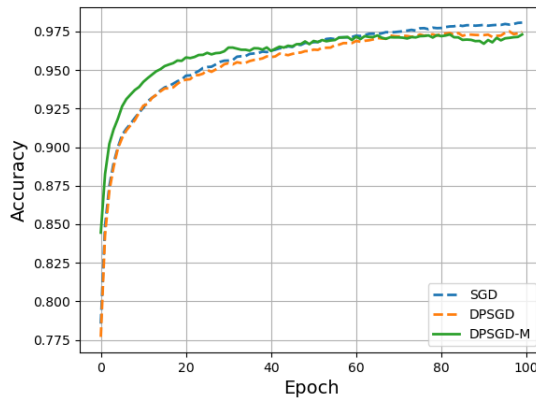


图 3-2 超调现象结果图

如图3-2所示，图像中共有三组不同的实验结果，该实验采用的是 MINIST 数据集及两层神经网络算法，三组实验分别采用 SGD、DPSGD 以及基于动量更新方法的 DPSGD 算法，这三组实验其他的所有超参数完全相同。图中的三种方法，SGD 算法的效果最好，这是因为其并未添加高斯噪声，使用的是真实的随机梯度进行训练，故其效果最优；其次 DPSGD 算法由于其高斯噪声的引入，给模型的收敛过程带来了部分的波动，但整体上模型仍然收敛到一个较好的精度，这是隐私和效用权衡的结果，也证明了差分隐私在机器学习的模型训练过程中有一定的可用性。DPSGD-M 算法由于动量的引入，相较于其它两组更新方法，其大大加快了模型的收敛速率；然而在模型的收敛的过程中尤其是模型收敛的后期由于由于动量不断累计高斯噪声，故模型的效果一直在不断地抖动，这与前文的理论分析结果是一致的，即动量更新方法会不断增加高斯噪声的引入，但是其不会无限增大，高斯噪声方差其有一个上限值，模型收敛的后期会无限接近于这个上限值。这便是超调现象，这大大影响了模型的收敛速率以及收敛精度。

3.2 基于卡尔曼滤波器的隐私保护机器学习优化方法

本节针对差分隐私机器学习优化方法中的梯度波动问题和动量超调问题，提出了基于卡尔曼滤波器的梯度过滤算法和基于卡尔曼滤波器的超调过滤算法，然后将二者结合为基于卡尔曼滤波器的隐私保护机器学习优化方法，并进行了大量实验验证。

3.2.1 梯度过滤方法

根据对梯度波动问题的分析可知，可以将使用 DPSGD 训练中的梯度变化过程近似看做是一个包含噪声的线性过程。正如图3-1所示，其中全局梯度变化较为稳定，DPSGD 的梯度由于随机噪声和额外添加的高斯噪声故其自身的波动比随机梯度的波动更大，这也造成了 DPSGD 方法训练模型时往往收敛较慢且最终收敛精度较差的问题。从动态系统的角度来讲，DPSGD 算法对梯度进行计算的过程可以近似看做是一个动态观测过程，每一步的全局梯度值便是每一个时间戳的真实状态，包含了多种噪声。

$$x_i(t) = r(t) + v_i(t) \quad (3-8)$$

如式(3-8)所示，其中 $x_i(t)$ 为第 t 轮的观测值即第 t 次迭代时求得的差分隐私随机梯度， $r(t)$ 为第 t 次迭代时的全局梯度，而 $v_i(t)$ 则为观测噪声。观测噪声 $v_i(t)$ 的方差由 σ_{dp} 和 σ_{sgd} 两部分组成。其中的 σ_{dp} 是为满足差分隐私而添加的高斯噪声的方差，而 σ_{sgd} 则为第 t 步时求的随机梯度时，由于进行了随机采样进而所带来的额外噪声方差，且与批大小成反比。基于上述结论，为了降低训练过程中梯度的波动方差，提升梯度自身的精确度，故基于全局梯度近似线性的稳定变化性质作出如下假设：

假设 3.1 (全局梯度线性变化假设) 在采用基于差分隐私的优化方法训练模型的过程中，可以将模型训练过程中其梯度的变化过程近似地看做是一个线性动态过程。

$$r(t+1) = A(t)r(t) + \omega(t) \quad (3-9)$$

式(3-9)中， $r(t+1)$ 代表了第 $t+1$ 次梯度下降时的全局梯度，而 $\omega(t)$ 则是造成梯度波动的噪声，其满足分布 $\mathcal{N}(0, \sigma)$ ，其中 σ 为梯度波动噪声的方差。

基于假设3.1，便可将模型训练过程中每一步求得的随机梯度推广为卡尔曼滤波场景下的线性传感器的观测值。如式(3-10)所示， $x_i(t)$ 为第 t 轮的观测值即第 t 次迭代时求得的随机梯度， $r(t)$ 为第 t 次迭代时的全局梯度， $H_i(t)$ 为观测占比，由于随机梯度是全局梯度的无偏估计故其在本场景下观测占比默认为 1，而 $v_i(t)$ 则为观测噪声。

$$x_i(t) = H_i(t)r(t) + v_i(t) \quad (3-10)$$

基于上述分析, 已经将基于差分隐私的随机梯度下降过程近似等价转换成一个存在高斯噪声的动态线性变换过程, 而卡尔曼滤波器作为常用于在动态线性系统中过滤高斯噪声的方法, 此时便可以迁移至模型训练的过程中。此时训练过程如下:

$$\theta_{t+1} \leftarrow \theta_t - \eta_t KFG\text{Grad}(\tilde{g}(t)); \quad (3-11)$$

如式(3-11)所示, θ_t 代表第 t 轮迭代结束后模型的参数, $\tilde{g}(t)$ 为第 $t+1$ 轮计算得到的梯度信息, $KFG\text{Grad}$ 为卡尔曼滤波器, 其具体细节可见 2.2.1。利用卡尔曼滤波器自身的性质用来过滤 DPSGD 过程中添加的高斯噪声, 以平滑噪声对梯度的影响, 使得梯度更为准确。其中卡尔曼滤波器的系统标准差与裁剪系数 C 相关, 观测方差为添加的高斯噪声方差 σ 。如图3-1所示, KF-Grad 便是将卡尔曼滤波器应用至 DPSGD 训练过程中的结果, 这在一定程度上降低了梯度的波动, 相较于 DPSGD 的原始梯度变得更为准确, 过滤后的梯度介于全局梯度和 DPSGD 梯度之间。明显降低了梯度在整体上的波动。在不降低隐私保护程度的前提下, 实现了近似于降低了差分隐私噪声方差的效果。综上, 基于卡尔曼滤波器的梯度过滤方法 (简称为 KF-Grad) 算法的具体过程如下:

算法 3-1 KF-Grad 算法

输入: 样本 x_1, x_2, \dots, x_n , 损失函数 $\mathcal{L}(\theta) = \sum_{i=1}^n \mathcal{L}(\theta, x_i)$, 学习率 η_t , 噪声方差 σ , 样本批大小 L , 梯度裁剪系数 C

输出: θ_T , 以及隐私损耗 (ϵ, δ)

- 1 随机初始化模型参数 θ_0 , 卡尔曼滤波器 $\mathcal{K}(2C, \sigma)$;
- 2 **while** $t < T$ **do**
- 3 以 L/N 的概率从全体样本中随机抽取样本集合 L_t ;
- 4 计算梯度;
- 5 针对 L_t 中的每个样本, 计算其梯度 $g_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$;
- 6 梯度裁剪;
- 7 $\bar{g}_t(x_i) \leftarrow g_t(x_i) / \max(1, \frac{\|g_t(x_i)\|_2}{C})$;
- 8 添加噪声;
- 9 $\tilde{g}(t) \leftarrow \frac{1}{L} (\sum_i \bar{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I))$;
- 10 KF-Grad 过滤;
- 11 $\theta_{t+1} \leftarrow \theta_t - \eta_t KFG\text{Grad}(\tilde{g}(t))$;
- 12 梯度下降;
- 13 $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{g}(t)$;
- 14 **end**

如算法3-1所示, 第一步先以 L/N 的概率从全体样本中进行采样, 此步引入了随机采样的噪声 σ_{sgd} , 其大小取决于批大小。第二步是计算这批样本的梯度信息 $\bar{g}_t(x_i)$ 。之后是差分隐私优化方法的步骤, 其中第三步对每个样本的梯度均进行梯度裁剪, 以控制模型训练过程的敏感度, 即限制单个样本在此次更新的作用, 与最基础的差分隐私定义相对应。第四步需要给每个样本的梯度中都添加高斯噪声, 其中高斯噪声取决于隐私预算和裁剪系数的大小, 并且添加完噪声后会对本次所有采样的样本的梯度进行

平均,以获得此次下降的梯度,据此可以得出,训练过程中的批大小越大,则经过平均后的梯度中的噪声的方差也就越低,同时,裁剪系数越大,则应用于梯度下降的噪声方差越小,在实践中,往往需要根据实验数据及模型的不同,根据实验情况动态调节批和裁剪系数的大小,以防止噪声过大进而影响模型的收敛。第五步是本节的主要创新点,即使用卡尔曼滤波器对添加完高斯噪声的梯度进行过滤;卡尔曼滤波器能够在动态线性系统中过滤高斯噪声,进而得到系统的最优估计,并且前文已经将 DPSGD 算法的训练过程建模为一个包含高斯噪声的线性动态系统,故此方法可以降低梯度偏差使得模型向着更为准确的方向进行更新。除此之外,与全局梯度下降相比,随机梯度下降由于噪声的引入故需要更低的学习率,同理, DPSGD 算法相较于随机梯度下降方法也需要更低的学习率,以防止模型向错误的方向更新太多影响模型收敛,然而使用 KF-Grad 方法后,梯度相较于原来更为平稳,在某些情况下甚至比随机梯度还要更为准确。因此为了进一步提升模型的收敛速率,该方法可以额外引入一个步长增强参数 $\alpha(t)$,该参数是一个动态值,在训练初期时可以通过增大步长增强参数来使其快速收敛,而在模型收敛后期通过降低步长增强参数进而减小更新防止错过模型最优点。

基于卡尔曼滤波器的梯度过滤方法,通过将梯度下降过程近似假设为一个线性动态过程,然后利用卡尔曼滤波器能够过滤线性动态系统中所存在的高斯噪声的性质,用来过滤模型训练过程中为了提供差分隐私保护进而给梯度所添加的高斯噪声,起到了类似于降低噪声方差的作用。并且尽管降低了噪声对模型更新过程的影响但并不会因此影响差分隐私的保护效果。因为此方法本质上是利用相邻两次梯度之间的时空相关性而进行的优化,即两次相邻的梯度变化之间的差别是有限的,故可以利用之前的梯度信息用来对此次梯度的准确性进行一定程度的校准,而非利用更多的额外样本信息,由于差分隐私具有后处理性质,即一旦添加噪声提供差分隐私保护后,对处理后的数据做任何其他额外的处理都不会影响差分隐私的保护性能,所以该方法不会出现额外的隐私信息泄露问题。本方法与其他基于差分隐私的隐私效用权衡方法不同点在于,基于卡尔曼滤波器的梯度过滤算法将梯度更新过程建模为线性变化的动态系统过程,利用相邻梯度之间不会波动过大的性质,基于卡尔曼滤波器对此处的梯度信息进行更为充分的校准。KF-Grad 算法在保证隐私保护程度即隐私预算不变的前提下,使得梯度变得更为准确且平稳,进一步加快了模型的收敛速率和最终的收敛精度,进而大大增加了基于差分隐私的优化方法的可用性,更好实现隐私和效用的权衡。

3.2.2 超调过滤方法

为了进一步解决隐私和效用权衡的问题,本节将引入动量优化方法用来给 DPSGD 方法进行加速,并针对其所带来的超调问题,本节提出新的基于卡尔曼滤波器的超调过滤方法用来缓解基于动量更新的 DPSGD 算法中的超调现象。机器学习模型的参数可

以被初始化成为一个标量 θ_0 ，并且它通过不断地迭代最终达到其最优解 θ^* 。此时，模型参数的变化便可以被视为是控制理论中的一个阶跃反应。根据文献 [57] 中的数学推导可以得到该优化过程的形式化表示，具体过程如下：

$$\theta(t) = \theta^* - \frac{(\theta^* - \theta_0) \sin(\omega_n \sqrt{1 - \zeta^2} t + \arccos(\zeta))}{e^{\zeta \omega_n t \sqrt{1 - \zeta^2}}} \quad (3-12)$$

如式 (3-12) 所示， ω_n 和 ζ 可以看做是此系统中的阻尼比和固有频率，在模型和数据集确定的情况下，该模型的收敛的过程中这两个值是一个常数值。由此可以得出，在 DPSGD 的场景下，使用动量优化方法进行模型训练的过程中，模型参数的变化过程可以被看做是一个是一个噪声不断降低的线性动态过程。由于存在方差指数级递减的超调噪声，因此在收敛后期会存在超调的现象，而差分隐私技术的使用又额外引入了更多的噪声进而加强了超调现象的影响。基于上述结论，本节作出如下假设：

假设 3.2 (超调噪声线性变化假设) 在采用 DPSGD+Momentum 方法训练模型的过程中，可以近似将模型参数的变化过程看做是一个近似的线性动态过程，具体过程如下：

$$r(t+1) = A(t)r(t) + \omega(t) \quad (3-13)$$

式(3-13)中， $r(t+1)$ 代表了第 $t+1$ 个 epoch 后模型的参数值，而 $\omega(t)$ 则是引发超调问题的噪声，其分布近似满足 $\mathcal{N}(0, \zeta)$ ， ζ 为超调噪声的方差，且该噪声的方差会随着迭代，不断指数级衰减直至降低为 0，此时超调现象也随之结束。

基于假设3.2, 模型训练过程中得到的模型参数值则也可推广为卡尔曼滤波情形下的线性传感器（或测量模型）的观测值, 具体过程如下：

$$x_i(t) = H_i(t)r(t) + v_i(t) \quad (3-14)$$

基于式(3-14)，此时可以得出模型训练时的参数变化过程可近似看作为是一个动态线性过程，且该过程存在一个方差指数衰减的超调噪声 $v_i(t)$ 。在真实的工业界场景中，该线性变化过程和所用数据集和以及选用的模型有直接的关联，在训练过程中的波动噪声往往也具有较强的不确定性，但其指数级衰减的性质往往普遍存在，因为使用动量更新方法时，尽管模型收敛的后期存在的不规则波动情况，但波动的幅度整体上不断地衰减变小。为了提高假设的普适性，本节将造成超调现象的噪声假设为方差指数级衰减的高斯噪声，以适应复杂的真实应用场景。如假设3.3所示：

假设 3.3 (超调噪声分布假设) 基于动量优化方法的 DPSGD 算法在训练模型的过程中，可以将造成模型参数超调的噪声近似地看做是一个方差指数级衰减的高斯噪声，该噪

声近似满足分布 $\mathcal{N}(0, \eta' \gamma(t))$ 。

假设3.3将存在不确定性的指数衰减的超调噪声假设为方差指数级衰减的高斯噪声，此假设的原因在于高斯噪声的分布在真实世界中普遍存在，因此具有较强的代表性意义。噪声分布 $\mathcal{N}(0, \eta' \gamma(t))$ 中的 η 为系统方差指数级衰减过程中的衰减系数，为了简化训练过程，在应用过程中可以将其固定化为一个小于 1 的实数值。与 KF-Grad 方法不同，此时的噪声是存在于模型参数中的噪声而非梯度中所包含的噪声，即动态过程中存在于模型参数当中的系统噪声，而非每一步梯度中所包含的观测噪声。梯度噪声会随着迭代次数的增加不断地进行累计，但系统噪声不会。基于上述分析，将模型训练时参数变化过程建模为线性系统，所以可以利用卡尔曼滤波器对其进行进一步优化。为了更适用于过滤超调噪声，首先需要对卡尔曼滤波器进行一定的修改，经过修改后的卡尔曼滤波器与传统的卡尔曼滤波器方法不同点在于，其额外引入了系统方差衰减系数 η ，用来模拟指数衰减的超调噪声，该参数的取值范围是 $(0, 1)$ ，当 $\eta=1$ 时，此时的算法就等价于传统的卡尔曼滤波器方法，该方法对应于超调现象中的衰减现象，使其更为适用于过滤动量更新引发的超调现象，加快模型的收敛。具体算法如下：

算法 3-2 Kalman Filter2 算法

输入：数据 $x_i(t)$, 观测矩阵 $H_i(t)$, 初始化方差 $P_i(0) = P_0$, 测量方差 $R_i(t)$, 总过程次数 T , 衰减系数 η

输出： $\hat{x}_i(T)$

```

1 对所有参数进行初始化设置;
2 while  $t < T$  do
3   计算卡尔曼增益;
4    $K(t) = P(t)H_i^T(R + H_i(t)P(t)H_i(t)^T)^{-1}$ ;
5   更新状态估计;
6    $\hat{x}_i(t) = \bar{x}_i(t) + K_i(t)(x_i(t) - H_i(t)\bar{x}_i(t))$ ;
7   更新状态协方差;
8    $M(t) = P(t) - P(t)H_i^T(R + H_i(t)P(t)H_i(t)^T)^{-1}H_i(t)P(t)$ ;
9   预测状态的协方差;
10   $P(t+1) = AM(t)A + Q(t)$ ;
11  进行系统方差衰减;
12   $Q(t+1) = \eta Q(t)$ ;
13  预测系统的状态;
14   $\bar{x}_i(t+1) = A(t)\hat{x}_i(t)$ ;
15 end
```

如算法 3-2 所示，该算法与本文 2.2.2 节中的卡尔曼滤波器算法不同点在于其额外增加了系统方差衰减的过程，即 $Q(t+1) = \eta Q(t)$ ，以此来更好地适应系统方差指数级衰减的超调噪声波动过程，进而有效地过滤其中的超调噪声。基于上述内容，本文针对动量优化方法所引发的超调问题，提出基于卡尔曼滤波器的参数超调过滤方法 (简称为 KF-Param 方法)，具体算法如下：

算法 3-3 KF-Param 算法

输入：样本 x_1, x_2, \dots, x_n , 损失函数 $\mathcal{L}(\theta) = \sum_{i=1}^n \mathcal{L}(\theta, x_i)$, 学习率 η_t , 噪声方差 σ , 样本批大小 L , 梯度裁剪系数 C

输出： θ_T , 以及隐私损耗 (ε, δ)

- 1 随机初始化模型参数 $\theta_0, \mathcal{K}(2C, \sigma)$;
- 2 **while** $t < T$ **do**
- 3 计算梯度;
- 4 针对 L_t 中的每个样本, 计算其梯度 $g_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$;
- 5 梯度裁剪;
- 6 $\bar{g}_t(x_i) \leftarrow g_t(x_i) / \max(1, \frac{\|g_t(x_i)\|_2}{C})$;
- 7 添加噪声;
- 8 $\tilde{g}(t) \leftarrow \frac{1}{L}(\sum_i \bar{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I))$;
- 9 Momentum 累计;
- 10 $V_{t+1} = \alpha V_t - \gamma \eta_t \tilde{g}(t)$;
- 11 梯度下降;
- 12 $\theta_{t+1} \leftarrow \theta_t + V_{t+1}$;
- 13 KF-Param 方法过滤;
- 14 $\theta_{t+1} = \text{KalmanFilter2}(\hat{\theta}_{t+1})$;
- 15 **end**

算法3-3展示了基于卡尔曼滤波器的超调过滤方法的算法细节, 第一步首先进行样本采样并进行梯度计算, 之后便是进行梯度裁剪和添加高斯噪声; 然后与 KF-Grad 方法不同的是, 第四步进行了动量的累计计算, 其将之前累计的动量信息通过乘以动量系数 α 进行固定衰减, 然后将最新梯度进行累计, 此步也会在一定程度上累计高斯噪声; 第五步是进行梯度下降, 模型基于最新的动量梯度进行更新; 最后基于 KF-Param 方法对模型更新后的参数进行过滤, 降低超调噪声对模型收敛的影响, 在模型收敛的后期遏制住超调现象的出现。在模型训练的过程中, 系统方差变化、超调过程以及模型的收敛过程这三个过程是同步的, 随着模型的不收敛, 系统方差也会随之不断进行衰减, 超调现象也会不断随之衰减, 当模型最终收敛时, 系统方差也会随之无限趋近于 0, 超调现象也会随之结束。基于卡尔曼滤波器的超调过滤方法 (KF-Param 方法) 通过利用卡尔曼滤波器将采用 DPSGD-Momentum 更新后的模型参数进行进一步的过滤, 以缓解动量优化方法所带来的超调副作用。在模型收敛的后期, 能够很好的抑制噪声累计所带来的超调现象, 进一步加快了模型的收敛速率。

3.2.3 隐私保护机器学习优化方法

基于上述提到的两个基于卡尔曼滤波器的优化方法, 由于二者的作用对象不相同且解决的问题也并非相同, 故可以将上两个章节提出的两个算法进行结合形成基于卡尔曼滤波器的隐私保护机器学习优化方法 (简称为 KF2 算法)。其中, 基于卡尔曼滤波器的梯度过滤方法作用于模型训练过程中包含高斯噪声的梯度, 主要用于过滤训练过程梯度中的高斯噪声, 在不降低隐私预算的前提下, 通过利用相邻两次梯度的相关性

以降低梯度中的噪声方差，进而加快模型的收敛速率并提高模型收敛的精度。基于卡尔曼滤波器的超调过滤方法其作用于模型训练过程中的模型参数，主要用于缓解动量优化方法所带来的超调现象，在加速的同时，通过利用方差指数衰减的卡尔曼滤波器降低造成超调的噪声方差，进而确保模型在训练后期能够稳定地收敛。二者所作用的模型对象并不相同并且可以相为补充，故本节将这两个算法进行结合，以进一步加快模型的训练速率，提升模型的最终精度。具体的算法过程如下：

算法 3-4 基于卡尔曼滤波器的优化算法

输入：样本 x_1, x_2, \dots, x_n , 损失函数 $\mathcal{L}(\theta) = \sum_{i=1}^n \mathcal{L}(\theta, x_i)$, 学习率 η , 噪声方差 σ , 样本批大小 L , 梯度裁剪系数 C

输出： θ_T , 以及隐私损耗 (ϵ, δ)

- 1 随机初始化模型参数 $\theta_0, \mathcal{K}(2C, \sigma)$;
- 2 **while** $t < T$ **do**
- 3 计算梯度;
- 4 针对 L_t 中的每个样本, 计算其梯度 $g_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$;
- 5 梯度裁剪;
- 6 $\bar{g}_t(x_i) \leftarrow g_t(x_i) / \max(1, \frac{\|g_t(x_i)\|_2}{C})$;
- 7 添加噪声;
- 8 $\tilde{g}(t) \leftarrow \frac{1}{L} (\sum_i \bar{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I))$;
- 9 KF-Grad;
- 10 $\theta_{t+1} \leftarrow \theta_t - \eta_t \text{KFGrad}(\tilde{g}(t))$;
- 11 Momentum 累计;
- 12 $V_{t+1} = \alpha V_t - \gamma \eta_t \tilde{g}(t)$;
- 13 梯度下降;
- 14 $\theta_{t+1} \leftarrow \theta_t + V_{t+1}$;
- 15 KF-Param 方法过滤;
- 16 $\theta_{t+1} = \text{KFParam}(\hat{\theta}_{t+1})$;
- 17 **end**

如算法 3-4 所示，在模型的训练过程中，计算完梯度后可先使用 KF-Grad 方法用于过滤模型的梯度，在降低梯度噪声方差的同时，也降低了动量优化累计动量的噪声方差，在完成此步梯度下降后，通过 KF-Param 方法来降低动量优化方法中所包含的超调噪声的方差。该算法通过将 KF-Param 方法与 KF-Grad 方法进行结合，不仅可以降低模型训练过程中的梯度噪声，还能降低动量更新方法所累计的噪声，进而缓解超调现象，进一步提升了收敛精度，更好地实现了隐私和效用的权衡。

3.3 实验结果与分析

本节的主要对本章所提出的三个基于卡尔曼滤波器的优化方法 (KF-Grad 方法、KF-Param 方法和 KF2 方法) 进行实验。首先介绍实验设置，包括数据集、实验环境、模型介绍；其次从多个不同的角度针对本章提出的三个基于卡尔曼滤波器的优化算法进行了系统化的实验，并分析各个算法在不同参数组合下的自身的性质和优势。

3.3.1 实验设置

1) 数据集介绍

本文主要采用了 MINIST 数据集和 CIFAR10 数据集这两个图片数据集进行实验。这两个数据集均属于典型的多分类问题，其对应的评价指标为准确率。

(1) MINIST 数据集是一个大型的手写数字数据集，该数据集通常被用于训练各种图像处理系统，数据集共包含了 60000 个训练样本和 10000 个测试样本，共计 70000 个样本。每个样本都是一个灰度图像，大小为 28x28 像素。这些图像展示了从 0 到 9 的手写数字，每个数字都有大约 6000 个样本。

(2) CIFAR10 数据集是一个彩色图像数据集，通常用于图像分类任务。该数据集共包含 60000 张 32x32 像素的 RGB 彩色图像，其中包含 50000 张训练图片和 10000 张测试图片，这些图片共涵盖了 10 个不同的类别，每个类别有 6000 个图像。这些类别分别为：飞机、汽车、鸟类、猫、鹿、狗、蛙、马、船和卡车。

2) 实验环境

本实验采用的编程语言为 Python，所使用的深度学习实验框架为 PyTorch，实验在 Linux 操作系统中利用 GPU 运行，具体的各类实验配置信息如表3-1所示。

表 3-1 实验环境配置信息

| 配置 | 参数 |
|------------|----------------------------|
| 操作系统 | Ubuntu 22.04.4 LTS |
| GPU | NVIDIA GeForce RTX 2080 Ti |
| Python 版本 | Python 3.11.7 |
| Pytorch 版本 | Pytorch 1.8.0 |

3) 模型介绍

本实验主要选用了三个深度学习网络，与 MINIST 数据集和 CIFAR10 数据集进行组合形成三组实验，具体的算法-数据集组合如下：

(1)MLP-MINIST: 基于 MINIST 数据集的两层全连接神经网络(简称为 MLP-MINIST)，其中激活函数使用的是 ReLU 激活函数，隐藏层节点的个数为 1000；

(2)CNN-MINIST: 基于 MINIST 数据集的两层卷积神经网络 + 一层全连接神经网络(简称 CNN-MINIST)，其中两个卷积层的输入通道数和输出通道数组合分别为 (1,16) 和 (16,32)，两个卷积核尺寸均为 5；

(3)CNN-CIFAR10: 基于 CIFAR10 数据集的两层卷积神经网络 + 一层全连接神经网络(简称 CNN-CIFAR10)，其中两个卷积层输入通道数和输出通道数组合分别为

(3,18) 和 (18,36)，两个卷积核尺寸均为 5。

基于上述三个深度学习网络，与选取的两个数据集组成了 MLP-MINIST、CNN-MINIST 和 CNN-CIAFAR10 三个组合，本章节所有实验均基于这三组实验组合进行。

3.3.2 梯度过滤方法实验

本节将对基于卡尔曼滤波器的梯度过滤方法进行实验验证，针对该方法的实验主要包含三个部分：1) 梯度波动过滤效果实验：该实验用于验证 KF-Grad 算法过滤梯度的效果；2) 梯度过滤方法实验：该实验用于验证 KF-Grad 算法相较于 DPSGD 算法所带来的提升；3) 基于不同高斯噪声方差下的梯度过滤方法实验：该实验选取多组高斯噪声方差，验证 KF-Grad 方法在不同噪声方差下的所有结果。具体实验过程如下：

1) 梯度波动过滤效果实验

本章提出的基于卡尔曼滤波器的梯度过滤方法之所以有效，是因为卡尔曼滤波器能够很好地过滤掉添加到梯度中的高斯噪声以及随机采样造成的随机噪声，进而使得梯度信息更为准确和稳定。尽管前文 3.1.2 中已经给出了部分实验结果论证，但为了更完善地验证基于卡尔曼滤波器的梯度过滤方法的有效性，本节将在三组不同的算法-数据集的组合下，通过绘制出模型训练过程中参数的多个梯度变化曲线，来验证该方法针对梯度波动问题所带来的准确性和稳定性的提升。

本实验基于 DPSGD 算法进行模型的训练，并针对三组不同的模型-数据组合，随机选取了其模型中的任意两个参数。在模型训练过程中，每一步除了需要计算添加高斯噪声的随机梯度外，还需要额外计算其全局梯度以及经过 KF-Grad 方法过滤后的梯度，并全程记录这三类梯度的变化曲线，实验参数设置为：高斯噪声的方差为 3，裁剪系数为 1，样本批大小为 50，训练迭代轮数为 500。具体结果如下：

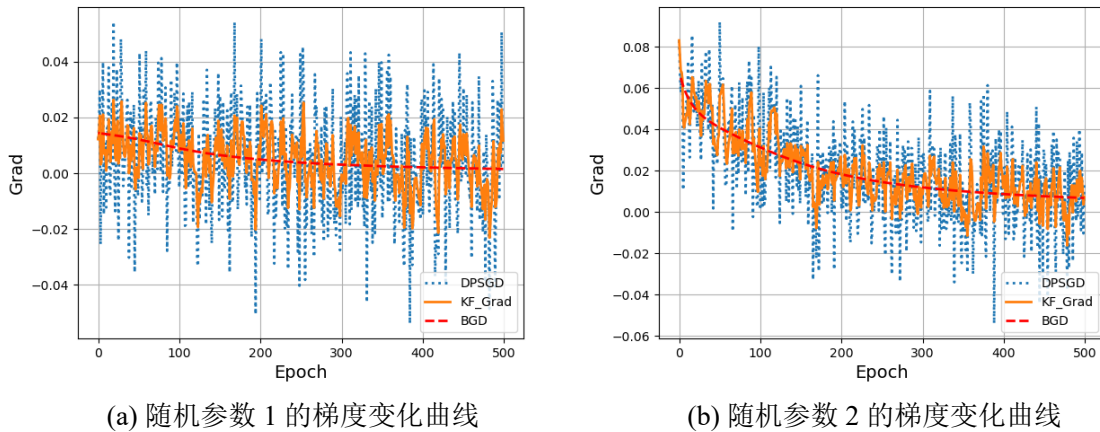


图 3-3 KF-Grad 方法在 MLP-MINIST 组合下的梯度变化曲线

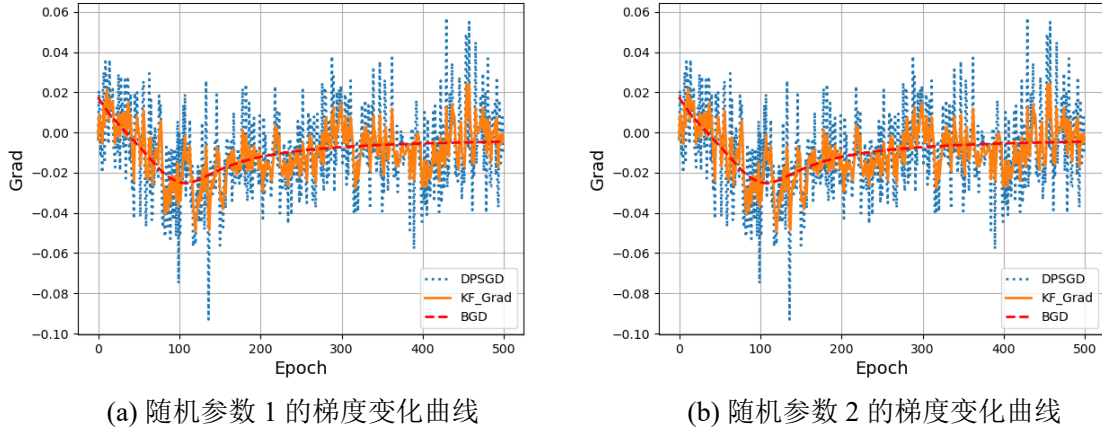


图 3-4 KF-Grad 方法在 CNN-MINIST 组合下的梯度变化曲线

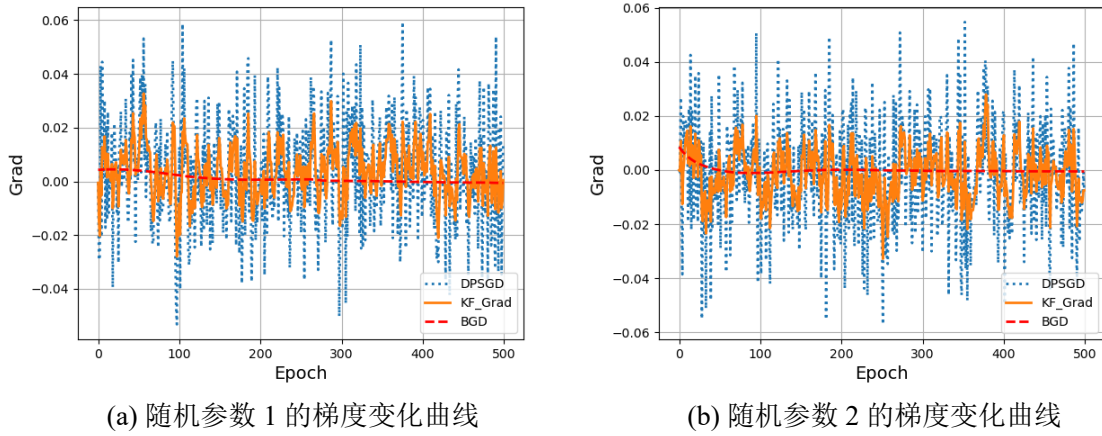


图 3-5 KF-Grad 方法在 CNN-CIAFAR10 组合下的梯度变化曲线

如图3-3、图3-4、图3-5所示，其中红色的线是全局梯度的变化曲线，蓝色的虚线是基于 DPSGD 算法的梯度变化曲线，橙色的线是经过卡尔曼滤波过滤后的 DPSGD 梯度变化曲线。从实验结果中不难看出，全局梯度 BGD 处于非常稳定的状态，在全局 500 个次数的迭代中，全局梯度的变化十分缓慢，在相邻几次的更新中，其梯度信息几乎未发生变化，这也印证了本文对其做出的全局梯度信息近似线性变化的假设。而 DPSGD 部分，由于其额外引入了两部分噪声，即随机采样噪声和满足差分隐私的高斯噪声，故其相较于全局梯度 BGD，发生了非常明显的波动，由图中的实验结果也可以观察到，在所对比的三组梯度中，DPSGD 的梯度波动幅度最大，其对应的噪声方差也是最大的。而 KF-Grad 实验组，其是基于 DPSGD 梯度使用了卡尔曼滤波器对其进行过滤，由于卡尔曼滤波器可以很好地过滤动态线性系统中的高斯噪声，故在此场景下也可以过滤掉梯度中所包含的差分隐私噪声以及随机抽样所带来的额外噪声，故相较于 DPSGD 实

验组，KF-Grad 组的梯度波动要明显更小，其与全局梯度的值也更为相近。从全局上来看，随着训练的不断进行，三组梯度的变化趋势是相同的，并且其梯度值全部逐渐趋于 0。这是因为尽管 DPSGD 组实验的梯度中包含了差分隐私高斯噪声以及随机采样噪声，但是由于这两个噪声是无偏的，因此 DPSGD 和 KFGrad 这两组梯度均围绕着全局梯度在上下波动，并且随着添加的高斯噪声方差越大，随机抽样的样本数量越小，则波动程度就越大。这种波动的噪声造成了 SGD 算法和 DPSGD 算法收敛所需要的迭代次数要远大于全局梯度下降算法所需要的迭代次数，而基于卡尔曼滤波器的梯度过滤方法可以有效缓解 DPSGD 训练过程中的梯度波动问题，使得训练过程中的每一步梯度都更为准确，进而可以大大提升模型收敛速率，并提升模型的收敛精度。

根据上述分析，可以得到在多个算法和数据集的不同组合下，KF-Grad 方法具有一定的普适性，都能够在较大程度上过滤掉梯度中所包含的高斯噪声，使得下降时的梯度跟全局梯度更为接近，进而降低模型训练过程中的误差。

2) 梯度过滤方法实验

为了评估基于卡尔曼滤波器的梯度过滤算法的效果，本文采用了三组实验对本方法进行验证，分别为 MLP-MINIST、CNN-MINIST 和 CNN-CIAFAR10。实验组为采用基于卡尔曼滤波器的梯度过滤算法，对照组为 DPSGD 算法。实验结果如下：

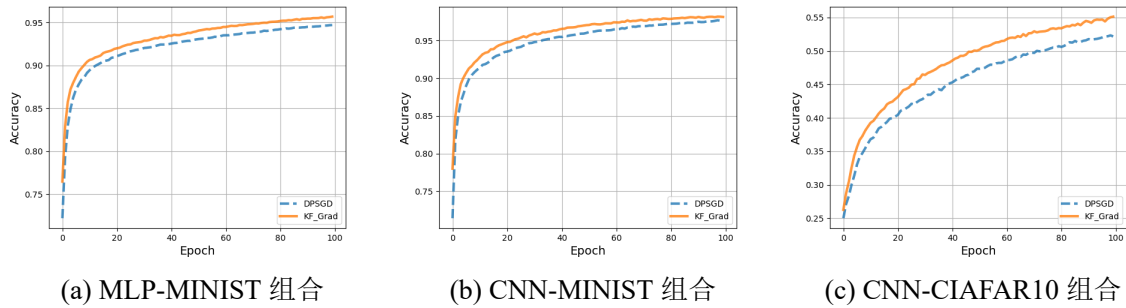


图 3-6 KF-Grad 方法在不同算法-数据集组合下的实验结果

实验结果如3-6所示，由图中的结果可以得知，KF-Grad 方法的使用大大加快了模型的收敛速率和收敛精度。其中 MLP-MINIST 任务采用 DPSGD 算法的准确率达到 94.71%，而 KF-Grad 算法达到了 95.69% 的准确率，相对提升了 0.98%；CNN-MINIST 任务的 DPSGD 算法达到了 97.62% 的准确率而 KF-Grad 算法达到了 98.13%，相对提升了 0.51%；CNN-CIFAR10 任务的 DPSGD 算法达到了 52.18% 的准确率，而 KF-Grad 算法达到了 55.13% 的准确率，相对提升了 2.95%。最终的实验结果显示，CNN-MINIST 组的实验结果最好，但 KF-Grad 方法相对的提升最小，只有 0.51%。而 CNN-CIAFAR10 任务的实验结果最差，但 KF-Grad 方法相对的提升却是最大，有 2.95%，相较于 CNN-MINIST

方法的提升增加了将近 6 倍。基于上述结果，可以分析得出当数据集和模型更为复杂时，KF-Grad 方法的提升效果更大，而数据集和模型较简单时，KF-Grad 方法的提升较低。本文分析认为造成该现象主要的原因是，当模型和数据集较为复杂时，给模型梯度中添加的噪声的个数也更多，故其受到的影响也会更大，而当模型和数据集较为简单的时候，由于添加噪声的个数变少，故其收到的影响也就更小。例如 CNN-CIAFAR10 这组实验，由于此任务较为复杂，模型难收敛且收敛精度低。同理其他两组实验的模型偏小且任务较为简单，所以 KF-Grad 算法带来的提升较小。

3) 基于不同高斯噪声方差下的梯度过滤方法实验

本节采用不同的噪声方差进行实验，以分析不同噪声方差情况下 KF-Grad 方法的效果表现。分别选取方差 σ 为 1、3、5 的高斯噪声进行 KF-Grad 的测试实验。其中实验组采用 KF-Grad 算法进行训练，而对照组采用 DPSGD 算法进行训练。除实验组额外使用 KF-Grad 对梯度进行过滤外，两者其他超参完全相同。具体的实验结果如下：

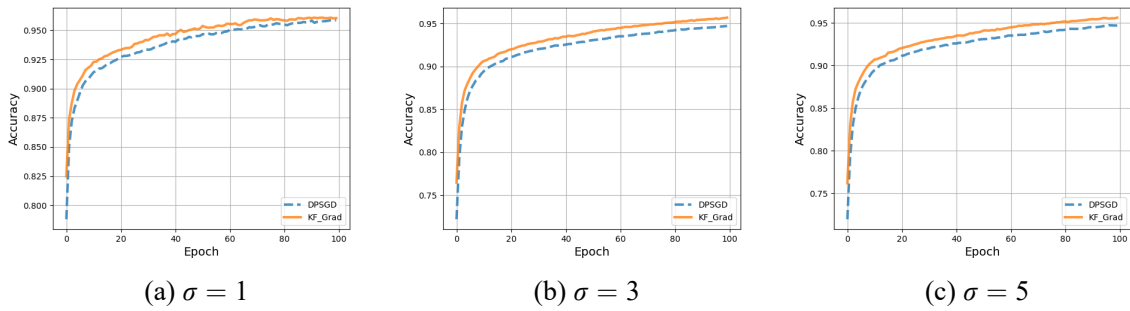


图 3-7 MLP-MINST 组合下 KF-Grad 方法基于不同噪声方差 σ 的实验结果

根据图3-7中结果显示，高斯噪声方差为 1 时，实验最终的精度最高，其中 DPSGD 的精度达到了 95.85%，而 KF-Grad 方法的精度达到了 96.04%，二者之间的差距仅有 0.19%；高斯噪声方差为 3 时，DPSGD 方法达到了 94.71% 的准确率，其相较于方差为 1 时的结果降低了 1.14%，这是由于方差增加进而导致在模型更新的过程中，添加了更多的噪声，进而影响模型的效果，而 KF-Grad 方法最终的精度为 95.69%，其相较于 DPSGD 方法提升了 0.98%；高斯噪声方差为 5 时，DPSGD 方法和 KF-Grad 方法之间的差距最大，二者准确率相差 1.1%。基于上述实验结果可分析得出，在训练过程中，向梯度中添加的高斯噪声越大，最终的实验精度越低，但 KF-Grad 方法相较于 DPSGD 方法的提升也越大。当噪声方差为 1 时，DPSGD 方法和 KF-Grad 方法之间的差距最小，而噪声方差为 5 时，DPSGD 方法和 KF-Grad 方法之间的差距最大。这说明了本方法通过卡尔曼滤波器有效地过滤掉了往梯度中添加的噪声，增加了梯度的准确性，进而提升了最终模型的效果。本方法在较高的噪声中的表现相较于传统 DPSGD 算法有更大的提升。

为了进一步验证此结论的准确性, 本文又选取了 CNN-MINIST 以及 CNN-CIAFAR10 两组实验进行了进一步验证, 实验结果如下:

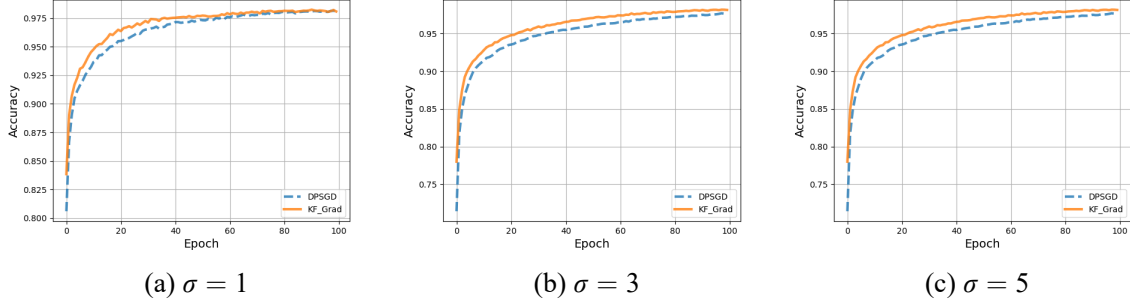


图 3-8 CNN-MINIST 组合下 KF-Grad 方法基于不同噪声方差 σ 的实验结果

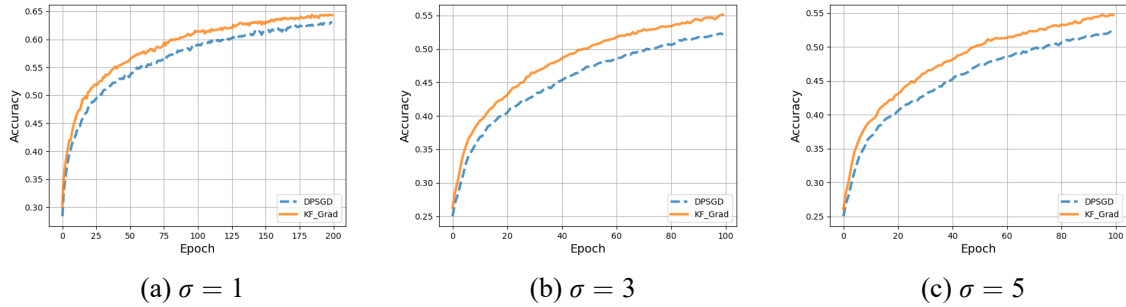


图 3-9 CNN-CIAFAR10 组合下 KF-Grad 方法基于不同噪声方差 σ 的实验结果

由图 3-8 和图3-9中的结果可得, KF-Grad 方法在这两个算法-数据集组合下均达到了最优。并且随着高斯噪声方差的增加, KF-Grad 方法在收敛速度和收敛精度上的提升也之增加, 当噪声方差为 5 时, 两个方法之间的差距最大。这进一步论证了本方法在基于差分隐私的优化过程中的有效性, 尤其是在高噪声的情况下, 本方法的提升更大。

本节的三个实验充分验证了 KF-Grad 方法的有效性。根据对这三组实验结果进行细致分析可以得知如下三个结论: (1) 是在采用 DPSGD 算法进行模型训练时, KF-Grad 方法可以有效地过滤其中存在的噪声, 使得梯度更为准确, 与全局梯度更为匹配; (2) 是在采用 DPSGD 算法进行模型训练时, 模型和数据集越复杂, 则 KF-Grad 方法所带来的提升越大; (3) 为在采用 DPSGD 算法进行模型训练时, 为了实现差分隐私保护所添加的高斯噪声方差越大, KF-Grad 方法相对于 DPSGD 的提升就越为明显。

3.3.3 超调过滤方法实验

本节将对基于卡尔曼滤波器的超调过滤方法进行实验, 共包含三个部分: 1) 基于卡尔曼滤波器对超调噪声过滤效果进行实验验证; 2) 对基于卡尔曼滤波器的超调过滤方

法进行实验验证；3) 在不同的动量系数下对基于卡尔曼滤波器的超调过滤方法进行实验验证。本实验采用的评价指标为准确率，高斯噪声的方差为 3，裁剪系数为 1，样本批大小为 50，动量系数 α 为 0.9。

1) 超调噪声过滤实验

超调过滤方法有效的前提是卡尔曼滤波器能够在基于差分隐私的动量优化场景下有效过滤模型参数中的超调噪声，为了进一步验证本章所做出的超调噪声线性变化假设和超调噪声分布假设，本节主要测试差分隐私场景下卡尔曼滤波器对超调噪声的过滤效果。其中实验的基线 1 名为 **OverShoot**，该组结果表示在不添加差分隐私噪声时超调噪声的波动曲线，即式 $\theta(t) = \theta^* - \frac{(\theta^* - \theta_0) \sin(\omega_n \sqrt{1-\zeta^2} t + \arccos(\zeta))}{e^{\zeta \omega_n t \sqrt{1-\zeta^2}}}$ 所表示的波动过程；实验的基线 2 名为 **DP**，该组结果表示添加过差分隐私噪声后的动量优化过程中的超调噪声波动曲线；实验组名为 **KF**，该组结果代表着使用卡尔曼滤波器 **KF** 对添加过差分隐私噪声的超调噪声进行过滤的实验结果；除此之外，为了验证 **PID** 方法在差分隐私场景下的不适用性，本实验还额外添加一组对照实验，其中对照组名为 **PID**，其使用 **PID** 方法对添加过 **DP** 噪声的超调噪声进行过滤；具体实验结果如下：

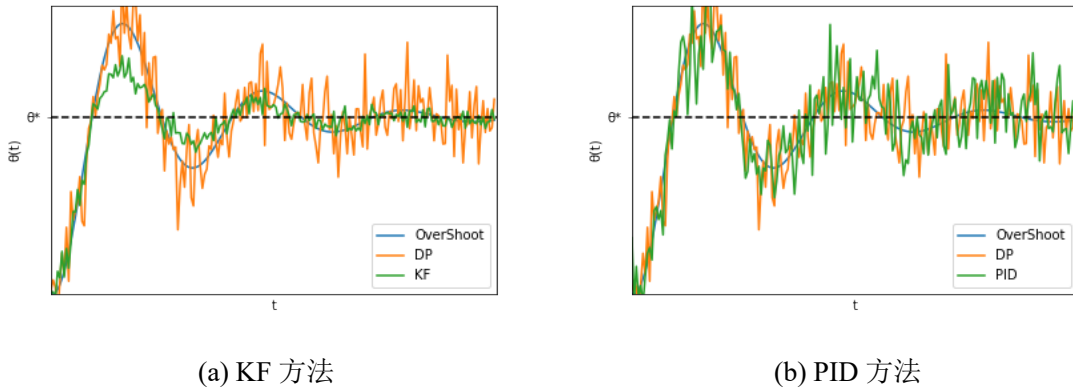


图 3-10 DP 场景下超调噪声过滤方法实验结果

如图3-10所示，首先，**OverShoot** 组的波动过程是一个方差指数衰减的线性波动过程，模型参数在最优参数 θ^* 上下不断波动，且波动范围不断变小，该过程对应着动量优化方法中的超调现象，随着模型的不断训练，超调噪声不断减小，直至使模型收敛至最优参数 θ^* ；其次，根据 **DP** 组的实验结果可以得知，由于差分隐私噪声的引入，进一步加剧了超调现象，尤其是在模型收敛到后期，但由于差分隐私噪声是无偏的，所以模型最终仍趋于收敛；然后，根据 **PID** 组的实验结果可以得知，**PID** 方法在添加差分隐私噪声的场景下无法有效过滤超调噪声，甚至会加重超调噪声的波动情况，因为其通过微分项引入了两倍的高斯噪声，进而加剧了超调问题的影响；最后，根据 **KF** 组的实验结果可以得知，卡尔曼滤波器在添加差分隐私的场景下可以有效过滤系统中的超

调噪声，降低超调噪声所造成的波动，进而缓解超调现象给模型更新过程带来的影响，该方法可以在模型收敛后期有效加快模型的收敛速率，并使得模型收敛至更高的精度。

2) 超调过滤方法实验

本节是对基于卡尔曼滤波器的超调过滤方法进行实验验证。与上节相同，同样采用三种不同的算法-数据集组合对方法进行实验。与上节实验的不同点在于，本节引入了动量优化方法。实验组名为 **KFParam** 算法，采用的是基于卡尔曼滤波器的超调过滤方法进行模型训练。除此之外，还有其他两组实验作为对比，对照组 1 是名为 **DPSGD-M** 的方法，采用基于动量优化方法的 **DPSGD** 算法，作为实验的基线；对照组 2 是名为 **PID** 的方法，在采用 **DPSGD-Momentum** 方法的基础上，进一步引入 **PID** 方法，用来验证 **PID** 方法在差分隐私的场景下的效果。实验结果如下：

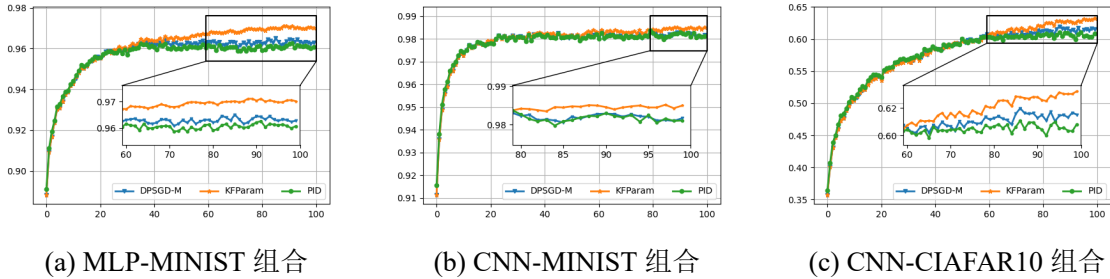


图 3-11 KF-Param 方法在不同算法 + 数据集组合下的实验结果

如图3-11所示, 相较于 **KFGrad** 方法, 由于动量更新方法的引入, 模型的收敛过程中出现了明显的抖动。并且由于 **KFParam** 算法主要作用于模型后期的收敛过程中所存在的超调现象, 而在模型训练的前期并未出现超调现象, 故在模型训练的前期, 三组实验的结果接近相同, 并未有较大的差异, 但在模型收敛的后期便出现了较大的差距。除此之外, 根据实验结果显示 **PID** 方法无法缓解 **DPSGD-M** 算法训练过程中的超调现象, 其效果甚至不如本实验的 **DPSGD-M** 算法。正如之前的分析所描述的一样, 由于 **PID** 方法借助两次梯度差值来缓解超调现象, 但是由于使用 **DPSGD** 方法更新时, 梯度中包含了额外的差分隐私噪声, 此时两次梯度差值造成了噪声的进一步累计, 进而较大程度上影响了 **PID** 方法的调节的作用。造成在差分隐私的场景下, 其不仅没有缓解动量优化方法的超调现象, 反而加剧了这种情况。**KFParam** 方法相较于其他两种方法, 在模型收敛的后期, 有效遏制住了超调现象对模型训练过程的影响, 使得模型精度并未出现明显波动, 进而加快模型的收敛速率的同时达到了更好的精度。并且引入了动量更新后, 本文发现其效果相较于传统的 **DPSGD** 方法有了更高的提升, 其中 **MLP-MINIST** 和 **CNN-CIAFAR10** 这两组实验的提升效果最大。该现象的原因本文分析和 **KFGrad** 方法中相同, 即在 **KFParam** 算法中, 任务越复杂, 效果越明显。

3) 不同动量系数下超调过滤方法实验

本小节主要测试不同动量系数 α 下 KFParam 算法的效果。动量优化方法引入了动量系数 α ，据前文分析，动量系数能在很大程度上影响超调噪声的方差，动量系数越大，则超调噪声越高。因此，为了更好地评估动量系数 α 对基于卡尔曼滤波器的超调过滤方法 KF-Param 的影响，本节针对不同的动量系数 α 取值 (分别选取 0.5、0.9，为相关实验的经验值^[57])，采用三种算法-数据集组合进行了实验，实验结果如下：

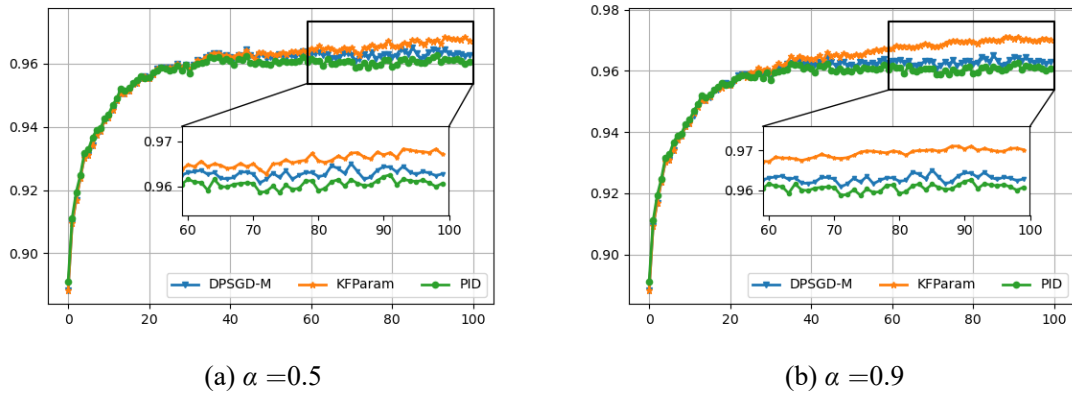
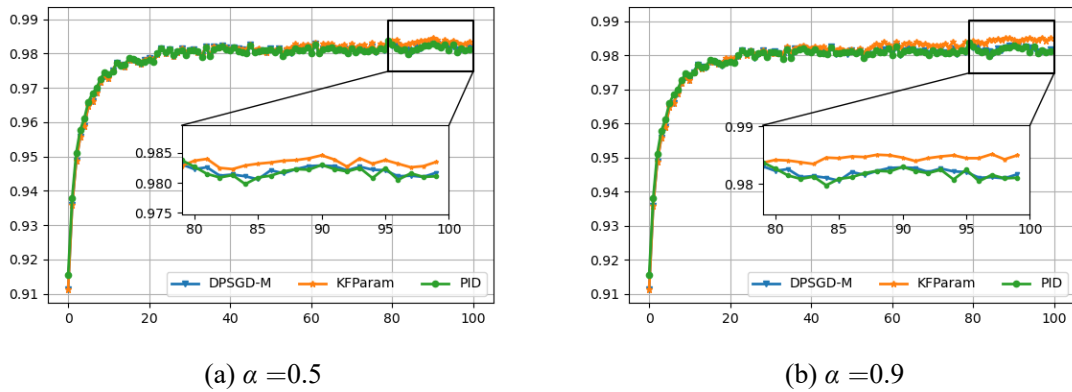
图 3-12 MLP-MINIST 组合下基于不同动量系数 α 的 KF-Param 方法

图 3-13 基于 CNN-MINIST 的 KF-Param 方法

如图3-12、图3-13和图3-14所示，首先根据三组不同的实验结果图像可以观察到，随着动量系数的增大，模型的抖动过程也随之增大，并且在模型收敛后期的超调现象对模型精度的影响也会随之增大，这是由于动量系数的增大使得模型参数中超调噪声也随之增大的原因所导致的。其次根据实验结果分析可以得出，动量系数的越大，模型训练速率会随之越大，但实验中出现超调问题的节点也会随之提前，其中 MLP-MINIST 这组实验中尤为明显，当动量系数为 0.9 时，在模型训练至 20 个 epoch 时便出现了超

调问题，比动量系数为 0.5 时提前了 20 个 epoch 出现超调问题。然后据实验结果显示，PID 方法在不同的动量系数下均无法有效过滤模型参数中的超调噪声，并给实验的最终精度带来了额外的损失，该结论进一步验证了 PID 方法不适用于差分隐私的机器学习优化场景。除此之外，可以观察到在三组实验中 KF-Param 算法相较于 KF-Grad 算法均有一定程度上的提升，分析认为这是降噪后的动量优化方法所带来的额外增益。最后综合实验结果可以分析得出，随之动量系数 α 的增大，KF-Param 算法的效果相较于 DPSGD-M 所带来的提升也随之增大。

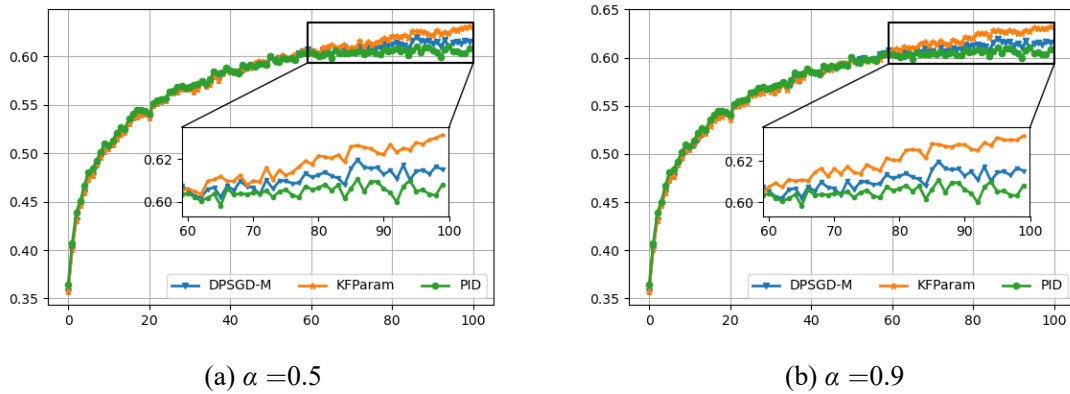


图 3-14 基于 CNN-CIFAR10 的 KF-Param 方法

基于上述的实验结果，可以分析得出如下四个结论：(1) DPSGD-Momentum 的训练方法在模型收敛的过程中会造成明显的抖动，即超调现象会明显加重；(2) 基于 DPSGD-M 的训练方法下，由于受到累计噪声的影响，PID 方法并不奏效，反而起到负面作用，加重超调现象；(3) KF-Param 方法相较于 PID 方法能更好的适用于缓解 DPSGD 场景下的超调现象；(4) 动量优化方法的动量系数越大，KF-Param 方法的效果越好。

3.3.4 隐私保护机器学习优化方法实验

本小节将对基于卡尔曼滤波器的隐私保护机器学习优化方法 (简称为 KF2 方法) 进行了实验验证。基于卡尔曼滤波器的隐私保护机器学习优化方法是 KF-Grad 算法和 KF-Param 两个算法的组合，故其自身也会具有其他两个算法的性质。本实验共有一个实验组和一个对照组，其中实验组名为 DPSGD-M-KF2 采用 KF2 算法进行模型的训练，对照组名为 DPSGD-M 算法采用基于动量更新算法的 DPSGD 算法训练。除此之外，为更完善地评估基于卡尔曼滤波器的隐私保护机器学习优化方法方法，本节在高斯噪声方差 σ 分别为 3 和 5 的情况下，对 KF2 方法的效果进行了细致评估。实验采用的评价指标为准确率，其他具体的基础参数设置为：高斯噪声的方差为 3、5，裁剪系数为 1，样本批大小为 50，动量系数 α 为 0.9，该实验的具体结果如下：

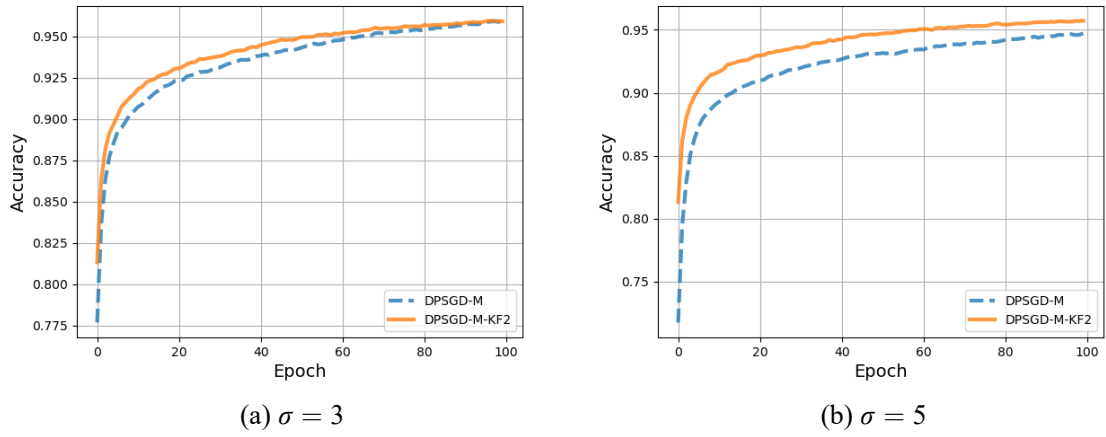


图 3-15 MLP-MINIST 组合下 KF2 方法基于不同噪声方差 σ 的实验结果

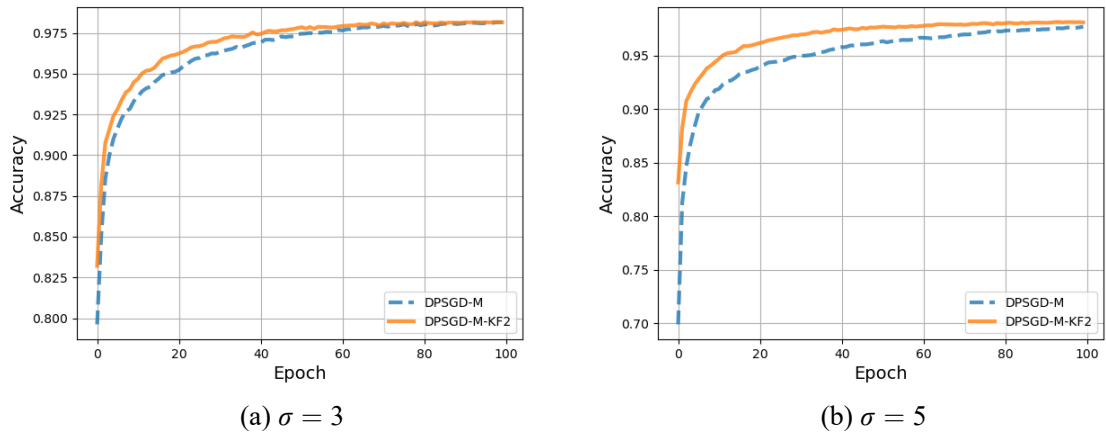


图 3-16 CNN-MINIST 组合下 KF2 方法基于不同噪声方差 σ 的实验结果

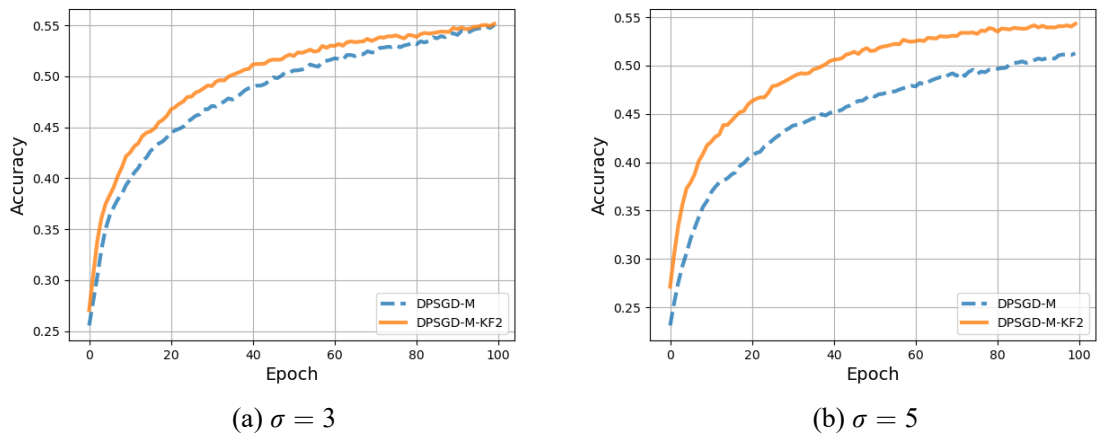


图 3-17 CNN-CIAFAR10 组合下 KF2 方法基于不同噪声方差 σ 的实验结果

如图3-15、图3-16和图3-17所示，基于卡尔曼滤波器的隐私保护机器学习优化方法 KF2 算法在三组不同的算法-数据集组合和不同的噪声方差下的模型收敛精度和收敛速率都要明显优于基于动量更新方法的 DPSGD 算法，根据对实验结果进一步的细致分析，可以得出如下三个结论：(1) KF-Grad 方法的引入能够帮助 KF-Param 方法更好的平滑动量优化方法所带来的超调现象，使得收敛过程的曲线变得更为平滑。(2) KF2 方法相较于传统的 DPSGD，大大加快了模型的精度，同时也能在一定程度上提升模型的最终效果。(3) KF2 方法与 KF-Grad 方法具有相似的性质，均更适用于高噪声的场景，向梯度添加的差分隐私高斯噪声方差越大，则 KF2 方法相对于 DPSGD-M 的提升也就越大。两个方法的结合后的基于卡尔曼滤波器的隐私保护机器学习优化方法，其相较于两个单独的方法，在收敛速率和最终收敛精度上均有了额外的提升。在基于差分隐私的机器学习场景中，更好地实现了隐私和效用之间的权衡，具有较强的实践意义。

3.4 本章小结

在基于差分隐私的机器学习模型优化过程中，为解决梯度波动较大的问题，本章提出了基于卡尔曼滤波器的梯度过滤方法，利用卡尔曼滤波器能够有效过滤高斯噪声的性质对梯度中的高斯噪声进行过滤，大幅度减小了梯度的波动，进而加快了模型的收敛速度以及最终的收敛精度。除此之外，为解决动量优化方法所造成的超调问题，本章提出了基于卡尔曼滤波器的超调过滤方法，利用卡尔曼滤波器对其模型中的超调噪声进行过滤，在模型收敛的后期降低其超调现象导致的波动，进而加快模型的收敛速率提升最终的模型精度。最后，将两个基于卡尔曼滤波器的方法进行结合，生成了基于卡尔曼滤波器的隐私保护机器学习优化方法，大大加快模型收敛速度的同时提升了模型的收敛精度，更好地实现了隐私和效用的权衡。

4 基于卡尔曼滤波器的隐私保护联邦学习算法研究

机器学习依赖于集中式数据管道，即数据需要先集中收集到中央服务器然后才能进行模型训练，然而该过程可能会造成严重的隐私泄露问题。去中心化的联邦学习方法能够在保证数据不出本地的情况下，通过各个相邻参与方之间互相交换模型参数来进行协同训练，进而能够在一定程度上保护用户的数据隐私。并且去中心化联邦学习方法由于不需要借助可信第三方用来做模型的统一聚合和分发，故不存在传统联邦学习中的信任依赖等一系列问题。然而，当前的去中心化的联邦学习方法仍然存在着系统难以达成共识的问题，进而给模型的收敛精度造成一定的损失。针对上述问题，本章提出了基于共识卡尔曼滤波器的去中心化联邦学习优化方法。首先将第三章提出的基于卡尔曼滤波器的隐私保护机器学习优化方法扩展至中心化的联邦学习场景下，然后进一步将其扩展至去中心化的联邦学习架构中，并且针对去中心化架构中系统在难以达成共识的问题，提出了基于共识卡尔曼滤波器的去中心化联邦学习优化方法，通过加快系统达成共识的速率，解决了去中心化联邦学习架构中所存在的共识问题，进而进一步提升了最终模型的收敛精度。

4.1 系统模型及问题定义

4.1.1 系统模型

本节主要描述横向联邦学习以及去中心化的联邦学习这两种联邦学习方法的系统模型。如图4-1所示，在横向联邦学习的系统场景中，存在 m 个参与方 $clinet_i$ (其中 $i \in (1, m)$) 以及一个聚合方。参与方所拥有样本的为 (x_j, y_j) ，其中 $j \in (1, n)$ ，即该参与方共有 n 个样本，聚合方的功能主要是负责模型的分发和聚合操作，其自身并不包含训练信息。在训练至第 t 个 epoch 时，各个参与方将在本地训练完的模型参数 $\omega(i, t)$ 上传至第三方聚合方；然后第三方聚合方将模型进行聚合平均，具体聚合过程如式(4-1)所示。然后可信第三方将平均后的参数 $\hat{\omega}_t$ 重新下发给各个参与方，直至最终模型收敛。

$$\hat{\omega}_t = \frac{\sum_{i=1}^n \omega(i, t)}{n} \quad (4-1)$$

如图4-1所示，在去中心化网络拓扑中，各个参与方之间的通信网络为无向连接的拓扑图 $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{W})$ ，其中 $\mathcal{N} = 1, 2, \dots, K$ 代表各个参与方的集合，共 K 个参与方。 $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ 代表通信通道的集合，每个通道链接两个不同的参与方。对于每条边 $(i, j) \in \mathcal{E}$ ，都对应着邻接矩阵 \mathcal{W} 中的一个位置，当 $\mathcal{W}_{ij} = 1$ 时表示参与方 i 和参与方 j 之间存在有通信信道， $\mathcal{W}_{ij} = 0$ 时则表示参与方 i 和参与方 j 之间不存在通信信道。对于

参与方 i , 当 $\mathcal{W}_{ij} = 1$ 时, 则参与方 j 和参与方 i 互为对方的邻居, 参与方 i 所有邻居的集合表示为 \mathcal{N}_i , 即 $\mathcal{N}_i = \{j | \mathcal{W}_{ij} = 1\}$ 。最终的评价指标选取各个参与方准确率的均值作为结果。在去中心化联邦学习场景中, 往往要求通信拓扑结构是连通图, 即对于图中的任意两个参与方 u 和 v , 均存在一条顶点序列 $u, v_1, v_2, \dots, v_n, v$, 其中 v_i 是序列中的第 i 个顶点, 则该图就是连通图。在去中心化的联邦学习架构下, 每个参与方 j 所拥有样本的为 (x_j, y_j) , 其中 $j \in (1, m)$, 本地训练模型参数为 ω_j 。

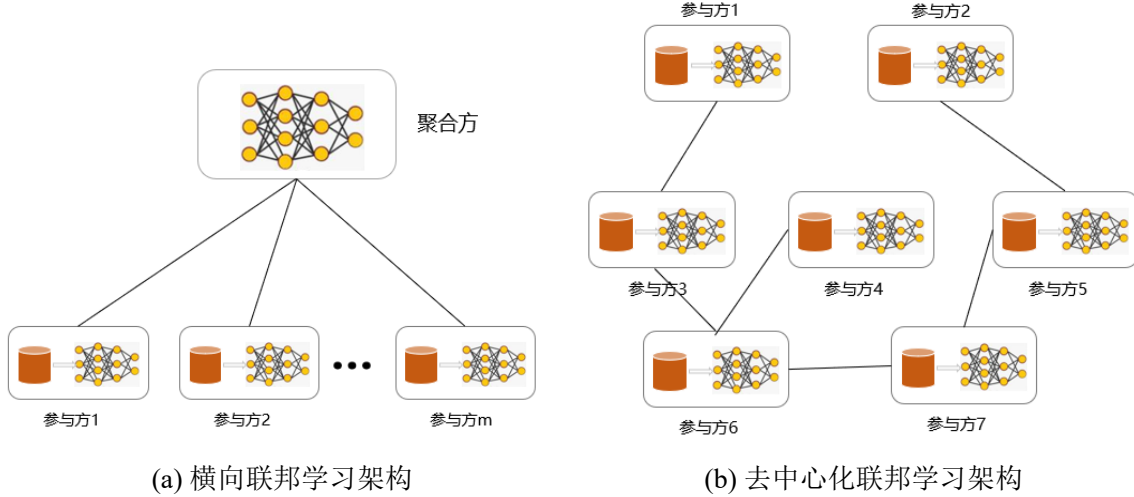


图 4-1 联邦学习架构

在去中心化联邦学习架构中, 并不存在任何的可信第三方, 每个参与方节点仅需与其相邻节点进行数据的交换, 如式(4-2)所示, 单次聚合时, 参与方 i 将相邻节点的信息以及自身的信息做平均聚合作为本次迭代的结果。

$$\omega_i(t+1) = \sum_{j \in \mathcal{N}_i \cup i} \omega_j(t) \quad (4-2)$$

尽管仅需和相邻节点交互, 但由于去中心化联邦学习架构中通信拓扑结构是连通图, 所以参与方 i 的信息可以直接或者间接地将信息传至所有其他参与方中。

4.1.2 问题定义

首先从理论上阐述去中心化联邦学习的优化问题, 假设有 K 个客户端, 其本地数据用于训练机器学习任务 \mathcal{D}_k , 其中 $k \in 1, 2, \dots, K$, 其中 \mathcal{D} 为模型自身的参数值, $\mathcal{D}_c \triangleq \cup_k \mathcal{D}_k$, 整体上去中心化的联邦学习系统的任务可以表述为如下的过程:

$$\mathcal{M}_c \triangleq \operatorname{argmin} F(\mathcal{D}; \mathcal{M}) \quad (4-3)$$

如式(4-3)所示, 各参与方使用相同的目标函数 F 来协作训练模型 \mathcal{D}_k , 并且根据 He 等人^[60]的分析, 可以证明在拓扑结构为连通图的情况下, 各参与方通过和相邻节点间交换信息便可近似获得全局的信息 \mathcal{M} , 使得系统最终完成收敛。但因为各参与方在分散系统中难以快速达成共识, 导致在更新时各参与方分别基于自身参数而向着不同的梯度方向更新, 进而大大影响了去中心化联邦学习架构中的收敛速率。

若要加速去中心化联邦学习架构的训练速度, 需要保证的是提升系统达成共识的速度以及参与方自身训练模型的速度, 以防止各参与方基于不同的自身参数向着不同的梯度方向更新。式(4-4)给出了系统达成共识的形式化表述, 即当每个参与方的模型 $\omega_i(t)$ 和全局所有参与方模型的平均 $\frac{1}{N} \sum_{j \in \mathcal{N}=1} \omega_j(t)$ 相同时, 系统便达成了共识。

$$\lim_{t \rightarrow \infty} (\omega_i(t) - \frac{1}{N} \sum_{j \in \mathcal{N}=1} \omega_j(t)) = 0 \quad (4-4)$$

其中系统达成共识的速率受拓扑结构影响较大, 网络拓扑越稠密即去中心化系统各个节点连接度越高, 信息交换的效率越高, t 值也随之越小。在去中心化的联邦学习的场景中, 各参与方之间的拓扑结构往往是较为固定且稀疏的, 因此, 如何加快参与方自身的训练速率并加快系统达成共识的速率是加速去中心化联邦学习方法的关键。

4.2 基于卡尔曼滤波器的隐私保护联邦学习算法

4.2.1 联邦架构下基于卡尔曼滤波器的联邦学习优化方法

本节将本文第三章所提出的三个基于卡尔曼滤波器的优化算法扩展至中心化的联邦学习架构下, 并给出具体的迁移方法。如图4-2所示, 在中心化联邦学习架构下, 系统中存在 n 个参与方以及一个可信第三方。可信第三方将各参与方本地训练更新后的最新结果进行平均聚合, 并将聚合后的最新参数分发给每一个参与方, 可信第三方的统一分发使得中心化的联邦学习架构中并不存在所谓的共识问题。

本文第三章所提出的基于卡尔曼滤波器的梯度过滤算法 KF-Grad 算法其只作用于自身样本所计算的梯度信息, 无需借助其他参与方信息, 因此本方法可以直接迁移至联邦学习的场景下, 即将参与方的本地更新算法替换为基于卡尔曼滤波器的梯度过滤算法, 便直接得到了中心化联邦学习场景下基于卡尔曼滤波器的梯度过滤算法。不同参与方之间使用各自的卡尔曼滤波器而无需进行协同, 这等价于在集中式场景下, 将整体样本分为多个相互独立的训练模块, 因此无需作额外的补充。

基于卡尔曼滤波器的超调过滤算法由于引入了动量优化方法, 而动量更新方法在联邦学习过程中需要进行全局聚合分发, 故无法像 KF-Grad 算法一样直接迁移。这是因为在联邦学习场景下, 可信第三方会对模型进行统一的聚合和分发, 所以在统一分发

后各个参与方的模型完全相同，但是受到自身数据分布和随机采样的影响，各个参与方的历史动量信息可能存在较大的差别，若不将动量信息也随参数进行聚合，则会造成在新的模型参数上使用旧的动量信息训练的现象，进而影响最终模型的效果。所以将基于卡尔曼滤波器的超调过滤方法迁移至联邦学习场景时，需要先将动量更新方法联邦化，Liu 等人^[61]将动量更新方法迁移至联邦学习的场景下，并进行了收敛性分析。与集中式动量优化方法不同的是，联邦学习场景下，动量优化方法也需要在固定训练间隔后，上传至可信第三方进行聚合并分发，以此来统一整个系统中的动量信息，确保整个系统朝着统一方向收敛。基于此方法，在每次聚合分发后，各个参与方的起点 (模型参数) 以及当前累计动量均达成一致，故能进一步加快联邦学习场景下模型的收敛速率。具体过程如下：

$$d(t) = \frac{\sum_{i=1}^N |\mathcal{D}_i| \hat{d}_i(t)}{|\mathcal{D}|} \quad (4-5)$$

如式(4-5)所示，其中 \mathcal{D} 为全量数据集，而 \mathcal{D}_i 代表参与方 i 的数据集；其中 $\hat{d}_i(t)$ 代表了参与方 i 此时的动量信息，而 $d(t)$ 则是经过可信第三方聚合后的全局动量信息。各个参与方按照自身数据占比来对动量进行聚合以作为全局梯度。综上所述，联邦架构下基于卡尔曼滤波器的超调过滤方法上传和分发的信息中不仅包含模型的参数信息，还应该额外包含着每个参与方自身的累计动量信息。

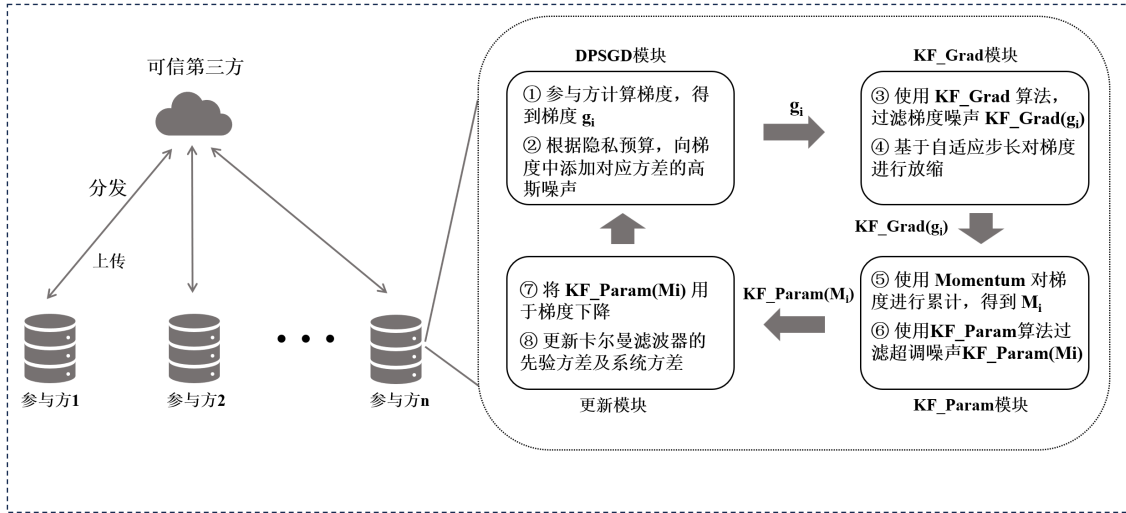


图 4-2 联邦架构下基于卡尔曼滤波器的隐私保护机器学习优化方法

经过上述对动量优化方法的联邦化变换，便可将 KF-Param 方法以及 KF2 方法均迁移到中心化联邦学习架构下，进而加快中心化联邦学习场景中各个参与方自身的收敛速率以及收敛精度。并且由于该系统中可信第三方仅仅只使用平均聚合的方法，则当各个节点的速率和精度均提升后，整个联邦学习系统的收敛速率以及收敛精度也会随之提升。除此之外，更快的收敛速率意味着更少的通信轮次，这不仅仅缓解了联邦学

习中的通信瓶颈问题，还通过减少数据交换轮数进而进一步降低了隐私泄露风险。

如图4-2所示，该图描述了联邦学习架构下基于卡尔曼滤波器的隐私保护机器学习算法的流程，该算法通过加快了各个参与方的收敛速率，进而在一定程度缓解了通信压力；并且由于各个参与方本地采用 DPSGD 算法进行模型训练，故第三方无法得知各参与方的准确消息，进而在一定程度上也缓解了信任依赖的问题。

4.2.2 去中心化联邦架构下的梯度过滤方法

Olfati-Saber 等人^[55]对共识卡尔曼滤波器做了严格的理论分析，相较于传统的卡尔曼滤波器，共识卡尔曼滤波器常被用作在存在高斯噪声的去中心化分散互联系统中，进行数据的聚合和过滤工作，除了能过滤系统中存在的高斯噪声外，共识卡尔曼滤波器还能够加快系统中各参与方达成共识的速率，这对于分散系统的快速收敛至关重要。因为在去中心化的联邦学习系统中，受到随机采样和数据分布异质性的影响，各参与方每一步所更新的梯度可能相差较大，若在系统未达成共识的前提下进行更新，则会减缓系统收敛速率，造成系统精度的损失。所以基于卡尔曼滤波器的梯度过滤方法并不能直接适用去中心化联邦学习架构。为此，本节将 KF-Grad 方法扩展至去中心化架构下基于共识卡尔曼滤波器的梯度过滤方法 (简称 KCIF-Grad 方法)。首先将卡尔曼滤波器过滤扩展至共识卡尔曼滤波器 (KCIF 滤波器) 并给出算法：

算法 4-1 KCIF 滤波器

输入：各相邻参与方梯度 $g_{i,1}, g_{i,2}, \dots, g_{i,n_i}$ ，数据占比 H_i ，邻居 N_i ， $J_i = N_i \cup i$ ，先验方差 $P_i(0) = P_0$ ， $\bar{x}_i(0) = 0$ ，观测方差 $R_i(t)$ ，系统方差 $Q(t)$ ，衰减系数 η ，步长 β ，相邻节点信息 $msg_j(t) = u_j(t), U_j(t), \bar{x}_j$

输出：后验聚合梯度 $\hat{\theta}_i$

- 1 计算 msg_i : $u_i(t) = H_i(t)g_i(t)/R_i(t)$, $U_i = (H_i(t))^2/R_i(t)$;
- 2 传递 msg_i : 将 $msg_i = u_i(t), U_i(t), \bar{x}_i$ 传递给各相邻节点;
- 3 接收 msg_j : 从各个相邻节点接收信息 $msg_j(t) = u_j(t), U_j(t), \bar{x}_j$ ，其中 $j \in N_i$;
- 4 聚合信息: $y_i(t) = \sum_{j \in J_i} U_j(t)$, $Y_i(t) = \sum_{j \in J_i} U_j(t)$;
- 5 计算后验估计方差: $M_i(t) = 1/(1/P_i(t) + Y_i(t))$;
- 6 计算共识增益: $\gamma = \beta/(|P_i(t)| + 1)$, $C_i(t) = \gamma P_i(t)$;
- 7 计算后验估计: $\hat{x}_i(t) = \bar{x}_i(t) + M_i(t)(y_i(t) - Y_i(t)\bar{x}_i(t)) + C_i(t) \sum_{j \in J_i} (\bar{x}_j(t) - \bar{x}_i(t))$;
- 8 计算先验估计估计: $\bar{x}_i(t) = A \cdot \hat{x}_i(t)$;
- 9 更新先验估计方差: $P_i(t) = A^2 M_i(t) + Q$;
- 10 更新系统方差: $Q(t+1) = \eta Q(t)$;

如算法 4-1 所示，针对算法的输入部分，其中 H_i 为各个参与方的样本量占比，参与方 i 的相邻节点集合为 N_i ，然而在 KCIF 方法中，每次聚合的信息不仅包含相邻节点的信息，还包含自身的信息，即 $J_i = N_i \cup i$ ；先验方差 $P_i(0) = P_0$ ，在真实场景中，先验方差往往是难以获得的，但根据卡尔曼滤波器的更新规则可以分析得到，先验方差的大小仅在前几个轮次中对系统的影响较大，随着该过程的不断进行，先验方差会迅速收敛到某一个固定值；观测方差 $R_i(t)$ 为所添加的高斯噪声方差和随机采样方差，其受

隐私预算和批大小的影响；步长 β 为控制共识程度的超参，该参数的取值越大则系统达成共识的速度就会越快，但同时也会导致系统训练不够充分的问题，当步长为 0 时，共识卡尔曼滤波器将会退化成传统的卡尔曼滤波器。该算法首先各个参与方将自己的数据计算打包发送给自己的相邻节点，同时也接收相邻节点传输来的信息，由于通信拓扑图是连通图，故至少会接受到一个其他参与方给它发来的信息；等到传输结束后，各个节点开始聚合自己所收到的信息，其中也包含自身的信息；之后便是基于系统方差以及步长等参数计算自己的后验估计方差和共识增益；然后计算后验估计，该过程主要包含两部分，一部分为过滤部分，用来过滤系统中存在的高斯噪声，而另一部分是共识部分，其主要用来加快系统达成共识的速率，以促进系统收敛；最后便是进行系统参数的更新，更新先验估计方差和系统方差。

现将 KCIF 滤波器引入 DPSGD 算法的训练过程中，给出具体的去中心化架构下基于共识卡尔曼滤波器的梯度过滤算法，该方法的过程与 KF-Grad 方法的主要区别在于，借助共识卡尔曼滤波器的共识项，在模型训练的过程中协同过滤并聚合梯度信息，而非单独过滤自身梯度。共识卡尔曼滤波器的这种协同过滤方法在某种程度上进一步增加了训练过程中每一步梯度的准确性。其具体算法过程如下：

算法 4-2 KCIF-Grad 算法

输入：各参与方样本 $x_{i,1}, x_{i,2}, \dots, x_{i,n_i}$, 损失函数 $\mathcal{L}(\theta) = \sum_{i=1}^n \mathcal{L}(\theta, x_i)$, 学习率 η_t , 噪声方差 σ , 批大小 L , 梯度裁剪系数 C

输出：模型参数 $\theta_{i,T}$

- 1 随机初始化模型参数 θ_0 , KCIFGrad 滤波器；
- 2 **while** $t < T$ **do**
- 3 参与方 i 以 L/N 的概率从自身的样本中随机抽取样本集合 L_t ；
- 4 计算梯度；
- 5 针对 L_t 中的每个样本，计算其梯度 $g_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$ ；
- 6 梯度裁剪；
- 7 $\bar{g}_t(x_i) \leftarrow g_t(x_i) / \max(1, \frac{\|g_t(x_i)\|_2}{C})$ ；
- 8 添加噪声；
- 9 $\tilde{g}(t) \leftarrow \frac{1}{L} (\sum_i \bar{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I))$ ；
- 10 KCIF 过滤；
- 11 $\theta_{t+1} \leftarrow \theta_t - \eta_t \text{KCIF}(\tilde{g}(t), \eta = 1)$ ；
- 12 梯度下降；
- 13 $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{g}(t)$ ；
- 14 **end**

如算法 4-2 所示，每个参与方首先从自己的样本中随机抽取部分样本用于计算梯度；其次对梯度进行裁剪，裁剪的目的是为了限制全局敏感度的大小；然后往模型中添加满足差分隐私的高斯噪声；之后使用 KCIF 方法进行过滤，通过与相邻节点的交互，进而得到更为准确的梯度估计，并同时过滤其中的高斯噪声。此过程中的共识卡尔曼滤波器 KCIF 还额外被传入了另一个参数，即系统方差衰减系数 $\eta = 1$ ，在此场景下，系统方差并不需要进行衰减，故其衰减系数设置为 1；最后基于聚合并过滤后的梯度，各

个参与方分别进行梯度下降更新。KCIF-Grad 算法与 KF-Grad 算法的相比,二者均具有过滤动态系统中差分隐私噪声的功能,进而保留了 KF-Grad 算法中加快模型收敛速率并提升收敛精度的效果;并且额外引入了共识项 β ,通过加快去中心化的联邦学习架构中各参与方之间达成共识的速率,进而进一步加快了模型的收敛速度以及系统最终的收敛精度。除此之外,相较于传统的联邦学习,该方法仅需和相邻节点交换先验信息,而无需进行全局广播和全局聚合,此性质可以在很大程度上缓解联邦学习的通信压力,解决了通信瓶颈和单点故障的问题,更加适用于真实的工业界场景。

4.2.3 去中心化联邦架构下的超调过滤方法

本节将介绍去中心化架构下基于共识卡尔曼滤波器的超调过滤方法(简称 KCIF-Param 算法)。正如 2.2 节所描述的那样,差分隐私的引入会降低模型的收敛速率和精度。同时,去中心化联邦学习架构由于没有第三方进行全局聚合和分发,故相较于集中式的训练方法,去中心化的联邦学习架构下需要更多的训练轮数才能确保模型的收敛,所以其更需要动量优化方法来为模型的训练过程进行加速。因此有必要将第三章所提出的基于卡尔曼滤波器的超调过滤方法迁移至联邦学习的场景中。然而在联邦学习的场景下,动量信息的延迟性以及参与方分布的异质性可能会造成超调现象的进一步加剧,延缓系统的训练时间,并且影响模型最终的训练效果。

在联邦学习或者去中心化的联邦学习系统中,每一步计算梯度时所随机抽样的样本并不一致,进而导致各个参与方每一步算得的梯度存在较大差异,并且各个参与方的数据分布也并非完全一致,这可能造成模型在不同参与方之间收敛困难^[62-63]。该问题也会导致使用动量优化方法时各个参与方的动量梯度存在较大差异,尤其是在去中心化联邦学习架构下,由于其自身拓扑结构大大限制了信息传递速度,参与方 i 的动量梯度信息可能需要更多的很多次迭代才能传达至参与方 j ,并且拓扑矩阵的稀疏度越高,数据分布受相邻性影响越大(距离越近,其模型信息越相似),这种限制就越强烈,甚至会造成系统一直无法达到共识的情况。基于此,如果直接将 KF-Param 算法迁移至去中心化的联邦学习场景,动量信息的延迟性以及参与方分布的异质性可能会造成超调现象的进一步加剧,剧烈的抖动又会很大程度上延缓系统达到共识所需的时间,进而大大影响 KF-Param 方法效果。为了缓解上述问题,本节将 KF-Param 算法中的方差指数衰减的卡尔曼滤波器替换为方差指数衰减的共识卡尔曼滤波器。通过利用共识卡尔曼滤波器中的共识项,加快动量梯度在去中心化联邦学习架构中的传递速度,同时加速各个参与方达成共识的速率,进而更好地发挥动量优化方法的加速效果。除此之外,针对各个参与方数据分布异质性以及分布相邻性的问题,本方法通过统一所有参与方的系统方差以及系统观测方差,进而防止各个分布相似的簇(即通信拓扑图中连通率较高的子图)影响达成全局共识的速率。

针对去中心化架构下引入动量优化方法造成超调问题加剧的情况，KCIF-Param 算法通过将方差指数衰减的卡尔曼滤波器更替为方差指数衰减的共识卡尔曼滤波器，将去中心化联邦学习架构中各个参与方的动量梯度进行动态聚合过滤，以降低梯度分布异质性和数据分布异质性给模型聚合过程带来的影响并缓解超调现象的波动。首先需要将基于联邦学习的动量优化方法迁移至去中心化的联邦学习架构下，针对某一参与方 i ，其具体的动量更新过程如下：

$$d(t) = \frac{\sum_{i \in \mathcal{N}_i \cup i} |\mathcal{D}_i| \hat{d}_i(t)}{|\mathcal{D}|} \quad (4-6)$$

如式(4-6)所示，其中 \mathcal{D} 为参与方 i 及其所有邻居数据集的并集。而 \mathcal{D}_i 代表参与方 i 的数据集；其中 $\hat{d}_i(t)$ 代表了参与方 i 此时的动量信息，而 $d(t)$ 则是相邻聚合后的动量信息。各个参与方按照自身数据占比来对动量进行聚合以作为全局梯度。现在将 KCIFGrad 滤波器引入 DPSGD-Momentum 方法的训练过程中，具体算法如下：

算法 4-3 KCIF-Param 算法

输入： 样本 x_1, x_2, \dots, x_n , 损失函数 $\mathcal{L}(\theta) = \sum_{i=1}^n \mathcal{L}(\theta, x_i)$, 学习率 η , 噪声方差 σ , 批大小 L , 梯度裁剪系数 C , 以及系统方差衰减系数 η

输出： θ_T , 以及隐私损耗 (ϵ, δ)

- 1 随机初始化模型参数 $\theta_0, \mathcal{K}(2C, \sigma)$;
- 2 **while** $t < T$ **do**
- 3 计算梯度；
- 4 针对 L_t 中的每个样本，计算其梯度 $g_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$ ；
- 5 梯度裁剪；
- 6 $\bar{g}_t(x_i) \leftarrow g_t(x_i) / \max(1, \frac{\|g_t(x_i)\|_2}{C})$ ；
- 7 添加噪声；
- 8 $\tilde{g}(t) \leftarrow \frac{1}{L} (\sum_i \bar{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I))$;
- 9 Momentum 累计；
- 10 $V_{t+1} = \alpha V_t - \gamma \eta \tilde{g}(t)$ ；
- 11 梯度下降；
- 12 $\theta_{t+1} \leftarrow \theta_t + V_{t+1}$ ；
- 13 KCIF-Param 方法过滤；
- 14 $\theta_{t+1} = \text{KCIFParam}(\hat{\theta}_{t+1}, \eta)$ ；
- 15 **end**

如算法4-3所示，先进行样本采样用来计算随机梯度，其中采样的样本数量便是批大小；然后进行梯度裁剪以及添加满足差分隐私条件的高斯噪声，此步骤的是为了确保算法满足差分隐私要求；之后进行动量的累计，此步累计的噪声除了差分隐私噪声和随机梯度噪声外，还多了数据分布异质性导致的偏差噪声；即在此场景下，由于各个参与方所包含的样本数量不一定相同，且各个参与方所包含的数据分布也不一定相同，故各个参与方都会存在一定量的偏差，此偏差也会随着 Momentum 动量的累计而进行累计，并且由于联邦学习的架构是去中心化的，导致系统可能无法及时的同步所有更

新,这也进一步加重了此种现象的出现。这也是为什么在去中心化联邦学习的场景下,动量优化方法所引发的超调现象进一步加剧的原因之一。之后,便是基于累计动量进行梯度下降,然后使用共识卡尔曼滤波器方法对更新后的参数进行过滤,以过滤其中所存在的超调噪声,进而缓解模型收敛后期所出现的超调现象,进一步加快模型的收敛速率。除此之外,相较于 KCIF-Grad 算法使用的共识卡尔曼滤波器, KCIF-Param 算法所使用的滤波器,其系统方差衰减指数 η 不再为 1,该参数能够在一定程度上反应出来系统收敛的情况, $\eta \in (0, 1)$,若 η 过大,则会造成该方法将在训练阶段正常的参数波动当做是噪声,并且尝试将其过滤掉,尤其是在寻优阶段达到鞍点处时,由于其往往伴随较大的模型参数波动,过大的 η 值很有可能会将模型限制在一个局部最优点处,进而造成模型效果的损失。若 η 过小,则超调现象得不到合理抑制,由于噪声的不断累计可能会导致模型一直抖动而不收敛,进而大大降低了该方法的增益效果,而当其为 0 时,该方法退化至动量优化方法。在实验过程中需在经验值的范围内进行微调,观测其效果并进行评估。同样,本方法也引入了共识项 β ,该共识项的目的在于控制各参与方分布异质性的程度。该项值越大,则说明该去中心化联邦学习架构中各个参与方分布的异质性越强;反之同理。

本方法针对去中心化联邦学习架构中超调问题加剧的情况,通过将卡尔曼滤波器扩展至共识卡尔曼滤波器,并将动量优化方法扩展至去中心化联邦学习架构下,进而将基于卡尔曼滤波器的超调过滤方法过渡到了去中心化联邦学习架构下的基于共识卡尔曼滤波器的超调过滤方法,通过额外引入共识项,加快了在分散互联的去中心化的联邦学习架构中各参与方动量信息达成共识的速率,进而加快了模型的收敛速率的同时也提升了最终的收敛精度,更好地实现了隐私和效用的权衡。

4.2.4 去中心化联邦架构下的联邦学习优化方法

在去中心化的联邦学习架构下,将本章节提出的两个算法(基于共识卡尔曼滤波器的梯度过滤方法和基于共识卡尔曼滤波器的超调过滤方法)进行结合,形成了去中心化联邦学习架构下的隐私保护联邦学习优化算法(简称为 KCIF2 算法)。与 KF2 算法同理,由于基于共识卡尔曼滤波器的梯度过滤方法和基于共识卡尔曼滤波器的超调过滤方法所作用的对象并不相同,其分别作用于过滤梯度的高斯噪声以及过滤参数的超调噪声,故可以将其两个进行结合,进而进一步加快模型的训练速率,提升模型的精度。

基于共识卡尔曼滤波器的梯度过滤方法其主要作用于模型训练过程中的各个参与方的梯度信息,主要是在聚合的过程中使用共识卡尔曼滤波器来过滤 DPSGD 训练过程中的高斯噪声,同时其额外的共识项确保各个参与方尽快达成共识。在不降低隐私预算的前提下,通过利用相邻两次梯度的时间相关性以降低梯度中的噪声方差,进而加快模型的收敛速率并提高模型最终收敛的精度;而基于共识卡尔曼滤波器的超调过滤

方法，则是针对由于差分隐私和动量更新所造成超调现象加剧的问题，使用共识卡尔曼滤波器来作用于模型训练过程中的模型参数，来过滤在使用动量优化方法过程中引发超调现象的噪声，以缓解动量优化方法所带来的超调现象。随着模型不断地收敛，模型参数中所包含的超调噪声也在不断地指数级衰减，与之相对应的，共识卡尔曼滤波器的系统方差也不断随之进行指数级衰减，以适应该线性动态过程的变化，更好地过滤其中的超调噪声。将二者结合后具体算法过程如下：

算法 4-4 KCIF2 算法

输入：样本 x_1, x_2, \dots, x_n , 损失函数 $\mathcal{L}(\theta) = \sum_{i=1}^n \mathcal{L}(\theta, x_i)$, 学习率 η , 噪声方差 σ , 批大小 L , 梯度裁剪系数 C

输出： θ_T , 以及隐私损耗 (ε, δ)

- 1 随机初始化模型参数 $\theta_0, \mathcal{K}(2C, \sigma)$;
- 2 **while** $t < T$ **do**
- 3 计算梯度;
- 4 针对 L_t 中的每个样本, 计算其梯度 $g_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$;
- 5 梯度裁剪;
- 6 $\bar{g}_t(x_i) \leftarrow g_t(x_i) / \max(1, \frac{\|g_t(x_i)\|_2}{C})$;
- 7 添加噪声;
- 8 $\tilde{g}(t) \leftarrow \frac{1}{L} (\sum_i \bar{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I))$;
- 9 使用 KCIF-Grad 算法过滤噪声;
- 10 $\theta_{t+1} \leftarrow \theta_t - \eta_t \text{KCIFGrad}(\tilde{g}(t))$;
- 11 Momentum 累计;
- 12 $V_{t+1} = \alpha V_t - \gamma \eta_t \tilde{g}(t)$;
- 13 梯度下降;
- 14 $\theta_{t+1} \leftarrow \theta_t + V_{t+1}$;
- 15 KCIF-Param 方法过滤;
- 16 $\theta_{t+1} = \text{KCIF}(\hat{\theta}_{t+1}, \eta)$;
- 17 **end**

如上述算法4-4所示，在去中心化联邦学习架构的模型训练过程中，单个参与方在每一步计算完梯度后，需要通过共识卡尔曼滤波器算法和自身的相邻节点互相交换梯度相关信息，待信息交换完毕后，参与方对相邻节点和自身的梯度信息同时进行聚合，在聚合的过程中完成高斯噪声的过滤并促进分散系统尽快达成共识。此过程可以在确保系统尽快达到共识的同时加快了模型的收敛速率，并且还可以降低累计动量的方差。之后，各个参与方基于过滤和聚合后的梯度各自进行动量梯度的累计，然后同样经过共识聚合，以降低数据分布异质性和超调现象的加剧，与基于共识卡尔曼滤波的梯度过滤方法不同的点在于，基于共识卡尔曼滤波的超调过滤算法不仅仅需要通过共识卡尔曼滤波器和其他参与方进行数据交互，还需要额外和参与方直接进行动量平均，这与将动量更新方法迁移至中心化联邦学习架构中的方法思想上一致。之后基于过滤聚合后的动量进行更新，最后便是通过 KCIF-Param 方法来缓解数据异质性分布以及动量优化方法所带来的超调现象。综上，这两个去中心化联邦架构下基于共识卡尔曼滤波器的过滤方法之间的具有较强的互补性质，其中去中心化的梯度过滤方法可以有效降

低每一步梯度信息的波动，进而减缓超调现象的出现，而去中心化的超调过滤方法不仅可以在一定程度上缓解数据异质性分布的问题，并且大大加快了整个系统的收敛速率，两者在去中心化架构下更好地实现了隐私效用权衡。

4.3 实验结果与分析

本节首先介绍实验的各种设置信息以及联邦实验的模拟方法，其次在中心化联邦学习场景下，针对中心化联邦学习架构下基于卡尔曼滤波器的隐私保护机器学习优化方法进行了实验验证；然后针对本章所提出的三个优化方法（即 KCIF-Grad 算法、KCIF-Param 算法以及 KCIF2 算法），在去中心化联邦学习架构下进行了系统化的实验验证，并分析各个算法在不同参数组合以及不同的拓扑结构下这三个算法的性质和优势。

4.3.1 实验设置

在数据集方面，本章主要采用的数据集与第三章相同，均为 MINIST 数据集和 CIFAR10 数据集这两个图片数据集；并且针对这两个数据集所采用的算法也相同，同样形成了三组不同的模型-数据集组合，即 MLP-MINIST、CNN-MINIST 和 CNN-CIFAR10。除此之外，具体的实验环境配置也与 3.3.1 相同，在此不再赘述。

在本节实验中，采用模拟联邦学习场景方法进行实验，参与方的数量为 20，并将数据集随机均匀划分至 20 个参与方中。在去中心化的联邦学习场景下，由于算法要求网络拓扑结构必须为双向连通图，故为了简化问题，将所有初始的参与方都串联起来。其他所有非链接位置都以 φ 概率联通， $\varphi \in (0, 1)$ 称为稀疏度。当 φ 为 1 时，该拓扑结构为全连通图，该算法也随之等价于 FedAVG；当 φ 为 0 时，该拓扑结构为一条线，所有参与方都是这条线是某个点，故本实验的参与方最少有两个相邻节点。

4.3.2 基于卡尔曼滤波器的联邦学习优化方法实验

本节将扩展至中心化联邦学习场景下的基于卡尔曼滤波器的梯度过滤算法、基于卡尔曼滤波器的超调过滤算法以及基于卡尔曼滤波器的隐私保护机器学习优化算法，同时进行实验验证，以证明这些基于卡尔曼滤波器的优化方法在联邦学习场景下也具有一定的普适性。在联邦学习的训练过程中，每次训练随机抽取其中 20% 的参与方执行本地更新，被选中的参与方在本地更新 10 个 epoch，然后将更新后的参数发送给聚合方，聚合方经过聚合后分发给各个参与方。该实验的基础参数设置为：高斯噪声的方差为 5，裁剪系数为 1，批大小为 50，动量系数 α 为 0.6。

本实验首先基于 MLP-MINIST 的这组数据集算法组合进行实验，采用的评价指标为准确率，即将所有参与方节点的准确率的均值作为最终的评价指标。通过在集中式场景下模拟联邦学习的过程进行不同参与方之间的协同训练，并同时绘制出其在测试

集上的平均准确率变化曲线。该实验的具体结果如图4-3所示：

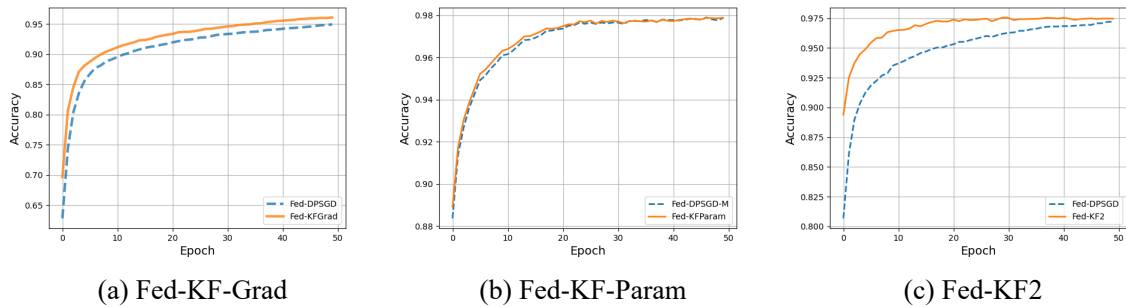


图 4-3 联邦卡尔曼滤波器优化方法实验结果

本实验的实验组分别采用联邦架构下的 KF-Grad 算法 (名为 Fed-KFGrad)、KF-Param 算法 (Fed-KFParam) 以及 KF2 算法 (Fed-KF2)，作为对照，这三个实验的对照组分别采用联邦化的 DPSGD 算法 (Fed-DPSGD)、基于动量更新方法的 DPSGD 算法 (Fed-DPSGD-M) 以及 DPSGD 算法 (Fed-DPSGD)。如实验结果图4-3所示，三种方法在联邦学习场景下的效果均要优于传统的 DPSGD 算法。Fed-KFGrad 在模型的收敛速率和最终的收敛精度上均有较大的提升。尽管 KF-Param 方法对模型精度的提升效果并不显著，但是根据实验结果可以观察到基于卡尔曼滤波器的超调过滤方法在模型训练的前中期在一定程度上加快了模型的收敛速率。二者方法的结合 Fed-KF2 算法相较于两个单独方法的模型效果均有了较大的提升，不仅大大加快了收敛速率，并且还有效地缓解了其中所存在的超调现象，提高模型最终的收敛速率，更好地实现了隐私和效用之间的权衡。针对 KF-Param 方法的效果并不显著的问题，本文分析这与联邦学习架构下数据异质性分布有关，即各个参与方每一步所计算的梯度互相之间存在一定偏差，并且各个参与方的数据分布也并非完全一致，进而导致各自累计的动量之间存在一定的偏差，这些偏差相互作用，可能导致整个系统中各个参与方均无法收敛至达到最优点，造成最终效果的衰减。为了进一步验证此方法的有效性，本文又进一步扩展了其他两个数据集 + 算法组合 (MLP-MINIST 和 CNN-CIAFAR10)，具体的实验结果如下：

表 4-1 Fed-KF 方法实验结果

| 算法 | 数据集 | KF-Grad 方法 | | KF-Param 方法 | | KF2 方法 | |
|-----|----------|------------|---------------|-------------|---------------|-----------|---------------|
| | | Baseline1 | KFGrad | Baseline2 | KFParam | Baseline3 | KF2 |
| MLP | MINIST | 0.9268 | 0.9231 | 0.9484 | 0.9496 | 0.9268 | 0.9465 |
| CNN | MINIST | 0.9495 | 0.961 | 0.9786 | 0.9788 | 0.9495 | 0.9746 |
| CNN | CIAFAR10 | 0.4721 | 0.4835 | 0.5264 | 0.5367 | 0.4721 | 0.5458 |

如表4-1所示，其中 Baseline1 和 Baseline3 均是采用 DPSGD 算法进行训练的结果，而 Baseline2 则是使用基于联邦动量更新方法的 DPSGD 算法进行训练的结果。根据实

验结果分析可知, KF-Grad 方法可以大幅增加模型的训练精度, 并且 KF-Param 算法也可以在一定程度上提升了最终的效果。将这两个方法进行结合后所生成的 Fed-KF2 算法在精度和收敛速度上均取得了较大的优势。除此之外, 本文发现使用 CIAFAR10 数据集时, 模型的效果整体都较差, 初步分析是由两个原因所造成的, 其一是因为该任务自身较为复杂, 其在集中式的场景在的效果就普遍较差, 难以达到 60%, 其二是因为差分隐私技术和去中心化的联邦学习架构给模型的精度带来了额外的损失。

综上所述, 将本文第三章提出的三个基于卡尔曼滤波器的优化方法迁移到集中式的联邦学习场景下时, 相较于传统的 DPSGD 和 DPSGD-M 算法, 这三个算法均取得了较优的效果, 这证明本文第三章所提出的方法在联邦学习场景下仍然适用, 这些方法通过提升各参与方的训练速率和收敛精度, 进而提升联邦系统的训练速率和收敛精度, 在中心化联邦学习场景在实现更好的隐私效用权衡。

4.3.3 去中心化联邦架构下的梯度过滤方法实验

本节主要测试在去中心化联邦学习架构下基于共识卡尔曼滤波器的梯度过滤方法方法的效果。该实验的对照组为 Decel 算法和 KFGrad 算法。Decel 算法指的是在去中心化的联邦学习场景下各参与方各自运行 DPSGD 的实验结果, 该算法的实验结果用于作为本实验的基线; 而 KFGrad 算法则指的是在去中心化的联邦学习场景下各参与方各自运行基于卡尔曼滤波器的梯度过滤算法, 相较于本节验证的 KCIF-Grad 方法, KFGrad 方法并未考虑如何加速达成系统共识。该实验的参与方的数量为 20, 且稀疏度 $\varphi = 0.3$, 全局 epoch 数量为 100, 所添加的噪声方差为 3, 裁剪系数为 1。最终的评价指标为各个参与方的准确率的均值。具体的实验结果表如下:

表 4-2 KCIF-Grad 方法实验结果

| 算法 | 数据集 | Decel 算法 | KFGrad 算法 | KCIFGrad 算法 |
|-----|----------|----------|-----------|---------------|
| MLP | MINIST | 0.8979 | 0.9079 | 0.9107 |
| CNN | MINIST | 0.9260 | 0.9328 | 0.9352 |
| CNN | CIAFAR10 | 0.4173 | 0.4257 | 0.4311 |

如表4-2所示, 根据上述实验结果分析可得, KFGrad 算法在三个不同的实验设置下均优于基线的 Decel 算法, 这证明了 KFGrad 算法在去中心化联邦学习的场景下仍然起效。KCIF-Grad 方法在三个不同的实验设置下的效果均是最优, 并且相较于 KFGrad 方法, 有了较为明显的提升, 这是由于 KCIF 滤波器中的共识功能给去中心化系统带来的额外增益。该实验证明了 KCIF-Grad 方法在去中心化联邦学习架构下的有效性。然而, 尽管 KF-Grad 方法的效果是次优, 但与集中式的情况相比, 去中心化联邦学习架构下, 它跟 DeceFL 方法之间最终收敛精度的差值被大大缩小。这也反映了在去中心化联邦学习架构下通过加速达成共识速率会给系统带来的额外增益。

4.3.4 去中心化联邦架构下的超调过滤方法实验

本节用来测试去中心化架构下基于共识卡尔曼滤波器的超调过滤方法的效果。该实验参与方数量为 20，且稀疏度 $\varphi = 0.3$ ，本实验采用的评价指标为选取各个参与方准确率的均值作为结果。基础参数设置为：高斯噪声的方差为 3，裁剪系数为 1，批大小为 50，动量系数 α 为 0.7，全局最大 epoch 数量为 100。具体实验结果如下：

表 4-3 KCIF-Param 方法实验结果

| 算法 | 数据集 | Decel 算法 | KFParam 算法 | KCIFParam 算法 |
|-----|----------|----------|------------|---------------|
| MLP | MINIST | 0.9104 | 0.9122 | 0.9134 |
| CNN | MINIST | 0.9411 | 0.9427 | 0.9438 |
| CNN | CIAFAR10 | 0.4173 | 0.4225 | 0.4282 |

本实验的实验组采用的是 KCIF-Param 算法，两个对照组分别采用基于动量优化方法的 Decel 算法以及联邦化 KFParam 算法，如表如4-3所示，经过对表4-3中实验结果分析可得，KFParam 算法在三个不同的算法-数据集的组合下的最终效果均优于本文的基线 Decel 算法，这说明本文第三章所提出的基于卡尔曼滤波器的超调过滤方法在去中心化的联邦学习架构下仍然有效，但是其相对于集中式机器学习场景下的提升在一定程度上被削弱。而 KCIF-Param 方法在三个不同的算法-数据集的组合设置下的效果仍然均是最优，其不仅优于基线 Decel 算法，并且全部优于联邦化的 KFParam 算法，这证明了 KCIF-Param 方法在去中心化联邦学习架构下的有效性，即通过加快去中心化的分散系统的达成共识的速率可以使得模型有额外的增益，减小了去中心化拓扑结构对模型最终精度的影响。除此之外，将该方法的实验结果与 KCIF-Grad 的结果相对比发现，当测试的模型数据集为 CNN-CIAFAR10 时，在使用 KCIF-Param 算法的效果 (准确率为 0.4282) 不如 KCIF-Grad(准确率为 0.4311) 的效果，该现象与集中式情况下的实验结果并不一致。这也对应了 4.2.3 节中提出的由于去中心化联邦学习架构的存在的梯度异质性会加重超调现象进而影响模型最终的精度的问题。

4.3.5 去中心化联邦架构下的联邦学习优化方法实验

本小节主要测试去中心化架构下基于共识卡尔曼滤波器的隐私保护联邦学习优化方法 (KCIF2 方法) 的效果，并在不同拓扑结构下对该方法的有效性进行了进一步验证。实验参与方数量为 20，且稀疏度 $\varphi = 0.3$ ，本实验采用的评价指标为选取各个参与方准确率的均值。其他的参数设置为：高斯噪声的方差为 3，裁剪系数为 1，批大小为 50，动量系数 α 为 0.7，全局最大 epoch 数量为 70。具体的实验结果表明如下：

由表4-4可知，在三组不同的算法-数据集组合中，KF2 算法均优于基线的 Decel 算法，该现象表明第三章提出的 KF2 算法在去中心化联邦学习框架下仍然有一定的效果，

表 4-4 KCIF2 方法实验结果

| 算法 | 数据集 | Decel 算法 | KF2 算法 | KCIF2 算法 |
|-----|----------|----------|--------|---------------|
| MLP | MINIST | 0.8887 | 0.9240 | 0.9250 |
| CNN | MINIST | 0.9144 | 0.9542 | 0.9547 |
| CNN | CIAFAR10 | 0.4173 | 0.4437 | 0.4501 |

但出现了效果衰减。KCIF2 算法在三组纵向对比的实验中，在每一组组合中的效果相较于其他两种方法均为最优，此现象证明了通过加快系统达成共识的速率能够给系统精度带来额外提升。KCIF2 算法保留了 KCIF-Param 算法的优点，即通过引入动量优化方法大大加快了模型的收敛速度，在一定程度上减少了模型收敛所需要的最大迭代轮数，并且 KCIF-Grad 通过过滤高斯噪声进而降低了超调噪声的累计，进而能够提升最终的收敛精度。并且二者的结合后要明显优于 KCIF-Grad 算法和 KCIF-Param 算法的效果，这也在某种程度上证明了两种方法结合的有效性。

基于上述结果，本节额外测试在不同拓扑结构下该方法对最终模型效果的影响。实验选取 KCIF2 算法，选用 MLP-MINIST 这组模型数据用于实验测试，拓扑结构中参与方数量为 20，且将拓扑图的稀疏度 φ 分别设置为 0.1, 0.3, 0.5, 0.7。训练过程中所添加的高斯噪声方差为 5，且动量优化方法的动量系数为 $\alpha = 0.7$ 。最终的评价指标选取各个参与方准确率的均值作为结果。为了对齐各组实验的结果，同时防止稀疏度高的时候，模型收敛过快导致过拟合，进而影响实验结果。本节均将全局迭代次数设置为 200，且将训练过程中最优的全局平均结果作为本实验的结果。实验结果如下：

表 4-5 KCIF2 方法实验结果

| 算法-数据集 | 拓扑图稀疏度 | Decel 算法 | KCIF2 算法 |
|--------------|--------|----------|---------------|
| MLP-MINIST | 0.1 | 0.8874 | 0.9252 |
| | 0.3 | 0.8887 | 0.9255 |
| | 0.5 | 0.8888 | 0.9271 |
| | 0.7 | 0.8889 | 0.9282 |
| CNN-MINIST | 0.1 | 0.9135 | 0.9532 |
| | 0.3 | 0.9144 | 0.9547 |
| | 0.5 | 0.9130 | 0.9554 |
| | 0.7 | 0.9140 | 0.9557 |
| CNN-CIAFAR10 | 0.1 | 0.3575 | 0.4467 |
| | 0.3 | 0.3605 | 0.4501 |
| | 0.5 | 0.3619 | 0.4653 |
| | 0.7 | 0.3695 | 0.4678 |

通过对表4-5中的结果分析可得，本文所提出的 KCIF2 算法在多个不同稀疏度的通信拓扑结构下均取得了更优的结果。在三个不同的算法-数据集组合中，CNN-CIAFAR10 这组实验的提升最高，KCIF2 算法给准确率所带来的提升将近 8%，该现象的出现的原

因可能是由于 CNN-CIFAR10 这组实验本身难以收敛，在集中式的场景下该组实验的精度就远低于其他两组，在去中心化的联邦学习场景下受到共识速率以及通信速率的影响，该组实验其更难收敛，然而 KCIF2 方法通过加速系统达成共识的速率进而有效缓解了该问题。这个结果也充分证明了 KCIF2 算法在去中心化的联邦学习架构下的有效性。通过对拓扑图稀疏度和 Decel 算法的最终精度之间的关系进行分析还可以发现，拓扑图本身的稀疏度对模型效果的影响整体上是正相关的，即通信拓扑图的稀疏度越高，则各个参与方之间的通信信道越多，信息交互速率越快，模型最终的收敛精度也会随之越高。但是该结论却具有一定的不确定性，因为在 CNN-MINIST 这组实验中，拓扑图稀疏度为 0.7 的 Decel 算法的效果反而没有拓扑图稀疏度为 0.3 的效果好，这与直觉相悖。造成这种现象的原因可能是去中心化联邦学习架构中的梯度异质性问题以及数据分布异质性问题所造成的。通过对拓扑图稀疏度和 KCIF2 算法的最终精度之间的关系进行分析还可以发现，KCIF2 算法相较于 Decel 算法所带来的效果提升与拓扑图稀疏度并未存在明显的关联，该方法在不同的拓扑图稀疏度下均取得了较为稳定的提升，且该提升与拓扑图稀疏度无关。这证明了去中心化联邦架构下基于共识卡尔曼滤波器的隐私保护机器学习方法具有一定的普适性，适用于多种不同的拓扑结构。

4.4 本章小结

本章在去中心化的联邦学习架构下提出了两个基于共识卡尔曼滤波器的隐私保护优化算法。首先将第三章所提出的两个卡尔曼滤波器算法扩展至共识卡尔曼滤波器算法，通过引入共识项，加快去中心化联邦学习架构中各参与方的模型训练过程中达到共识的速率，进而提升模型的收敛精度。除此之外，针对本章节中提出的两个效果增益不如集中式场景的原因进行了较为细致的分析，并在不同的拓扑结构下进行了实验。实验结果表明，本章节提出的 KCIF-Grad 方法、KCIF-Param 方法以及二者结合后的 KCIF2 方法在多种场景下均有不同程度的增益，证明了该方法具有一定的普适性。

5 结论和展望

5.1 研究结论

联邦学习^[4]作为一种分布式学习方法，它允许在不共享原始数据的情况下进行模型训练和更新，这种分布式学习方式避免了数据的中心化，有效减少了隐私泄露的风险。差分隐私^[10]通过向训练过程中的梯度添加噪声，进而给联邦学习过程提供更强的隐私保护。本文针对差分隐私所造成的隐私效用权衡问题以及去中心化联邦场景下的系统共识问题展开研究，利用卡尔曼滤波器和共识卡尔曼滤波器来解决训练过程中的梯度波动问题、超调波动问题以及系统共识问题，主要的研究内容和创新点总结如下：

1) 基于卡尔曼滤波器的隐私保护机器学习优化方法研究

针对差分隐私造成的隐私效用权衡问题，提出了基于卡尔曼滤波的梯度过滤方法，基于全局梯度近似线性变化的假设，利用卡尔曼滤波器能够在动态线性系统中过滤高斯噪声的性质，用来过滤所添加的差分隐私噪声，使得梯度信息变得更为准确平稳，大大加快模型的收敛速率和最终的收敛精度。经过大量的实验验证，证明本方法在 DPSGD 的场景下有较大的提升。除此之外，为了进一步加快模型的收敛速率，本研究额外引入了动量优化方法用来给模型训练过程加速，并针对其额外带来的超调问题，本研究进一步提出了基于卡尔曼滤波器的超调过滤方法，确保模型在收敛的后期能够快速平稳的收敛。上述两种方法分别针对梯度和针对模型参数进行过滤，将两个方法进行结合便形成了基于卡尔曼滤波器的隐私保护机器学习优化方法，在确保隐私保护程度不变的前提下，大大加快了模型的收敛速率，提升了模型的效用，进而更好地实现了隐私保护和效用的权衡。

2) 基于卡尔曼滤波器的隐私保护联邦学习优化方法研究

针对去中心化联邦学习中的共识问题，提出了基于共识卡尔曼滤波的去中心化隐私保护联邦学习优化方法，首先将基于卡尔曼滤波的隐私保护机器学习优化方法扩展至联邦学习架构下，然后进一步迁移至去中心化的联邦学习架构下，并针对去中心化联邦系统中的共识的问题，通过将基于卡尔曼滤波器的隐私保护机器学习优化方法中的卡尔曼滤波器扩展为共识卡尔曼滤波器，进而在过滤系统中梯度噪声和超调噪声的同时，加快系统达成共识的速率。最后，在多个不同的去中心化联邦拓扑结构下进行了大量实验，实验结果显示，扩展后的 KCIF-Grad 方法、KCIF-Param 方法以及 KCIF2 方法在多个数据集和拓扑结构上均取得了最优，实验结果充分证明了 KCIF2 方法可以有效解决去中心化联邦系统中的共识问题，进而给模型精度带来额外的提升。

5.2 未来展望

本文主要对基于差分隐私的联邦学习方法进行了深入的探索和研究，研究主要包含两个方面，其一是基于卡尔曼滤波对隐私保护机器学习算法进行了优化；其二是基于共识卡尔曼滤波器解决了去中心化联邦学习中的共识问题。本文的研究工作具有较高的现实意义和应用价值，但面对现实中复杂的场景与多样化的需求，该工作仍存在需要进一步完善和优化的方向，未来的研究工作可以从以下两个方面展开：

1) 梯度非线性变换问题研究

本文所提出的方法是基于将全局梯度的变换过程和超调噪声波动过程假设为一个动态线性过程所得。但在机器学习的场景中，梯度的变化曲线可能是极为复杂的。在这样的场景下，本文的方法会限制梯度的剧烈变化进而影响模型的训练过程。针对此问题，可以尝试利用扩展卡尔曼滤波器等适用于非线性系统的方法，但这些方法也会增大训练过程的计算复杂度。因此如何将卡尔曼滤波器扩展至非线性变换的动态系统中的同时不会额外增加复杂度可以作为本工作的第一个未来研究方向。

2) 去中心化联邦架构下数据非独立同态分布研究

本研究并未考虑各个参与方节点数据非独立同态分布的问题，若各个参与方的数据分布存在较大差异，则可能导致联邦学习的作用被削弱，甚至起到反作用，即参与方单独使用自身数据的最终效果要优于联邦训练的模型效果。针对该问题，研究如何将适用于传统横向联邦学习中的非独立同态分布解决方案扩展至去中心化的联邦学习架构下，进而实现个性化的训练过程可以作为本工作的第二个未来研究方向。

致 谢

在硕士论文即将完成之际，我怀着一颗感恩的心，向所有在我求学路上给予帮助和支持的人表达衷心的感谢。

首先，我要特别感谢我的导师任老师。任老师在我研究生阶段的学术道路上，给予了我悉心的指导和无私的帮助。他严谨的治学态度、深厚的学术造诣，使我深受启发。在我迷茫时，为我指明了方向；在我遇到困难时，也及时给予我鼓励和帮助。任老师的教诲和关怀，我将铭记在心，永远感激。

其次，我要感谢我的朋友们。在研究生期间，认识了很多很好的朋友，他们一直陪伴在我身边，给予我陪伴和鼓励。我们一起度过了许多难忘的时光，共同面对挑战，分享喜悦。大家的友情是我宝贵的财富，让我在求学路上不再孤单。

此外，我还要特别感谢在京东实习期间指导我的 mentor。她耐心地解答我在工作中的疑惑，帮助我快速适应职场环境。给了我悉心指导和宝贵建议，让我在实习期间收获颇丰，为未来的职业生涯奠定了坚实的基础。

最后，我要感谢我的家人。父母一直是我坚强的后盾，默默支持我、鼓励我。在我求学的过程中，始终给予我无私的爱和关怀，让我能够专心于学业。他们的付出和支持，是我能够顺利完成学业的重要保障。

总之，我的硕士毕业论文能够顺利完成，离不开导师的悉心指导、朋友们的陪伴鼓励、实习期间的 mentor 的耐心帮助以及家人的无私支持。在此，我再次向他们表示衷心的感谢！在未来的道路上，我将继续努力，不辜负他们的期望，为学术事业和人生道路贡献自己的力量。

参考文献

- [1] Li C, Chen Y, Shang Y. A review of industrial big data for decision making in intelligent manufacturing[J]. Engineering Science and Technology, an International Journal, 2022, 29: 101021.
- [2] Le Ny J, Pappas G J. Differentially private filtering[J]. IEEE Transactions on Automatic Control, 2014, 59(2): 341-354.
- [3] Le Ny J, Pappas G J. Differentially private Kalman filtering[J]. Differential Privacy for Dynamic Data, 2020, 68(10): 55-75.
- [4] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Artificial Intelligence and Statistics. 2017: 1273-1282.
- [5] Kairouz P, McMahan H B, Avent B, et al. Advances and open problems in federated learning[J]. Foundations and Trends® in Machine Learning, 2021, 14(1-2): 1-210.
- [6] Liu Y, Huang A, Luo Y, et al. Fedvision: An online visual object detection platform powered by federated learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 34: 08. 2020: 13172-13179.
- [7] Kumar R, Khan A A, Kumar J, et al. Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging[J]. IEEE Sensors Journal, 2021, 21(14): 16301-16314.
- [8] Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models [C]//2017 IEEE Symposium on Security and Privacy. 2017: 3-18.
- [9] Melis L, Song C, De Cristofaro E, et al. Exploiting unintended feature leakage in collaborative learning[C]//2019 IEEE Symposium on Security and Privacy. 2019: 691-706.
- [10] Yang M, Guo T, Zhu T, et al. Local differential privacy and its applications: A comprehensive survey [J]. Computer Standards & Interfaces, 2023, 89: 103827.
- [11] Wei K, Li J, Ding M, et al. Federated learning with differential privacy: Algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454-3469.
- [12] 刘艺璇, 陈红, 刘宇涵, 等. 联邦学习中的隐私保护技术[J]. 软件学报, 2021, 33(3): 1057-1092.
- [13] Zhu Y, Yu X, Chandraker M, et al. Private-knn: Practical differential privacy for computer vision[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11854-11862.
- [14] Khodarahmi M, Maihami V. A review on Kalman filter models[J]. Archives of Computational Methods in Engineering, 2023, 30(1): 727-747.
- [15] Ren X, Yu C M, Yu W, et al. Dperowd: privacy-preserving and communication-efficient decentralized statistical estimation for real-time crowdsourced data[J]. IEEE Internet of Things Journal, 2020, 8(4): 2775-2791.
- [16] 王恺祺, 洪睿琦, 毛云龙, 等. 基于区块链构建安全去中心化的联邦学习方案[J]. 中国科学: 信息科学, 2024, 54: 316-334.
- [17] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Artificial Intelligence and Statistics. 2017: 1273-1282.

-
- [18] Hegedűs I, Danner G, Jelasity M. Decentralized learning works: An empirical comparison of gossip learning and federated learning[J]. *Journal of Parallel and Distributed Computing*, 2021, 148: 109-124.
- [19] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy[C]//*Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 2016: 308-318.
- [20] Mironov I. Rényi differential privacy[C]//*2017 IEEE 30th Computer Security Foundations Symposium*. 2017: 263-275.
- [21] Choudhury O, Gkoulalas-Divanis A, Salonidis T, et al. Differential privacy-enabled federated learning for sensitive health data[J]. *arXiv preprint arXiv:1910.02578*, 2019: 1-6.
- [22] Phan H, Thai M T, Hu H, et al. Scalable differential privacy with certified robustness in adversarial learning[C]//*International Conference on Machine Learning*. 2020: 7683-7694.
- [23] Kairouz P, McMahan B, Song S, et al. Practical and private deep learning without sampling or shuffling[C]//*International Conference on Machine Learning*. 2021: 5213-5225.
- [24] Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning[C]//*Proceedings of the 2019 IEEE Symposium on Security and Privacy*. 2018: 1-15.
- [25] Gong M, Pan K, Xie Y, et al. Preserving differential privacy in deep neural networks with relevance-based adaptive noise imposition[J]. *Neural Networks*, 2020, 125: 131-141.
- [26] Du J, Li S, Chen X, et al. Dynamic differential-privacy preserving sgd[J]. *arXiv preprint arXiv:2111.00173*, 2021: 1-16.
- [27] Xu Z, Shi S, Liu A X, et al. An adaptive and fast convergent approach to differentially private deep learning[C]//*IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. 2020: 1867-1876.
- [28] Xu D, Du W, Wu X. Removing disparate impact on model accuracy in differentially private stochastic gradient descent[C]//*Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021: 1924-1932.
- [29] Wu X, Fredrikson M, Jha S, et al. A methodology for formalizing model-inversion attacks[C]//*2016 IEEE 29th Computer Security Foundations Symposium*. 2016: 355-370.
- [30] Yang Z, Zhang J, Chang E C, et al. Neural network inversion in adversarial setting via background knowledge alignment[C]//*Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019: 225-240.
- [31] Gong M, Feng J, Xie Y. Privacy-enhanced multi-party deep learning[J]. *Neural Networks*, 2020, 121: 484-496.
- [32] Kim M, Günlü O, Schaefer R F. Federated learning with local differential privacy: Trade-offs between privacy, utility, and communication[C]//*ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2021: 2650-2654.
- [33] Liu R, Cao Y, Yoshikawa M, et al. FedSel: Federated sgd under local differential privacy with top-k dimension selection[C]//*Database Systems for Advanced Applications: 25th International Conference*. 2020: 485-501.
- [34] Shi Y, Liu Y, Wei K, et al. Make landscape flatter in differentially private federated learning[C]

- //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 24552-24562.
- [35] Cheng A, Wang P, Zhang X S, et al. Differentially private federated learning with local regularization and sparsification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10122-10131.
- [36] He L, Bian A, Jaggi M. Cola: Decentralized linear learning[J]. Advances in Neural Information Processing Systems, 2018, 31: 4541-4551.
- [37] Billah M, Mehedi S T, Anwar A, et al. A systematic literature review on blockchain enabled federated learning framework for internet of vehicles[J]. arXiv preprint arXiv:2203.05192, 2022: 1-36.
- [38] Gupta R, Alam T. Survey on federated-learning approaches in distributed environment[J]. Wireless Personal Communications, 2022, 125(2): 1631-1652.
- [39] Saraswat D, Verma A, Bhattacharya P, et al. Blockchain-based federated learning in UAVs beyond 5G networks: A solution taxonomy and future directions[J]. IEEE Access, 2022, 10: 33154-33182.
- [40] Chen H, Wang H, Long Q, et al. Advancements in federated learning: Models, methods, and privacy [J]. arXiv preprint arXiv:2302.11466, 2023: 1-35.
- [41] Witt L, Heyer M, Toyoda K, et al. Decentral and incentivized federated learning frameworks: A systematic literature review[J]. IEEE Internet of Things Journal, 2022, 10(4): 3642-3663.
- [42] Cyffers E, Bellet A. Privacy amplification by decentralization[C]//International Conference on Artificial Intelligence and Statistics. 2022: 5334-5353.
- [43] Tran A T, Luong T D, Karnjana J, et al. An efficient approach for privacy preserving decentralized deep learning models based on secure multi-party computation[J]. Neurocomputing, 2021, 422: 245-262.
- [44] Qu Y, Uddin M P, Gan C, et al. Blockchain-enabled federated learning: A survey[J]. ACM Computing Surveys, 2022, 55(4): 1-35.
- [45] Naseri M, Hayes J, De Cristofaro E. Local and central differential privacy for robustness and privacy in federated learning[J]. arXiv preprint arXiv:2009.03561, 2020: 1-20.
- [46] 赵禹齐, 杨敏. 差分隐私研究进展综述[J]. 计算机科学, 2023, 50: 265-276.
- [47] 孙一帆, 张锐, 陶杨, 等. 本地化差分隐私综述[J]. 数据与计算发展前沿, 2023, 5: 74-97.
- [48] Ren H, Deng J, Xie X. Grnn: generative regression neural network—a data leakage attack for federated learning[J]. ACM Transactions on Intelligent Systems and Technology, 2022, 13(4): 1-24.
- [49] Dwork C, Kenthapadi K, McSherry F, et al. Our data, ourselves: Privacy via distributed noise generation[C]//Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques. 2006: 486-503.
- [50] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography: Third Theory of Cryptography Conference. 2006: 265-284.
- [51] Mironov I. Rényi differential privacy[C]//2017 IEEE 30th Computer Security Foundations Symposium. 2017: 263-275.
- [52] El Ouadrhiri A, Abdelhadi A. Differential privacy for deep and federated learning: A survey[J]. IEEE Access, 2022, 10: 22359-22380.

-
- [53] Shokri R, Shmatikov V. Privacy-preserving deep learning[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015: 1310-1321.
- [54] 付学瀚, 燕贺云, 朱立东, 等. 基于卡尔曼和扩展卡尔曼滤波的耦合载波跟踪方法[J]. 移动通信, 2024, 48: 118-124.
- [55] Olfati-Saber R. Kalman-consensus filter: Optimality, stability, and performance[C]//Proceedings of the 48th IEEE Conference on Decision and Control held jointly with 2009 28th Chinese Control Conference. 2009: 7036-7042.
- [56] Menghani G. Efficient deep learning: A survey on making deep learning models smaller, faster, and better[J]. ACM Computing Surveys, 2023, 55(12): 1-37.
- [57] An W, Wang H, Sun Q, et al. A PID controller approach for stochastic optimization of deep networks [C]//Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition. 2018: 8522-8531.
- [58] Ziegler J G, Nichols N B. Optimum settings for automatic controllers[J]. Transactions of The American Society of Mechanical Engineers, 1942, 64(8): 759-765.
- [59] Ogata K. Discrete-time control systems[M]. Prentice-Hall, Inc., 1995.
- [60] Yuan Y, Liu J, Jin D, et al. DeceFL: A principled decentralized federated learning framework[J]. arXiv preprint arXiv:2107.07171, 2021: 1-58.
- [61] Liu W, Chen L, Chen Y, et al. Accelerating federated learning via momentum gradient descent[J]. IEEE Transactions on Parallel and Distributed Systems, 2020, 31(8): 1754-1766.
- [62] Karimireddy S P, Kale S, Mohri M, et al. Scaffold: Stochastic controlled averaging for on-device federated learning[J]. arXiv preprint arXiv:1910.06378, 2019: 1-12.
- [63] Li X, Huang K, Yang W, et al. On the convergence of fedavg on non-iid data[J]. arXiv preprint arXiv:1907.02189, 2019: 1-26.
- [64] Kim J, Lee K. Unscented Kalman filter-aided long short-term memory approach for wind nowcasting [J]. Aerospace, 2021, 8(9): 236.
- [65] Wang T, Liu Y, Zheng X, et al. Edge-based communication optimization for distributed federated learning[J]. IEEE Transactions on Network Science and Engineering, 2021, 9(4): 2015-2024.
- [66] Wu J, Drew S, Dong F, et al. Topology-aware federated learning in edge computing: A comprehensive survey[J]. arXiv preprint arXiv:2302.02573, 2023: 1-36.

攻读学位期间取得的研究成果

答辩委员会会议决议

论文围绕联邦学习的隐私保护优化方法进行了研究，提出了基于卡尔曼滤波的联邦学习优化方法，研究成果具有一定的理论意义和应用前景。

论文主要工作包括：

- 1) 针对差分隐私造成的梯度波动问题，提出了一种基于卡尔曼滤波的梯度过滤方法，并在多个数据集下进行了实验验证；
- 2) 针对动量优化方法造成的超调问题，提出了一种基于卡尔曼滤波的超调过滤方法，并在多个数据集下进行了实验验证；
- 3) 针对去中心化联邦架构下的共识问题，提出了一种基于共识卡尔曼滤波器的联邦学习优化方法，并在多个数据集和不同的拓扑结构下进行了实验验证。

论文写作认真，结构合理，论述清楚。论文工作表明作者具有较扎实的基础理论和系统的专业知识，以及独立从事科研工作能力。

论文答辩过程中，讲述清楚，回答问题正确。

答辩委员会根据学位申请人提交的材料、评阅人的意见和答辩情况，并经投票表决，一致同意授予王炜飞同学工学硕士学位。

常规评阅人名单

本学位论文共接受 2 位专家评阅，其中常规评阅人 0 名。

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1. 交回学校授予的学位证书；
2. 学校可在相关媒体上对作者本人的行为进行通报；
3. 本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。
4. 本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）：

任雪峰

日期：

年

月

日

学位论文独创性声明（2）

本人声明：研究生 所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1. 学校可在相关媒体上对本人的失察行为进行通报；
2. 本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。
3. 本人接受学校按照有关规定做出的任何处理。

指导教师（签名）：

任雪峰

日期：

年

月

日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）：

日期：

年

月

日

指导教师（签名）：

任雪峰

日期：

年

月

日

（本声明的版权归西安交通大学所有，未经许可，任何单位及任何个人不得擅自使用）