异构网络中的联合优化

天丽1 Anit Kumar Sahu2 Manzil Zaheer3 Maziar Sanjabi 4 Ameet Talwalkar 1 5 Virginia Smith 1

摘要

联邦学习是一种分布式学习范式，与传统分布式优化相比，存在两个关键挑战：(1)网络中每个设备上的系统特性存在显著差异（系统异质性），以及(2)网络中数据的非同分布性（统计异质性）。在这项工作中，我们引入了一个框架，FedProx，以应对联邦网络中的异质性问题。FedProx可以视为FedAvg方法的泛化和再参数化，FedAvg是当前最先进的联邦学习方法。尽管这种再参数化仅对方法本身进行了微小修改，但这些修改在理论和实践上都具有重要意义。理论上，我们为我们的框架提供了收敛保证，当数据来自非同分布（统计异质性）时，并且通过允许每个参与设备执行可变的工作量来遵守设备级别的系统约束（系统异质性）。实际上，我们证明了FedProx在一系列现实的联邦数据集上比FedAvg具有更强的收敛能力。特别是在高度异构环境中，与FedAvg相比，FedProx表现出显著更稳定和准确的收敛行为，平均绝对测试精度提高了22%

1.导言

联邦学习作为一种在远程设备网络中分发机器学习模型训练的有吸引力的范式应运而生。尽管在机器学习的分布式优化领域已有大量研究，但联邦学习与传统分布式优化相比，存在两个关键挑战：系统高度复杂和统计异质性。(McMah an等人，2017;列支敦士登以及其他人2019).

为了处理异质性和解决高通信成本，允许本地更新和低参与度的优化方法是联邦学习的流行方法(McMahan等人， 2017;Smith 等 人， 2017)。特别是，FedAvg (McMahan等人，2017）是一种迭代方法，在联邦设置中已经成为事实上的优化方法。在每次迭代中，FedAvg首先局部执行E个随机梯度的E个周期

1卡内基梅隆大学2博世人工智能中心3谷歌研究4Facebook AI5确定的人工智能。通信地址：Tian Li< tianli@cmu.edu>。

第三届MLS ys会议，美国德克萨斯州奥斯汀市，
2020.

隐私是联邦设置下的第三个关键挑战。虽然这不是本文的重点，但标准的隐私保护方法，如差分隐私和安全多方通信，可以自然而然地与本文提出的方法结合使用——特别是因为我们的框架仅提出了对先前工作的轻量级算法修改。

设备上的dient descent（SGD）——其中E是一个小常数，K是网络中总设备的一个小部分。然后，这些设备将它们的模型更新发送到一个中央服务器，在那里进行平均。

虽然FedAvg在均质环境中展示了实证成功，但它并未完全解决异质性带来的潜在挑战。在系统异质性的背景下，FedAvg不允许参与设备根据其底层系统约束执行不同数量的本地工作；相反，通常的做法是简单地丢弃未能在指定时间窗口内计算E个周期的设备(Bonawitz等人，2019)从统计学的角度来看，在数据在不同设备上不等同分布的情况下，FedAvg在经验上出现了分歧(例如，McMahon以及其他人2017，第3节)。不幸的是，在这种现实情况下，很难从理论上分析FedAvg，因此缺乏收敛性保证来描述其行为(见第2节更多详情请参见）。

在这项工作中，我们提出了FedProx，一种联邦优化算法，旨在从理论和实证两方面应对异质性挑战。我们在开发FedProx时的一个重要见解是，在联邦学习中，系统之间存在相互作用以及统计异质性。事实上，无论是丢弃落后的节点（如FedAvg），还是简单地将部分信息纳入落后节点（如FedProx中将近端项设为0），都会隐式地增加统计异质性，并可能产生不利影响。

收敛行为。为了缓解这一问题，我们建议在目标函数中加入一个近端项，这有助于提高方法的稳定性。该项为服务器提供了一种合理的方法来考虑部分信息相关的异质性。理论上，这些修改使我们能够为我们的方法提供收敛保证，并分析异质性的影响。实证研究表明，这些修改提高了异构网络中联邦学习的稳定性和整体准确性，在高度异构环境中平均绝对测试准确率提高了22%。

本文其余部分组织如下。第2节，我们提供了联邦学习的背景和相关工作的概述。然后，我们在第3节中介绍了我们提出的框架，FedProx并在第4节中推导出框架的收敛保证，该框架同时考虑统计异质性和系统异质性。最后，在第5节中，　我们对FedProx在一系列合成数据集和真实世界的数据集上进行了详尽的经验评估。我们的经验结果有助于说明和验证我们的理论分析，并证明了FedProx在异构网络中相对于FedAvg的实际改进。

2背景及相关工作

大规模机器学习，特别是在数据中心设置中，已经激发了过去十年中许多分布式优化方法的发展(见，例如，Boyd等人。2010;Dekel等人，2012;Dean等人，2012;张等人，2013;李等人，2014a;Shamir等人，2 014;Reddi等人，2016;张等人，2015;Richt rik & Tak2016;Smith等人，2018)然而，随着手机、传感器和可穿戴设备等计算基板在性能和普及率上的增长，学习分布式设备网络中的本地统计模型变得越来越有吸引力，这与将数据转移到数据中心的做法形成鲜明对比。这一问题被称为联邦学习，需要应对隐私、异构数据和设备以及大规模分布式网络带来的新挑战(Li等人，2019).

最近提出了针对联邦设置中特定挑战的优化方法。这些方法在传统分布式方法如ADMM (Bo yd以及其他人2010年）或小批量方法(Dekel等人，2012)允许在不精确的局部更新中平衡通信和计算，以及在任何通信周期内让一小部分设备处于活动状态(McMahan等人，2017;Smith等人，2017)例如，Smith等人。（2017）提出了一种通信效率高的方法原始-对偶优化方法，通过多任务学习框架为每个设备学习独立但相关的模型。尽管所提出的方法具有理论保证和实际效率，但这种方法并不可推广到非凸问题，例如深度学习，在这种情况下，强对偶性不再得到保证。在非凸环境中，基于原始问题中局部随机梯度下降（SGD）更新平均的启发式方法——联邦平均（FedAvg），已被实证证明效果良好(McMahan等人，2017).

不幸的是，FedAvg的分析相当具有挑战性，这不仅因为其本地更新方案，还因为每轮只有少数设备处于活动状态，以及数据在网络中经常以异构形式分布的问题。特别是，由于每个设备生成自己的本地数据，统计异质性普遍存在，数据在设备之间分布不均。一些研究已经在更简单的非联邦设置下尝试分析FedAvg。例如，parallelSGD及其相关变体（张以及其他人2015;Shamir等人，2014;Reddi等人，2016;周&Cong2018;诗行2019;王和乔希，2018;木材价值以及其他人2018;林等人，2020)，这些方法使得局部更新类似于FedAvg，在IID框架下进行了研究。然而，研究结果依赖于每个局部求解器都是同一随机过程的副本这一前提（基于IID假设）。这种推理方式不适用于异构环境。

尽管最近的一些研究(余等人，2018;王等人，2019;郝等，2019;Jiang& Agrawal，2018年）已退出-在统计上异质的环境中，他们要求收敛保证，并假设所有设备都参与每一轮通信，这在现实的联邦网络中通常是不可行的(McMahan等人，2017)此外，它们依赖于特定的求解器在每个设备上使用（无论是SGD还是GD），而与本文提出的求解器无关的框架相比，它们增加了凸性的额外假设(Wang等人，（2019）或统一有界梯度(Yu等人，　2018)用于他们的分析。也有启发式方法旨在通过共享本地设备数据或服务器端代理数据来解决统计异质性问题(Jeong等人，2018;赵等人，2018;黄以及其他人2018)然而，这些方法可能不现实：除了给网络带宽带来负担外，将本地数据发送到服务器(Jeong等人，2018年）违反了联邦学习的关键隐私假设，以及向所有设备发送全球共享的代理数据(Zhao等人，2018;黄等人，2018)需要努力仔细生成或收集此类辅助数据。

除了统计异质性外，系统异质性也是联邦网络中的一个关键问题。由于硬件(CPU、内存）、网络连接（3G、4G、5G、wifi）和电源（电池电量）的差异，每个设备在联邦网络中的存储、计算和通信能力可能不同。这些系统级特性极大地加剧了诸如缓解延迟和容错等挑战。一种策略是

在实践中，就是忽略那些受限制的设备无法完成一定量的训练(Bonawitz等人，2019)然而(如我们在第5节中所证明的)，这可能会对收敛产生负面影响，因为它限制了参与训练的有效设备数量，并且如果丢失的设备具有特定的数据特性，可能会导致设备采样过程中的偏差。

在这项工作中，受FedAvg启发，我们探索了一个更广泛的框架FedProx，该框架能够在处理异构联邦环境的同时，保持相似的隐私和计算优势。我们通过分析局部函数之间的统计不相似性来研究该框架的收敛行为，同时考虑实际系统约束。我们的不相似性特征受到随机化Kaczmarz方法解决线性方程组的启发(Kaczmarz，1993;Strohmer & Vershynin，（2009），一个类似的这种假设已被用于分析其他环境下的SGD变体(见Schmidt& Roux，2013;Vaswani等人，2019;尹等人，2018)我们提出的框架为异构联合网络中的优化提供了更好的鲁棒性和稳定性。

最后，在相关工作方面，我们注意到所提出工作的两个方面——FedProx中的近端项和分析中使用的有界不相似性假设——之前已在优化文献中有所研究，尽管这些研究往往具有非常不同的动机和非联邦设置。为了完整性，我们在附录B中提供了进一步的讨论。在这样的背景下工作。

3联合优化方法

在本节中，我们介绍了最近的联邦学习方法背后的关键成分，包括FedAvg，然后概述了我们提出的框架，FedProx。

联邦学习方法（例如，McMahan等人，20 17;Smith等人，2017）旨在处理多个设备收集数据和一个协调网络中全局学习目标的中央服务器。特别是，其目的是最小化：

$$\min_w f(w) \quad \sum_{k=1}^{N} p_k F_k(w) = \mathbb{E}_k[F_k(w)], \quad (1)$$

其中N是设备数量，pk≥0，andΣk p k= 1。通常情况下，局部目标衡量的是在可能不同的数据分布D k上的局部风险，即F k (w)：= E x k~D k[f k（w；x k）]，有nk个样本可用。
在每个设备k上。因此，我们可以设置p k= n n k，其中n=Σk n k是数据点的总数。在本研究中，我们考虑F k (w)可能是非凸的。

为了减少通信，一种常见的技术是，在每个设备上，采用局部目标函数进行优化

基于设备数据的功能被用作全局目标函数的代理。在每次外部迭代中，选择部分设备，并使用局部求解器优化这些选定设备上的局部目标函数。然后，设备将其局部模型更新发送到中央服务器，中央服务器汇总这些信息并相应地更新全局模型。在这种情况下实现灵活性能的关键在于每个局部目标函数都可以不精确求解。这使得可以根据执行的局部迭代次数来调整局部计算与通信的比例（更多的局部迭代对应于更精确的局部解）。我们将在下文中正式介绍这一概念，因为它将在本文中多次使用。

定义1（√-不精确解）.对于函数h（w；w₀ ）=-w₀ Ⅱ2和√∈[0,1]，我们称w*是min wh（w；w₀ ）ifⅡ▽h（w*；w₀ ）Ⅱ≤√Ⅱ▽h（w₀ ；w₀ ）Ⅱ的一个√-不精确解，其中$F(w) + \frac{\mu}{2}\|w$
▿h (w ; w 0) =▿F (w) +μ(w—w 0)。注意，较小的√对应更高的精度。

我们在分析中使用了√-不精确性（第4节)定时确保每轮本地求解器的局部计算量。如前所述，由于系统条件的变化，不同的设备在解决局部子问题时可能会取得不同的进展，因此允许√随设备和迭代次数变化非常重要。这是我们接下来讨论的框架的一个动机。为了便于记号，我们首先假设一个均匀的√来推导主要的收敛结果，如本文第4节所述。 ），
然后在推论9中给出结果，其中包含变量√'。

3.1联盟平均（FedAvg）

在Federated Averaging（FedAvg）(McMahan等人，2017)，局部目标函数的全局代理函数为F k（·），局部求解器采用随机梯度下降（SGD），每个设备使用相同的步长和局部训练轮数。每一轮中，从所有设备中选择一部分K冬N进行本地SGD训练E轮，然后将结果模型更新进行平均。FedAvg的详细过程总结在算法1中。

McMahan等人。（2017）从实证角度表明这一点至关重要 调整FedAvg的优化超参数。特别是，FedAvg中的局部训练轮数对收敛性起着重要作用。一方面，执行更多的局部训练轮次可以增加局部计算量并减少通信，这在通信受限的网络中可以显著提高整体收敛速度。另一方面，当局部目标函数F k不同（异构）时，更多的局部训练轮次可能导致每个设备趋向于其局部最优解。

算法1联盟平均法（FedAvg）

输入：K、T、η、E、w 0、N、p k、k=1、…、N，t=0、…、T—1
服务器随机选择K台设备的子集S t（每台设备k以概率p k被选中）
服务器向所有选定设备发送w t
每个设备k∈S t在E个SGD周期中更新w t
在F k上用步骤sizeη求得w t k+1
每个设备k∈S t将w t k+1发送回服务器
服务器将w聚合为wt+1=k∈St w t k+1$\frac{1}{K}\sum$
结束

目标与全局目标——可能损害收敛性，甚至导致方法发散。此外，在具有异构系统资源的联邦网络中，设置较高的本地训练轮数可能会增加设备无法在给定通信轮内完成训练的风险，因此必须退出该过程(Bonawitz等人，2019).

在实际应用中，因此重要的是找到一种方法，将局部周期设置得较高（以减少通信），同时允许稳健的收敛。更根本的是，我们注意到局部周期数量的最佳设置可能会随着每次迭代和每个设备的变化而变化——这取决于局部数据和可用系统资源。事实上，一种比强制固定局部周期数量更为自然的方法是，让周期根据网络特性变化，并通过考虑这种异质性来仔细合并解决方案。我们在下面介绍的FedProx中形式化了这一策略。

3.2提出的框架：FedProx

我们提出的框架，FedProx（算法2），类似于FedAvg，每轮选择一部分设备进行本地更新，然后将这些更新平均化以形成全局更新。然而，FedProx进行了以下简单但关键的修改，这不仅带来了显著的经验改进，还使我们能够为该方法提供收敛性保证。

容忍部分工作。如前所述，联邦网络中的不同设备在计算硬件、网络连接和电池电量方面通常具有不同的资源限制。因此，强制每个设备执行相同数量的工作（即运行相同数量的本地周期E）是不现实的，就像FedAvg那样。在FedPr ox中，我们通过允许根据可用系统资源在各设备上执行不同数量的工作来推广FedAvg，然后汇总来自落后的设备的部分解决方案（而不是丢弃这些设备）。换句话说，在整个训练过程中，FedPr ox并没有假设所有设备都具有相同的√，而是隐式地

可适应不同设备的可变√和不同的设备不同的迭代。我们正式定义√t k-不精确性
设备k在t次迭代之后，这是从定义1的自然延伸。

定义2（√tk不精确解）.对于函数hk（w；w t）= Fk—wt Ⅱ 2和√∈[0,1]，我们称w*是最小w hk（w；w t）ifⅡ▽hk（w*；w t）Ⅱ≤√t kⅡ▽hk（w t；w t）Ⅱ的一个√tk不精确解，其中$(w)+\frac{\mu}{2}\|w$
▿h k（w；w t）=▿F k(w)+μ（w—w t）。注意，a
较小的√t k对应较高的精度。

类似于定义1，√tk测量了局部的多少
计算用于在第t轮中解决设备k上的局部子问题。变量数量
局部迭代可以看作是√t k的代理。利用
更灵活的√t k-不精确性，我们可以很容易地扩展
定义1下的收敛结果 (定理4)考虑与系统异质性相关的问题，如滞后者（见推论9）.

近端术语。如第3.1节所述在容忍不同设备间工作量不均匀的情况下，也有助于减轻系统异质性带来的负面影响。然而，过多的本地更新仍可能（潜在地）导致方法因底层异质数据而发散。我们建议在本地子问题中加入一个近似项，以有效限制可变本地更新的影响。具体来说，不是仅仅最小化本地函数Fk（·），设备k使用其本地求解器来近似最小化以下目标函数h k：

$$\text{m win hk (w；w t）=Fk(w)+ⅡwⅡ2。(2)}\frac{\mu}{2}$$

近端项在两个方面是有益的： (1)它通过限制局部更新更接近初始（全局）模型来解决统计异质性问题，而无需手动设置局部epoch的数量。

(2)它允许安全地将由系统异质性产生的不同数量的本地工作整合在一起。我们在算法2中总结了FedProx的步骤。

算法2 FedProx（建议框架）

输入：K，T，μ，√，w 0，N，p k，k=1，…，N，t=0，…，T—1
服务器随机选择K台设备的子集S t（每台设备k被选中的概率为p k）
服务器向所有选定设备发送w t
每个选定的设备k∈S tfi nds a w t k+1，它是w t k+1≈的a√t k-不精确最小值。
arg min w h k（w；w t）= F k tⅡ 2$(w)+\frac{\mu}{2}\|w-w$
每个设备k∈S t将w t k+1发送回服务器，服务器
将w的值作为wt+1=k∈St w t k+1聚合$\frac{1}{K}\sum$
结束

我们注意到，像上面这样的近端项是优化文献中常用的非流行工具；为了完整性，我们在附录B中提供了更详细的讨论。我们建议、探索和分析这样一个术语，目的是解决联邦网络中的异质性问题。我们的分析(第4节)在分布式环境中解决此类目标的独特之处在于：(1)非IID分割的数据，(2)使用任何本地求解器，(3)在设备间进行变量不精确更新，以及(4)每轮有部分设备处于活动状态。这些假设对于在实际联邦场景中描述此类框架至关重要。

在我们的实验中(第5节)，我们证明了在系统异质性存在的情况下，容忍部分工作的方法是有益的，并且我们在FedProx中改进的局部子问题比普通的FedAvg在异构数据集上具有更稳健和稳定的收敛性。在第4节我们还发现，使用近似项使得FedProx更易于进行理论分析（即局部目标函数可能表现得更好）。特别是，ifμ被相应地选择，使得h km的海森矩阵可能是半正定的。因此，当Fk是非凸时，h k将是凸的；而当Fk是凸时，它将是becomesμ强凸的。

最后，我们注意到，由于FedP仅对FedAvg进行了轻量级的修改，这使我们能够对广泛使用的FedAvg方法的行为进行推理，并使FedProx能够轻松集成到现有的软件包/系统中，如TensorFlow Federated和LEAF (TFF；Caldas等人，201 8)尤其需要注意的是，FedAvg是FedProx的一个特例，具有以下特点：(1)μ=为0，(2)局部求解器特别选择为SGD，以及(3)设备和更新轮次之间存在常数√（对应于局部迭代次数，即系统异质性不存在）。实际上，FedProx在这方面更为通用，因为它允许部分工作在设备之间进行，并且每个设备可以使用任何局部（可能是非迭代的）求解器。

4 F ED P RO X：收敛性分析

FedAvg和FedProx本质上是随机算法：每轮中，只有部分设备被采样以执行更新，且每个设备上的更新可能不精确。众所周知，为了使随机方法收敛到一个稳定点，需要采用递减的步长。相比之下，非随机方法如梯度下降可以通过使用恒定的步长来找到稳定点。为了分析具有恒定步长的方法（通常在实际应用中实现）的收敛行为，我们需要量化这些方法之间的差异程度。

局部目标函数。这可以通过假设数据为IID，即同质且跨设备来实现。然而，在现实的联邦网络中，这一假设并不实用。因此，我们首先提出一个度量方法，专门用于衡量局部函数之间的差异（第4.1节），然后在允许变量√变化的情况下分析FedProx（第4.2节）.

4.1地方差异

这里我们介绍了一种在联邦网络中设备间不相似性的度量，这足以证明收敛性。也可以通过更简单和更严格的梯度方差假设来满足（推论10），我们在第5节的实验中进行了探索。有趣的是，类似的假设(例如，施密特&Roux，2013;Vaswani等人，2019;尹等人，2018)已经在其他地方被探索过，但目的不同；我们在附录B中对这些作品进行了讨论。

定义3（B局部不相似）。如果E k[Ⅱ∇F k(w)Ⅱ2]≤Ⅱ∇f (w)Ⅱ 2 B 2，则局部函数F k在w处是B局部不相似的。我们进一步定义B (w)= $\sqrt{\frac{\mathbb{E}_k[\|\nabla F_k(w)\|^2]}{\|\nabla f(w)\|^2}}$

2小时Ⅱ∇ f (w)Ⅱ 0 .≠

这里E k[·]表示质量为pk = n k p k=1的设备的期望值（如公式1所示）$/n$ and $\sum_{k=1}^{N}$. Defini- tion 3可以视为具有有界不相似性的IID假设的推广，同时允许统计异质性。作为合理性检查，在所有局部函数相同的情况下，我们有B (w) =1对于所有w。然而，在联邦设置中，数据分布通常具有异质性，即使假设样本是IID的，由于采样差异也会导致B>1。我们还考虑F k（·）与经验风险目标相关的情况。如果所有设备上的样本都是同质的，即它们以IID的方式被采样，则随着mink nk→∞，可以得出结论，对于每个w，B (w)→1，因为所有局部函数在大样本极限下收敛到相同的期望风险函数。因此，B (w)≥1，且B (w)的值越大，局部函数之间的不相似性就越大。

使用定义3，我们现在陈述我们的形式不相似性假设，我们在我们的收敛分析中使用它。这仅仅要求定义在定义3中的不相似性是有限的。如后文所述，我们的收敛速度是网络中统计异质性/设备不相似性的函数。

假设1（有界不相似性）。对于某些E > 0，存在一个BE，使得对于所有点w∈S={w|Ⅱ∇f (w)Ⅱ2 > E}，B(w)≤B E。

2作为例外，当E k[Ⅱ▽F k(w)Ⅱ2]=Ⅱ▽f (w)Ⅱ 2时，我们定义B (w)=1，即w是一个所有局部函数s Fk都一致的稳定解。

对于大多数实际的机器学习问题，没有必要求解高度精确的稳定解，即E通常不是非常小。事实上，众所周知，解决这个问题超过某个阈值甚至可能因过拟合而损害泛化性能(Yao以及其他人2007)尽管在实际的联邦学习问题中，样本并非IID，但它们仍然来自并非完全无关的分布（如果情况如此，例如，在设备之间拟合单一全局模型将是不合适的）。因此，可以合理假设在整个训练过程中，本地函数之间的差异保持有限。我们还在第5.3.3节中通过实证方法测量了真实和合成数据集上的差异度量。并表明该度量能够捕捉到真实世界的统计异质性，并且与实际性能相关（不相似度越小，收敛性越好）。

4.2 FedProx分析

使用有界不相似性假设（假设1)，我们现在分析执行一个步骤的FedProx时目标函数预期减少的数量。我们的收敛率（定理6)可以直接从每轮更新预期减少的结果中得出。我们
假设对于任意k，t，都有相同的√t k
随后的分析。

定理4（非凸FedProx收敛：B局部不相似性）.设备假设1*假设函数Fk非凸，L-利普希茨光滑，并且存在L-> 0，使得∇2 F k≥−L-1，其中：=μ−L-> 0。假设wt不是稳定解，局部函数Fk是B-不相等的，即B（wt）≤B。Ifμ，K，在算法2中μ̄选择这样的*

$$P=---\left(\frac{1\gamma BB(1+\gamma)\sqrt{2}}{\mu\,\mu}\frac{}{\bar{\mu}\sqrt{K}}-\frac{LB(1+\gamma)}{\bar{\mu}\mu}\right.$$

$$--\ 2\sqrt{\ }\ 2\ K+2\ ))>\ 0\ ,\ \frac{L(1+\overline{\gamma)^2}B^2}{2\bar{\mu}^2}$$

$$\frac{LB^2(1+\gamma)^2}{\bar{\mu}^2K}\Big($$

然后在算法2的迭代t中，我们预计全球目标将出现以下下降：

$$E\ S\ t[f\ (w\ t+1\ )]\leq f\ (w\ t\ )-P\amalg\nabla\ f\ (w\ t\ )\amalg\ 2\ ,$$

其中，St是第t次迭代中选择的K个设备的集合。

我们建议读者参见附录A.1详细证明见下文。关键步骤包括应用我们对√-不精确性的概念（定义1)对每个子问题进行处理，并使用有界不相似性假设，同时允许每轮中只有K个设备处于激活状态。特别是最后一步引入了E St，即关于第t轮设备选择S t的期望值。我们注意到，在我们的理论中，

要求>0，这是FedProx收敛的充分但非必要条件。因此，一些otherμ（不一定满足>0）也可能使收敛成为可能，我们将在实验中探索这一点（第5节）。μ̄μ̄.

定理4使用定义3中的不相似性 toiden-tifysufficient decrease of objective value at each iteration for FedProx. 见附录A.2我们提供了一个更常见的（尽管稍微严格一些）有界方差假设来表征性能。这一假设通常用于分析如SGD等方法。接下来，我们提供充分（但非必要）条件以确保P> 0在定理中。

4在每一轮之后，这样的足够减少是可实现的。

备注5.对于Pin定理4*为了得到正解，我们需要√B<1<1。这些条件有助于量化不相似性(B)和算法参数（√，K）之间的权衡。and* $\frac{B}{\sqrt{K}}$

最后，我们可以利用上述充分条件来表征在有界不相似性假设下，即假设1，收敛到近似稳定解集Ss={w|E[ⅡΣf(w)Ⅱ2]≤E} 的收敛速度。请注意，这些结果适用于一般的非凸F k（·）。

定理6（收敛率：FedProx)。给定某个E> 0，假设对于B≥B E，μ，√和K，满足定理4的假设*在每次迭代FedProx中保持不变。此外，f（w 0）−f* =△。然后，在T= O（）次迭代FedProx之后[Ⅱ∇f（w t）Ⅱ2]≤E。*
$\frac{\Delta}{\rho\epsilon}$, *we have* $\frac{1}{T}\sum_{t=0}^{T}$ −

虽然迄今为止的结果适用于非凸Fk（·），我们也可以用局部目标函数来表征凸损失函数的精确最小化情况下的收敛性（推论7）证明见附录A.3。

推论7（收敛性：凸情况）。设定理4中的断言*保持。此外，设Fk（·）为凸函数*
并且√t k= 0对于任意k，t，即所有局部问题都是
如果1冬B≤0.5√K，那么我们可以选择μ≈6 LB 2，由此可知P≈。$\frac{1}{24LB^2}$

注意假设1中的小E转换为更大的B E。推论7建议，为了使用FedProx以越来越高的精度解决问题，需要increaseμ适当。我们在第5.3节中通过实证验证了thatμ> 0会导致更稳定的收敛。此外，在推论7中，如果我们对B E的上界进行插值，在有界变化假设下（推论10)，达到主精度所需的步数为 $O(\frac{L\Delta}{\epsilon}+\frac{L\Delta\sigma^2}{\epsilon^2})$。我们的分析有助于描述当局部函数不同时，FedProx和类似方法的性能。

备注8（与SGD的比较）.我们注意到FedProx实现了与SGD相同的渐近收敛保证：在方差有界假设下，对于小∈，如果用推论中的上界替换B∈，则
10并且选择足够大的μ，当子问题被精确求解且F k（·）是凸函数时，FedProx的迭代复杂度为O（+），与SGD (Ghadimi& Lan，$\frac{L\Delta}{\epsilon}\frac{L\Delta\sigma^2}{\epsilon^2}$2013).

为定理6中的比率提供背景在第8条中，我们将它与凸情况下的SGD进行比较。 总体而言，我们对FedProx的分析并未得出优于经典分布式SGD（不包括本地更新）的收敛速度——尽管FedProx在每次通信轮次中可能进行更多的本地更新。事实上，当数据以非同分布方式生成时，像FedProx这样的本地更新方案可能会表现得比分布式SGD更差。因此，我们的理论结果并不一定证明FedProx优于分布式SGD；相反，它们提供了FedProx收敛的充分（但非必要）条件。我们的分析是我们所知的第一个用于分析任何联邦（即使用本地更新方案和低设备参与度）优化方法的分析，针对问题（1）。在异构环境中。

最后，我们注意到，先前的分析假设系统没有异质性，并且所有设备和迭代使用相同的γ。然而，我们可以扩展这些分析，允许γto随设备和迭代变化（如定义2所述。），这相当于允许设备根据当地系统条件执行可变数量的工作。我们提供收敛结果，其中γ值可变。
推论9（收敛性：Variableγ）。假设函数F k非凸，L-利普希茨光滑，并且存在L-> 0，使得▽2 F k≥—L-I，其中：=μ—L->$\bar{\mu}$
0.假设w t不是平稳解，且局部函数F k是B-不相似的，即B（w t）≤B。Ifμ，K，
以及算法2中的γt k选择这样的

$$\rho t=\frac{1}{\mu}\left(\frac{\gamma^t BB(1+\gamma^t)\sqrt{2}}{\bar{\mu}\sqrt{K}}-\frac{LB(1+\gamma^t)}{\bar{\mu}\mu}\right)$$

$$-2\sqrt{2}K+2>0\left(\frac{L(1+\gamma^t)^2B^2}{2\bar{\mu}^2}+\frac{LB^2(1+\gamma^t)^2}{\bar{\mu}^2K}\right)$$

然后在算法2的迭代t中，我们预计全球目标将出现以下下降：

$$E\,S\,t[\,f\,(w\,t+1)]\leq f\,(w\,t)-\rho t\,\|\nabla f\,(w\,t)\|\,2\,,$$

其中S t是第t次迭代中选择的K个设备andγ t= max k∈S tγ t k。

证明可以很容易地从对the-的证明中扩展出来
定理4，注意到E k[（1 + γ t k）‖▽F k（w t）‖]≤
(1 + max k∈ S tγ t k) E k[‖▽ F k (w t)‖] .

5个实验

现在我们给出广义FedProx框架的实证结果。在第5.2节，我们展示了FedProx在面对系统异质性时，能够容忍部分解决方案的性能改进。在第5.3节中，我们展示了FedProx在具有统计异质性（无论系统异质性）的设置中的有效性。我们还研究了统计异质性对收敛的影响（第5.3.1节)并展示经验收敛与我们的理论上的差异性假设（假设1 ）（第5.3.3节）在第5.1节中，提供实验装置的详细信息以及附录C。所有代码、数据和实验均可在github.com/litian96/FedPro x上公开获取。

5.1实验细节

我们评估了FedProx在多种任务、模型和真实世界联邦数据集上的表现。为了更好地表征统计异质性并研究其对收敛的影响，我们还在一组合成数据上进行了评估，这使得统计异质性的操控更加精确。我们通过向不同设备分配不同数量的本地工作来模拟系统的异质性。

合成数据。为了生成合成数据，我们遵循Shamir等人中的类似设置。（2014），添加 -ally imposing heterogeneity among devices. In particular, for each device k , we generate samples (X k , Y k)accord- ing to the model y= argmax (softmax (W x+ b)) ,x ∈ R 60 , W∈ R 10× 60 , b∈ R 10. We model W k~ N (u k, 1) , b k~ N (u k , 1), u k~ N (0 ,α); x k~ N (v k ,Σ), wherethe covariance matrixΣ is diagonal withΣj,j = j- 1 . 2. Each el- ement in the mean vector v k is drawn from N (B k, 1) , B k~ N (0 ,β) . Therefore, α controlshow much local models dif-fer from eachother and β controlshow much the local dataat each device differs from thatof othe r devices. We varyα,β to generate three heterogeneous distributed datasets, denoted Synthetic (α,β), as shown in Figure 2 . 我们还通过设置所有设备的相同的W、b并设置Xk以遵循相同的分布来生成一个IID数据集。我们的目标是学习一个全局的W和b。详细信息见附录C.1。

真实数据。我们还探索了四个真实数据集；统计信息总结见表1。这些数据集是根据联邦学习的先前工作以及最近的联邦学习基准(McMahan等人，2017;冷度以及其他人2018)我们研究了一个凸分类问题，使用MNIST (LeCun等人，1998)使用多项式逻辑回归。为了引入统计异质性，我们将数据分布在1,000个设备中，每个设备只有两位数的样本，且每个设备的样本数量遵循幂律分布。然后我们研究了一个更复杂的62类联邦扩展MNIST (Cohen等人，

2017;Caldas等人， 2018)（女性主义）数据集使用相同的模型。对于非凸设置，我们考虑在Sentiment140（Go等，2009）（Sent140）上的推文进行文本情感分析任务，其中每个推特账户对应一个设备。我们还研究了在《莎士比亚全集》数据集上进行下一个字符预测的任务(McMahan等，2017)（莎士比亚）。戏剧中的每个说话角色都与不同的设备相关联。数据集、模型和工作负载的详细信息见附录C.1。

表1.四个真实联邦数据集的统计信息。

数据集	设备	样品	样品/器械	
			意思	stdev
MNIST	1,000	69,035	69	106
女权主义者	200	18,345	92	159
莎士比亚	143	517,106	3,616	6,808
发送140	772	40,783	53	32

实现。我们实现了FedA vg（算法1）和FedProx（算法2)在Tensorflow中(Abadi等人，2016)为了与FedAvg进行公平比较，我们采用SGD作为FedProx的局部求解器，并采用与算法1略有不同的设备采样方案和2：均匀采样设备，然后用与局部数据点数量成比例的权重对更新进行平均(如McMahan等人最初提出的那样。(2017))虽然我们的分析不支持这种采样方案，但我们观察到无论是否使用，FedProx和FedAvg的相对行为相似。有趣的是，我们还发现本文提出的采样方案实际上使两种方法的表现更加稳定（见附录C.3.4，图12)这表明，该拟定框架具有额外的好处。详细信息见附录C.2。

超参数评估指标。对于每个数据集，我们在FedAvg上调整学习率（使用E=1且不考虑系统异质性），并在该数据集的所有实验中使用相同的初始学习率。我们还为所有数据集的所有实验设置了10个选定设备。在每次比较中，我们固定随机选择的设备、落后的设备和小批量顺序。我们报告基于全局目标函数f (w)的所有指标。请注意，在我们的模拟中（见第5.2节对于细节)，我们假设每个通信轮次对应一个特定的聚合时间戳（以现实世界的全球时钟时间来衡量）——因此，我们报告的结果是基于轮次而不是FLOP或时钟时间。详见附录C.2中的超参数细节。

5.2系统异质性：容忍部分工作

为了测量允许部分解决方案被发送到处理系统异质性与FedProx的效应，我们模拟了具有不同系统异质性的联邦设置，如下所述。

系统异构性模拟。我们假设训练过程中存在一个全局时钟，每个参与设备根据该时钟周期及其系统约束确定局部工作的量。这种指定的局部计算量对应于某种隐含的值√t k在第t次迭代时的设备k。在我们的模拟中，我们固定一个全局的周期数E，并根据当前系统约束，强制某些设备执行少于E个周期的更新。特别是，在不同的异构设置下，每轮我们分别为0%、50%和90%的选定设备分配x个周期（随机均匀选择[1，E]）。其中，0%的设备执行少于E个周期的工作对应于没有系统异构性的环境，而90%的设备发送部分解决方案则对应于高度异构的环境。当达到全局时钟周期时，FedAvg会简单地丢弃这些0%、50%和90%的落后的设备，而FedProx则会整合这些设备的部分更新。

如图1所示，我们把E设为20，研究aggre-门控部分工作对其他被删除设备的影响。合成数据集取自图（1,1）中的合成数据集

2 .我们发现，在所有数据集上，系统异质性对收敛有负面影响，且异质性越大，收敛效果越差（FedAvg）。与放弃更受限的设备（FedAvg）相比，分配不同量的工作(FedProx，μ=0）是有益的，这能带来更加稳定和更快的收敛。我们还观察到，在FedProx中设置settingμ> 0可以进一步提高收敛速度，具体讨论见第5.3节。

我们还研究了两种不那么异质的设置。首先，我们通过将E设置为1（即所有设备最多运行一个本地周期）来限制所有设备的能力，并以类似的方式施加系统异质性。我们在图9中展示了训练损失测试准确度见图10 附录中。即使在这些设置下，允许部分工作也能比FedAvg提高收敛性。其次，我们探索了一种没有统计异质性的设置，使用了等分布的合成数据集（Synthetic IID）。在这个IID设置中，如图5所示在附录C.3.2中，FedAvg在设备故障下表现相当稳健，能够容忍不同量级的本地工作，这可能不会带来重大改进。这为严格研究统计异质性对联邦学习新方法的影响提供了额外的动力，因为仅仅依赖IID数据（在实际应用中不太可能出现）可能无法揭示全部情况。

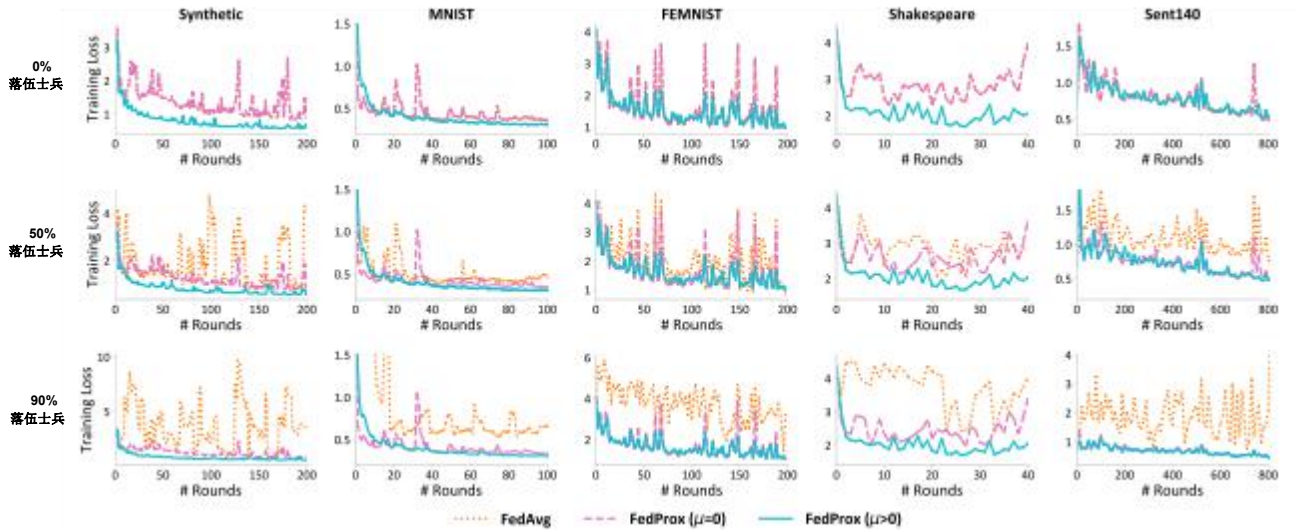

图1.在异构网络中，FedProx的收敛改进相对于FedAvg不显著。我们通过强制0%、50%和90%的设备成为落后的节点（被FedAvg摆脱）来模拟不同水平的系统异质性。(1)比较FedAvg和FedProx（μ = 0），我们发现允许执行可变数量的工作可以在系统异质性存在的情况下帮助收敛。(2)比较FedProx（μ> 0）和FedProx（μ = 0），我们展示了增加近似项的好处。当μ> 0时，FedProx导致更稳定的收敛，并使其他可能发散的方法在系统异质性存在(50%和90%落后节点）和无系统异质性(0%落后节点）的情况下都能收敛。请注意，当withμ=0且无系统异质性（没有落后节点）时，对应于FedAvg。我们在图7中还报告了测试准确性。，附录C.3.2 并且表明FedProx提高了所有数据集的测试精度。

5.3统计异质性：近端术语

为了更好地理解近端项如何在异质环境中带来益处，我们首先展示随着统计异质性的增加，收敛性可能会变得更糟。

5.3.1统计异质性的影响

见图2（第一行），我们研究了统计异质性如何影响收敛，使用四个没有系统异质性的合成数据集（固定E为20)。从左到右，随着数据变得更加异质，当FedProx withμ= 0时（即FedAvg)，收敛性能变差。尽管对于IID数据可能会减缓收敛速度，但我们发现settingμ>0在异构环境中特别有用。这表明，在具有不同统计异质性的实际联邦设置中，FedProx引入的修改子问题可以带来益处。对于完全IID的数据，一些启发式方法如decreasingμif损失继续减少可能有助于避免收敛减速（见图11见附录C.3.3)在接下来的部分中，我们在非合成实验中看到了类似的结果。

5.3.2效果ofμ> 0

影响FedProx性能的关键参数是局部工作量（由局部周期数E参数化），以及由μ缩放的近端项。直观上，较大的E可能导致局部模型偏离太远远离初始起点，从而导致潜在的分歧(McMahan等人，2017)因此，为了处理非IID数据下的FedAvg的发散或不稳定问题，仔细调整E是有帮助的。然而，E受到底层系统环境对设备的影响，很难为所有设备确定一个合适的统一E。相反，允许设备特定的E（变量√）并调整bestμ（可以视为E的重新参数化的一个参数），以防止发散并提高方法的稳定性是有益的。properμ可以通过限制迭代轨迹更接近全局模型的轨迹来实现这一点，从而引入不同数量的更新并保证收敛（定理6）。.

我们在图1中展示了FedProx（μ> 0）中近端项的影响。对于每次实验，我们比较了FedProx与μ = 0的结果以及FedProx与最佳μ的结果（关于如何selectμ的讨论见下一段）。对于所有数据集，我们观察到appropriateμ可以提高不稳定方法的稳定性，并能迫使发散方法收敛。无论系统异质性如何(50%和90%的滞后者），还是没有系统异质性(0%的滞后者），这一效果均成立。μ> 0在大多数情况下也能提高准确性（见图6）。图7见附录C.3.2)尤其在高度异质性环境(90%的落后的用户）中，FedProxim相比FedAvg的绝对测试准确率平均提高了22%（见图7）.

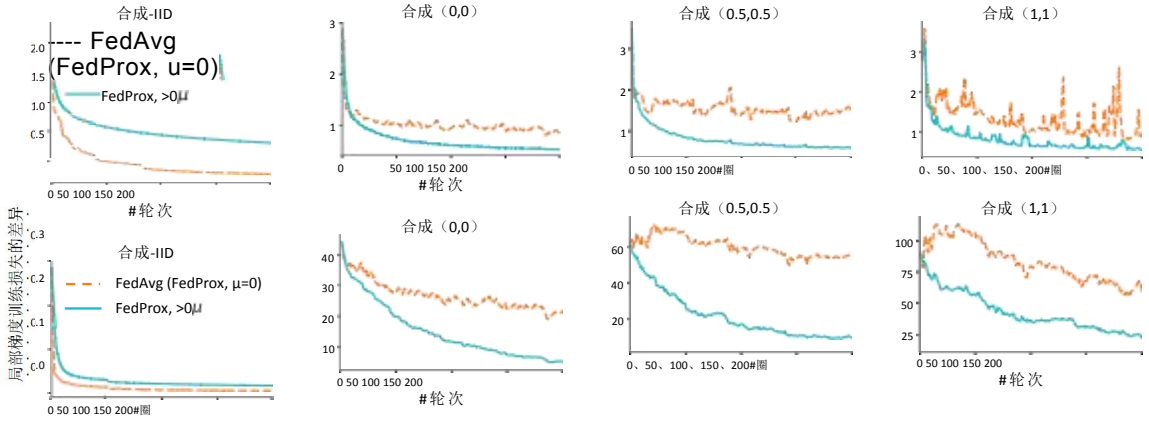

图2.数据异质性对收敛的影响。我们通过强制每个设备运行相同数量的周期来消除系统异质性的影响。在这种设置下，μ = 0的FedProx将简化为FedAvg。(1)上排：显示训练损失（详见附录C.3中的测试准确率结果。，图6)在四个合成数据集上进行测试，这些数据集的统计异质性从左到右逐渐增加。请注意，当μ = 0时，该方法对应于FedAvg。异质性的增加会导致收敛效果变差，但settingμ>0有助于缓解这一问题。(2)底部行：我们展示了四个合成数据集对应的不相似性度量（梯度方差）。这一指标捕捉了统计异质性，并且与训练损失一致——不相似性越小，表示收敛效果越好。

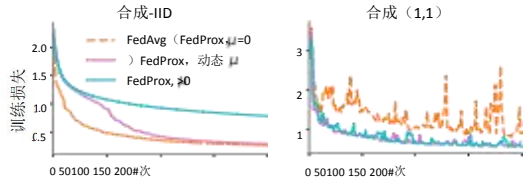

图3.基于当前模型性能的settingμ有效性。每当损失增加时，我们increaseμ0.1；每当损失减少时，我们减小0.1

5个连续的轮次。我们对合成IID的initializeμ值设为1（为了使我们的方法与我们的方法一致），而对合成IID的initializeμ值设为0（1,1）。这个简单的启发式方法在实验中表现良好。

选择μ。一个自然的问题是确定如何设置近端项中的惩罚常数μ。较大的μ可能会通过迫使更新接近起始点而减缓收敛速度，而较小的smallμ可能不会产生任何影响。在所有实验中，我们从有限候选集{0.001,0.01,0.1,1}中调整bestμ。对于图1中的五个联邦数据集bestμ值分别为1、1、1、0.001和0.01。虽然根据我们的理论结果直接实现自动调整μ较为困难，但在实际应用中，我们注意到thatμ可以根据模型当前性能自适应地选择。例如，一种简单的启发式方法是在损失增加时增加μ，在损失减少时增加decreasingμ。如图3所示我们使用两个合成数据集展示了这一启发式方法的有效性。请注意，这些initialμ值与我们的方法相对立。我们在附录C.3.3中提供了完整结果，展示了该方法的竞争性能。未来的工作包括开发方法，以自动调整该参数，用于异构数据集，例如基于这里提供的理论基础。

5.3.3Dissimilarity测量和发散

最后，在图2中 （最后一行），我们证明了我们在定义3中的B局部不相似性测量捕捉数据集的异质性，因此是性能的一个恰当代理。特别是，我们跟踪每个设备上的梯度方差，E k[‖▽F k(w)-▽f(w) ‖^2]，其下限由B E给出（见有界方差等价性推论10)。实证观察表明，增加μ会导致局部函数F k之间的差异减小，且差异度量与训练损失一致。因此，较小的差异意味着更好的收敛性，这可以通过适当使用settingμap来实现。我们还在附录C. 3.2中展示了差异度量在实际联邦数据中的表现。

6 C ONCLUSION

在这项工作中，我们提出了FedProx，一个优化框架，旨在解决联邦网络中固有的系统和统计异质性问题。FedProx允许在不同设备之间进行可变量的工作，并依赖于近似项来帮助稳定方法。我们提供了在设备差异假设下，FedProx在实际联邦设置中的收敛保证，同时考虑了诸如落后的节点等实际问题。我们在一系列联邦数据集上的实证评估验证了我们的理论分析，并证明了FedProx框架可以显著改善实际异构网络中联邦学习的收敛行为。

致谢

我们感谢Sebastian Caldas、Jakub Kone n‘ y、Brendanč McMahan，Nathan Srebro和Jianyu Wang对他们的帮助-

进行了充分的讨论。AT和VS部分得到了DARPA FA875017C0141、美国国家科学基金会资助IIS 1705121和IIS 1838017、冈川奖学金、谷歌教员奖、亚马逊网络服务奖、日本摩根人工智能研究教员奖、卡内基博世研究所研究奖以及CONIX研究中心的支持，后者是JUMP中六个中心之一，由DARPA赞助的半导体研究公司（SRC）计划。本文中表达的任何观点、发现、结论或建议均属于作者（们）所有，并不一定反映DARPA、美国国家科学基金会或任何其他资助机构的观点。

R EFERENCES

Tensorflow联邦：机器学习去中心化数据。URLhttps://www.tensorflow.org/使结成联邦

Abadi，M.，Barham，P.，Chen，J.，Chen，Z.，Davis，A.，Dean，J.，Devin，M.，Ghemawat，S.，Irving，G.，Isard，M.，Kudlur M. K.，Levenberg，J.，Monga，R.，Moore，S.，Murray，D. G.，Steiner，B.，Tucker，P.，Vasudevan，V.，Warden，P.，Wicke，M.，Yu，Y.和Zheng，X. Tensorfl ow：大规模机器学习系统。在操作系统设计与实现中，2016年。

Allen-Zhu，Z.如何使梯度最小化随机地：更快的凸和非凸sgd。在神经信息处理系统进展，2018年。

博纳维茨，K.，艾希纳，H.，格里斯坎普，W.，胡巴，D.，英格曼，A.，伊万诺夫，V.，基顿，C.，科内奇尼，J.，马佐奇，S.，麦克马汉，H. B.，奥弗维尔特，T. V.，彼得鲁，D.，拉姆扎德，D.和罗森兰德，J.走向大规模联邦学习：系统设计。在机器学习与系统会议上，2019年。

Boyd，S.，Parikh，N.，Chu，E.，Peleato，B.和Eckstein，J.通过交替乘子法实现分布式优化和统计学习。《机器学习基础与趋势》，2010年。

Caldas，S.，Wu，P.，Li，T.，Kone n`y，J.，McMahon，H.B.č Smith，V.和Talwalkar，A.Leaf：fed-federated settings的基准。arXiv预印本arXiv：1812.01097,2018。

Cohen，G.，Afshar，S.，Tapson，J.，和van Schaik，A. Em-nist：mnist到手写字母的扩展。arXiv *预印本arXiv：1702.05373,2017。*

Dean，J.，Corrado，G.，Monga，R.，Chen，K.，Devin，M.，Le Q.V.Mao，M.Ranzato，M.Senior，A.Tucker，P.Yang，K.和Ng，A.大规模分布式去电网络。《神经信息处理系统进展》，2012年。

Dekel，O.，Gilad- Bachrach，R.，Shamir，O.和Xiao，L.使用Mini-batch进行最优分布式在线预测。机器学习研究杂志，2012年。

Ghadimi，S.和Lan，G.非凸随机规划的随机一阶和零阶方法. SIAM Journal on Optimization，2013年。

G.O.A.，B.HAYANI和Huang，L.，使用远距离监督的推特情绪分类。CS 224 N项目报告，斯坦福大学，2009年。

Goldblum，M.，Reich，S.，Fowl，L.，Ni，R.，Cherepanova，V.和Goldstein，T.揭示元学习：理解少样本任务的特征表示。arXiv预印本arXiv：2002.06753,2020。

郝宇，荣杰，宋燕。分布式非凸优化中通信效率动量sgd的线性加速分析。国际机器学习会议，2019年。

黄磊、尹宇、付志、张超、邓海、刘德. Loadaboost：基于损失的基于医学数据的基于广告的联邦机器学习。arXiv预印本arXiv：1811.12629,2018。

Jeong E.、Oh S.、Kim H.、Park J.、Bennis M.和Kim S. L.设备上的通信效率高的机器学习：非独立同分布私有数据下的联邦蒸馏和增强。arXiv预印本arXiv：1811.11479,2018。

Jiang，P.和Agrawal，G.分布式深度学习的稀疏和量化通信的线性加速分析。在神经信息处理系统进展，2018年。

Kaczmarz，S.线性方程组的近似解.《国际控制杂志》，1993年。

Khodak，M.，Balcan，M.-F. F.，和Talwalkar，A.S.Adaptive gradient-based meta-learning methods. In Advances in Neural Information Processing Systems，2019.

LeCun，Y.，Bottou，L.，Bengio，Y.和Haffner，P.基于梯度的学习应用于文档识别.IEEE会议录，1998年。

李，M.，安德森，D. G.，斯莫拉，A. J.和Yu，K.《参数服务器的通信高效分布式机器学习》，见《神经信息处理系统进展》，2014a。

李敏、张涛、陈勇、斯莫拉，A. J.高效随机优化的微批训练。见2014b年知识发现与数据挖掘会议。

Li，T.，Sahu，A.，Talwalkar，A.和Smith，V. Federated学习：挑战、方法和未来方向。*arXiv预印本arXiv：1908.07873,2019*。

李，T.，萨胡，A. K.，扎希尔，M.，桑贾比，M.，塔尔瓦卡尔，A.和史密斯，V. Feddane：一种联邦牛顿型方法。arXiv预印本arXiv：2001.01920,2020。

林，T.，Stich，S.U.和Jaggi，M.不要使用大的小批量，使用局部梯度下降法。在国际学习表示会议，2020年。

McMahan，H. B.，Moore，E.，Ramage，D.，Hampson，S.和Arcas，B. A. y.《从分散数据中高效学习深度网络》.《2017年国际人工智能与统计会议》。

Pennington，J.，Socher，R.和Manning，C. Glove：全球词汇表示向量。见自然语言处理中的经验方法，2014年。

Reddi，S. J.，Kone n`y，J.，Richt rik，P.，P cz s，B.čáóó Smola，A. Aide： Fast and Communicationefficient Distributed optimization. arXiv preprint arXiv：1608.06879,2016年。

Richt rik，P.和Tak，M.分布式协调下降法在大数据学习中的应用，机器学习研究杂志，2016年。ááč

施密特，M.和鲁克斯，N. L.强增长条件下的随机梯度下降的快速收敛。arXiv预印本arXiv：1308.6370,2013。

Shamir O.，Srebro N.和Zhang T.使用近似牛顿型方法的通信效率高的分布式优化。在2014年国际机器学习会议上。

Smith，V.，Chiang，C.-K.，Sanjabi，M.和Talwalkar，A. S. Federatedmulti-task learning。见《神经信息处理系统进展》，2017年。

Smith，V.，Forte，S.，Ma，C.，Takac，M.，Jordan，M.I.和Jaggi，M. Cocoa：一种通用的通信高效分布式优化框架。机器学习研究杂志，2018。

Stich，S.U. Local sgd co co nverges fast and communicates little。在2019年国际学习表示会议中。

Strohmer，T.和Vershynin，R.随机kaczmarz算法具有指数收敛性，傅里叶分析与应用杂志，2009年。

Vaswani，S.，Bach，F.和Schmidt，M. Fast andfaster convergence of sgd for over-parameterized models（and accelerated perceptron)。在国际人工智能与统计会议，2019年。

王，J.和乔希，G. Cooperative sgd：A统一的框架用于设计和分析通信效率gd算法s。arXiv预印本arXiv：1808.07576,2018。

王，S.，图尔，T.，萨洛尼迪斯，T.，梁，K. K.，马卡亚，C.，何，T.，和陈，K.《资源受限边缘计算系统中的自适应联合学习》。《IEEE选择领域通信杂志》，2019年。

伍德沃斯，B. E.，王，J.，史密斯，A.，麦克马汉，B.和斯雷布罗，N.并行随机优化的图预言模型、下界和差距。见《神经信息处理系统进展》，2018年。

Yao，Y.，Rosasco，L.和Caponnetto，A.关于梯度下降学习中的提前停止。构造性逼近，2007年。

Yin，D.，Pananjady，A.，Lam，M.，Papailiopoulos，D.，Ram- chandran，K.，和Bartlett，P.梯度多样性：可扩展分布式学习的关键成分。在国际人工智能与统计会议，2018年。

余，H.，杨，S.，和朱，S.并行重启sgd用于非凸优化，具有快速收敛和较少通信。在AAA I人工智能会议，2018年。

张，S.，Choromanska，A. E.，和LeCun，Y. Deep learning with elastic averaginggd.在神经信息处理系统进展中，2015年。

张宇、杜奇、Wainwright M. J.《统计优化的通信效率算法》，《机器学习研究杂志》，2013。

赵、李、赖、苏达、辛、钱德拉V.非iId数据的联合学习。arXiv prep rint arXiv：1806.00582,2018。

周福，Cong G.关于非凸优化中k步平均随机梯度下降算法的收敛性。国际人工智能联合会议，2018年。

周鹏，袁翔，徐浩，严胜，冯杰.通过小蝙蝠近端更新实现高效的元学习.《神经信息处理系统进展》，2019年。

完整证明

A.1定理4的证明

证明。使用我们对每个局部求解器的ν不精确性的概念（定义1），我们可以定义e t k+1，使得：

$$\nabla F_k(w_k^{t+1}) + \mu(w_k^{t+1} - w^t) - e_k^{t+1} = 0,$$

$$\|e_k^{t+1}\| \leq \nu \|\nabla F_k(w^t)\|. \quad (3)$$

现在我们来定义w-t+1 = E k[w t k+1]。根据这个定义，我们知道

$$\bar{w}^{t+1} - w^t = \frac{-1}{\mu}\mathbb{E}_k\left[\nabla F_k(w_k^{t+1})\right] + \frac{1}{\mu}\mathbb{E}_k\left[e_k^{t+1}\right]. \quad (4)$$

我们定义=μ-L->0和w^ t k+1 = arg min w h k（w；w t）。那么，由于h k的theμ强凸性，我们有 $\bar{\mu}$

$$\|\hat{w}_k^{t+1} - w_k^{t+1}\| \leq \frac{\gamma}{\bar{\mu}}\|\nabla F_k(w^t)\|. \quad (5)$$

请注意，由于h k的theμ强凸性，我们知道thatǁ w^ t k+1-wtǁ≤ǁ ▽Fk（w t）ǁ。现在我们可以使用$\frac{1}{\bar{\mu}}$三角不等式得到

$$\|w_k^{t+1} - w^t\| \leq \frac{1+\gamma}{\bar{\mu}}\|\nabla F_k(w^t)\|. \quad (6)$$

因此

$$\|\bar{w}^{t+1} - w^t\| \leq \mathbb{E}_k\left[\|w_k^{t+1} - w^t\|\right] \leq \frac{1+\gamma}{\bar{\mu}}\mathbb{E}_k\left[\|\nabla F_k(w^t)\|\right] \leq \frac{1+\gamma}{\bar{\mu}}\sqrt{\mathbb{E}_k[\|\nabla F_k(w^t)\|^2]} \leq \frac{B(1+\gamma)}{\bar{\mu}}\|\nabla f(w^t)\|,$$

最后一个不等式是由于有界不相似性假设。

现在我们定义Mt+1，使得w-t+1-wt=▽f（w t）+ Mt+1 ），即Mt+1=Ek[▽F k（w t k+1）-▽F k（w t）- e t k+1]。$\frac{-1}{\mu}$(

我们可以通过boundǁ Mt+1ǁ：

$$\|M_{t+1}\| \leq \mathbb{E}_k\left[L\|w_k^{t+1} - w_k^t\| + \|e_k^{t+1}\|\right] \leq \left(\frac{L(1+\gamma)}{\bar{\mu}} + \gamma\right) \times \mathbb{E}_k\left[\|\nabla F_k(w^t)\|\right] \leq \left(\frac{L(1+\gamma)}{\bar{\mu}} + \gamma\right) B\|\nabla f(w^t)\|, \quad (7)$$

最后一个不等式也是由于有界不相似性假设。根据L-Lipschitz平滑性和泰勒展开，我们有

$$\begin{aligned} f(w\text{-}t+1) &\leq f(w\ t) + (\nabla\ f(w\ t),\ w\text{-}t+1\ -\ w\ t\rangle\ +\ \text{ǁ}\ w\text{-}t+1\ -\ w\ t\text{ǁ}\ 2\frac{L}{2} \\ &\leq f(w\ t)\ -\ \text{ǁ}\nabla\ f(w\ t)\text{ǁ}2\ -\ (\nabla\ f(w\ t),\ M\ t+1\rangle\ +\text{ǁ}\nabla\ f(w\ t)\text{ǁ}\ 2\ \frac{1}{\mu}\ \frac{1}{\mu}\ \ \frac{L(1+\gamma)^2B^2}{2\bar{\mu}^2} \\ &\leq f(w^t) - \left(\frac{1-\gamma B}{\mu} - \frac{LB(1+\gamma)}{\bar{\mu}\mu} - \frac{L(1+\gamma)^2B^2}{2\bar{\mu}^2}\right) \times \|\nabla f(w^t)\|^2. \end{aligned} \quad (8)$$

从上述不等式可以得出，如果我们设置惩罚项parameterμ足够大，就能使f（w-t+1）- f（w t）的目标函数值减少，且减少量与toǁ▽f（wt）ǁ^2成正比。然而，这并不是算法的实际工作方式。在算法中，我们仅使用随机选择的K个设备来近似w-t。因此，在寻找E[f（wt+1）]时，我们利用了函数f的局部利普希茨连续性。

$$f(w\ t+1) \leq f(w\text{-}t+1) + L\ 0\text{ǁ}\ w\ t+1\ -\ w\text{-}t+1\text{ǁ}\ , \quad (9)$$

其中L0是函数f的局部利普希茨连续常数，我们有

L0≤ǁ▽f（w t）ǁ + L max（ǁ w-t+1 - w tǁ，ǁ w t+ 1- w tǁ）≤ǁ▽f（w t）ǁ + L（ǁ w-t+1- w tǁ +ǁ w t+1 - w tǁ）。因此，如果我们对第t轮中设备的选择进行期望值计算，需要限制

$$E\ S\ t[f(w\ t+1)] \leq f(w\text{-}t+1) + Q\ t\ , \quad (10)$$

其中Q t= E S t[L0 Ⅱ w t+1 - w-t+1 Ⅱ]。注意，期望是在随机选择要更新的设备时取的。

$$\begin{aligned} Q_t &\leq \mathbb{E}_{S_t}\left[\left(\|\nabla f(w^t)\| + L(\|\bar{w}^{t+1} - w^t\| + \|w^{t+1} - w^t\|)\right) \times \|w^{t+1} - \bar{w}^{t+1}\|\right] \\ &\leq \left(\|\nabla f(w^t)\| + L\|\bar{w}^{t+1} - w^t\|\right)\mathbb{E}_{S_t}\left[\|w^{t+1} - \bar{w}^{t+1}\|\right] + L\mathbb{E}_{S_t}\left[\|w^{t+1} - w^t\| \cdot \|w^{t+1} - \bar{w}^{t+1}\|\right] \\ &\leq \left(\|\nabla f(w^t)\| + 2L\|\bar{w}^{t+1} - w^t\|\right)\mathbb{E}_{S_t}\left[\|w^{t+1} - \bar{w}^{t+1}\|\right] + L\mathbb{E}_{S_t}\left[\|w^{t+1} - \bar{w}^{t+1}\|^2\right] \end{aligned} \tag{11}$$

从（7)，我们有that Ⅱ w-t+1-wt Ⅱ≤Ⅱ ▽f（w t） Ⅱ。此外，$\frac{B(1+\gamma)}{\bar{\mu}}$

$$\overline{\text{E S t[Ⅱ w t+1 — w-t+1 Ⅱ]}}\leq\sqrt{\text{E S t[Ⅱ w t+1 — w-t+1 Ⅱ 2]}} \tag{12}$$

和

$$\begin{aligned} &\text{E S t[Ⅱ w t+1 — w-t+1 Ⅱ 2]}\leq \text{E k[Ⅱ w t k+1 — w-t+1 Ⅱ 2]}\frac{1}{K} \\ &\leq \text{Ek[ⅡwⅡ/tk+1-wtⅡ2]，（因为w-t+1=Ek[w/tk+1]]）}\frac{2}{K} \\ &\leq \frac{2}{K}\frac{(1+\gamma)^2}{\bar{\mu}^2}\mathbb{E}_k\left[\|\nabla F_k(w^t)\|^2\right] \quad \text{(from (6))} \\ &\leq \frac{2B^2}{K}\frac{(1+\gamma)^2}{\bar{\mu}^2}\text{Ⅱ▽ f (13)}(w^t)\|^2, \end{aligned}$$

第一个不等式是由于随机选择K个设备来获得wt而产生的，最后一个不等式是由于有界差异假设而产生的。如果我们用(11)我们得到了

$$Q_t \leq \left(\frac{B(1+\gamma)\sqrt{2}}{\bar{\mu}\sqrt{K}} + \frac{LB^2(1+\gamma)^2}{\bar{\mu}^2 K}\left(2\sqrt{2K}+2\right)\right)\|\nabla f(w^t)\|^2 \tag{14}$$

组合(8), (10), (9)和(14)并使用notationα=我们得到$\frac{1}{\mu}$

$$\begin{aligned} &\mathbb{E}_{S_t}\left[f(w^{t+1})\right] \leq f(w^t) - \left(\frac{1}{\mu} - \frac{\gamma B}{\mu} - \frac{B(1+\gamma)\sqrt{2}}{\bar{\mu}\sqrt{K}} - \frac{LB(1+\gamma)}{\bar{\mu}\mu}\right. \\ &\text{—2√(2K+2))Ⅱ▽f(wt)Ⅱ2.}\frac{L(1+\gamma)^2B^2}{2\bar{\mu}^2}\frac{LB^2(1+\gamma)^2}{\bar{\mu}^2K}\Big(\end{aligned}$$

Ⅰ Ⅰ

A.2有界方差的证明

推论10（有界方差等价性）。设假设1*那么，在有界方差的情况下，即≤σ2，对于任意E> 0，有B*
$E\leq\sqrt{\mathbb{E}_k\left[\|\nabla F_k(w) - \nabla f(w)\|^2\right]1 + \frac{\sigma^2}{\epsilon}}$.

证明：我们有，

$$\begin{aligned} &\text{E k[Ⅱ▽ F k (w) - ▽ f (w)Ⅱ 2] = E k[Ⅱ▽ F k (w)Ⅱ 2] - Ⅱ▽ f (w)Ⅱ 2≤ σ 2} \\ &\text{) E k[ⅡⅡ▽ F k (w)ⅡⅡ 2]≤ σ 2+ⅡⅡ▽ f (w)ⅡⅡ 2} \\ &\text{) B E= ≤ √ 1 + .}\sqrt{\frac{E_k[\|\nabla F_k(w)\|^2]\sigma^2}{\|\nabla f(w)\|^2 \quad \epsilon}} \end{aligned}$$

根据推论10在原位，我们可以重述定理4中的主要结果在有界方差假设方面。

定理11（非凸FedProx收敛：有界方差）。设定理4的断言 *保持不变。此外，令迭代w t满足that Ⅱ▽f（w t）Ⅱ2≥E，并且令E k[Ⅱ ▽F k (w) - ▽f(w) Ⅱ2]≤σ2满足不相似性条件。如果算法2中的μ、K和√选择如此之好*

$$\rho=\left(\frac{1}{\mu}-\left(\frac{\gamma}{\mu}+\frac{(1+\gamma)\sqrt{2}}{\bar{\mu}\sqrt{K}}+\frac{L(1+\gamma)}{\bar{\mu}\mu}\right)\sqrt{1+\frac{\sigma^2}{\epsilon}}-\left(\frac{L(1+\gamma)^2}{2\bar{\mu}^2}+\frac{L(1+\gamma)^2}{\bar{\mu}^2K}\left(2\sqrt{2K}+2\right)\right)\left(1+\frac{\sigma^2}{\epsilon}\right)\right)>0,$$

然后在算法2的第t次迭代中，我们有以下预期的全球目标下降：

$$E\ S\ t[f\ (w\ t+1)]\leq f\ (w\ t\ )-P\amalg\bigtriangledown\ f\ (w\ t\ )\amalg\ 2\ ,$$

其中，St是第t次迭代中选择的K个设备的集合。

定理11的证明由Theo rem 4的证明得出通过注意到有界方差假设与由推论10所描述的不相似性假设之间的关系。

A.3推论7的证明

在凸情况下，当L- = 0且=μ，如果√=0，即所有子问题都精确求解时，我们可以通过to Ⅱ▽f（w t）Ⅱ 2获得相应的减少，前提是B<√K。在这种情况下，如果我们假设1<<B≤0.5√K，则可以表示为 $\bar{\mu}$ —

$$\mathbb{E}_{S_t}\left[f(w^{t+1})\right]\ \text{乏}\ f\amalg\triangledown\ f\amalg\triangledown\ f(w^t)-\frac{1}{2\mu}(w^t)\|^2+\frac{3LB^2}{2\mu^2}(w^t)\|^2\,. \tag{15}$$

在这种情况下，如果我们chooseμ≈6 LB2，我们得到

$$\mathbb{E}_{S_t}\left[f(w^{t+1})\right]\lessapprox f(w^t)-\frac{1}{24LB^2}\|\nabla f(w^t)\|^2\,. \tag{16}$$

注意（16）中的期望值是基于前一次迭代的条件期望值。对两边取期望值，然后展开，我们得到至少生成一个梯度平方范数小于E的解所需的迭代次数为O（）。$\frac{LB^2\Delta}{\epsilon}$

与其它单机和分布式方法的连接

所提出的两个方面——FedProx中的近端项和我们在分析中使用的有界不相似性假设——在优化文献中已经得到研究，但动机非常不同。为了完整性，我们在下面讨论了我们与这些先前工作的关系。

近端项。在FedProx中提出的改进目标函数与弹性平均SGD（EASGD）(Zhang等人， 2015)，这是在数据中心环境中训练深度网络的一种方法，以及在目标函数中使用了类似的近端项。虽然直觉上类似于EASGD（该项有助于防止每个设备/机器上的大幅偏差），但EASGD采用了一个更复杂的移动平均更新参数，仅限于使用SGD作为局部求解器，并且仅对简单的二次问题进行了分析。我们引入的近端项在先前的优化文献中也以不同的目的进行了探索，例如Allen-Zhu（2018），加快（小批量）在单台机器上进行SGD训练，并在Li等人中进行了研究。（2014b）在单台机器上实现SGD的有效训练以及分布式设置。然而，Li等人中的分析。（2014b）仅限于具有不同设置的单台机器假设（例如，IID数据和每轮精确解决子问题）。

此外，DANE (Shamir等人，2014)和AIDE (Reddi等人，2016)，为数据中心设计的分布式方法在局部目标函数中设置一个类似的近端项，但同时增加一个额外的梯度校正项。这两种方法都假设所有设备在每个通信轮次中都参与其中，在联邦设置中这是不切实际的。事实上，由于完全梯度（即▽φ（w（t-1））的估计不精确，Shamir等人指出。(2014 , Eq （13））)与设备子采样方案和梯度校正项的陈旧性(Shamir等人，2014 , Eq (13)),这些方法并不直接适用于我们的场景。尽管如此，我们探索了这种应用方法的一个变体，在联邦设置中发现梯度方向项在此情况下并无帮助——尽管需要额外计算，但其表现仍逊于提出的FedProx框架在异构数据集上的表现（见图4）。 我们建议感兴趣的读者参考Li等人。（2020）详细讨论.

最后，我们注意到元学习方法和联邦优化方法之间有一个有趣的联系(Khodak等人，2019)，最近在元学习的背景下，类似的近端项也被研究用于提高少样本学习任务的表现(Goldblum等人，2020;周等，201 9).

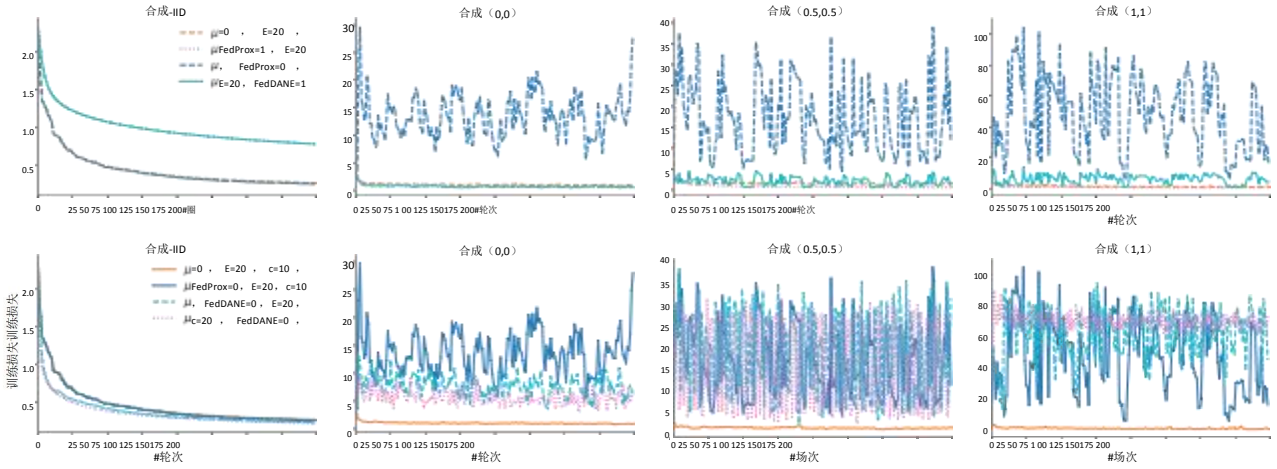

图4. DANE和AIDE (Shamir等人，2014;Reddi等人，2016)提出的方法在数据中心设置中使用了与FedProx相似的近似项以及额外的梯度校正项。我们修改了DANE以适应联邦环境，允许本地更新和设备参与度低的情况。我们在合成数据集上展示了这种改进方法的收敛性，我们称之为FedDane。在顶部图表中，我们在所有数据集中从30个设备中抽取1 0 个进行FedProx和FedDane的样本化。尽管FedDane在IID数据集上的表现与FedProx相似，但在非IID数据集上却存在较差的共收敛性。在底部图表中，我们展示了当增加选定设备数量以缩小估计的全梯度与真实全梯度（在梯度校正项中）之间的差距时，FedDane的结果。请注意，在实际环境中与所有（或大多数）设备通信已经不切实际。我们观察到，虽然每轮抽样更多设备可能在一定程度上有所帮助，但FedDane仍然不稳定且容易发散。这为我们在FedProx中提出的特定子问题提供了额外的动力。

有界不相似性假设。我们在假设1中讨论的有界不相似性假设以不同的形式出现，例如在Schmidt和Roux中(2013); 尹等。(2018);Vaswani等人。(201 9)在Yin等人中。（2018），在断言梯度多样性以及量化IID数据中小批量SGD的均方误差缩放带来的好处时，使用了有界相似性假设。在Schmidt和Roux的研究中(2013);Vaswani等人。（2019），作者使用了类似的假设，称为强增长条件，这是假设1的更强版本　当E= 0时，他们证明了一些有趣的实际问题满足这样的条件。他们还利用这一假设证明了具有常数步长的SGD算法具有最优且更好的收敛速度。需要注意的是，这与我们的方法不同，因为我们分析的算法不是SGD，尽管假设相似，但我们的分析方式有所不同。

c模拟详细信息和附加实验

C.1数据集和模型

在这里，我们提供了实验中使用的数据集和模型的详细信息。我们计算了一组多样化的非合成数据集，包括之前关于联邦学习的工作(麦克马汉等人，2017)，一些人提出了din LEAF，这是联邦设置的一个基准(Caldas等人，2018)我们还创建了合成数据，直接测试异质性对收敛的影响，如第5.1节所述。

合成：分别设置（α，β）=（0,0）、（0.5,0.5）和（1,1），生成三个分布不相同的非相同数据集（图2）在IID数据（图5中)，在所有设备上设置相同的W，b~ N（0,1），Xk遵循相同的分布N（v，Σ），其中均值向量的每个元素为零，Σ是对角矩阵，Σj；j=j-1.2。对于所有合成数据集，共有30个设备，每个设备上的样本数遵循幂律分布。

MNIST：我们研究了MNIST中手写数字0-9的图像分类(LeCun等人，1998)使用多项式逻辑回归。为了模拟异质环境，我们将数据分布在1000个设备上，每个设备只有2位数的样本，且每个设备的样本数量遵循幂律分布。模型的输入是一个784维（28×28)的图像，输出是一个0到9之间的类别标签。

·femnist：我们研究了62类EMNIST数据集上的图像分类问题(Cohen等人，2017)使用多项逻辑回归。为了生成异构数据分区，我们从EMNIST中抽取10个小写字母字符（'a'-'j'），并仅向每个设备分配5个类别。我们将这种EMNIST联邦版本称为EMNIST联邦主义。共有200个设备。模型的输入是一个784维（28×28)的图像，输出是一个0到9之间的类别标签。

·莎士比亚：这是一个基于《威廉·莎士比亚全集》(McMahan等人，2017)每种角色在戏剧中代表不同的装置。我们使用了一个包含100个隐藏单元的两层LSTM分类器，其中包含一个8维嵌入层。任务是预测下一个字符，共有80个字符类别。模型以80个字符的序列作为输入，将每个字符嵌入到学习到的8维空间中，并在经过2个LSTM层和一个全连接层后，输出每个训练样本的一个字符。

•Sent140：在非凸环境中，我们考虑了Sentiment140 (Go等人，2009)（发送140）使用包含256个隐藏单元的两层LSTM二分类器，预训练300 DGloVe嵌入(Pennington等人，2014)每个推特账号对应一个设备。模型以25个字符的序列作为输入，通过Glove查找将每个字符嵌入到300维空间中，在经过2个LSTM层和一个全连接层后，输出每个训练样本的一个字符。

C.2实施细节

（实现）为了与FedAvg进行公平比较，我们使用SGD作为FedProx的局部求解器，并采用与算法1中略有不同的设备采样方案和2：均匀采样设备，并按与局部数据点数量成比例的权重平均更新(最初由McMahan等人提出。(2017))虽然我们的分析不支持这种采样方案，但我们观察到无论是否采用FedProx还是FedAvg，其相对行为都相似（图12)有趣的是，我们还观察到，本文提出的采样方案使两种方法的性能更加稳定。这表明了所提出的框架具有额外的优势。

（机器）Wesimulate federated learning setup（1台服务器和N台设备）在一台商品机器上，该机器配备2个Intel o R Xeon o R E5-2650 v4 CPU和8个NVidia o R 1080 Ti GPU。

（超参数）我们在每个本地设备上随机将数据分为80%的训练集和20%的测试集。我们固定每轮选择的设备数量为10，适用于所有数据集的所有实验。我们还基于Feder平均值对学习率进行了网格搜索。我们没有在整个过程中衰减学习率。对于所有合成数据实验，学习率为0.01。对于MNIST、FEMNIST、Shakespeare和Sent140，我们使用的学习率分别为0.03、0.003、0.8和0.3。我们所有实验均使用10作为批量大小。

（库）所有代码都是在Tensorfl ow 1.10.1版本中实现的(Abadi等人，2016)请参见github. com/ litian96/ FedProx详情请参见。

C.3其他实验和完整结果

C.3.1系统异质性对IID数据的影响

我们展示了允许部分工作在完美的IID合成数据（Synthetic IID）上的效果。

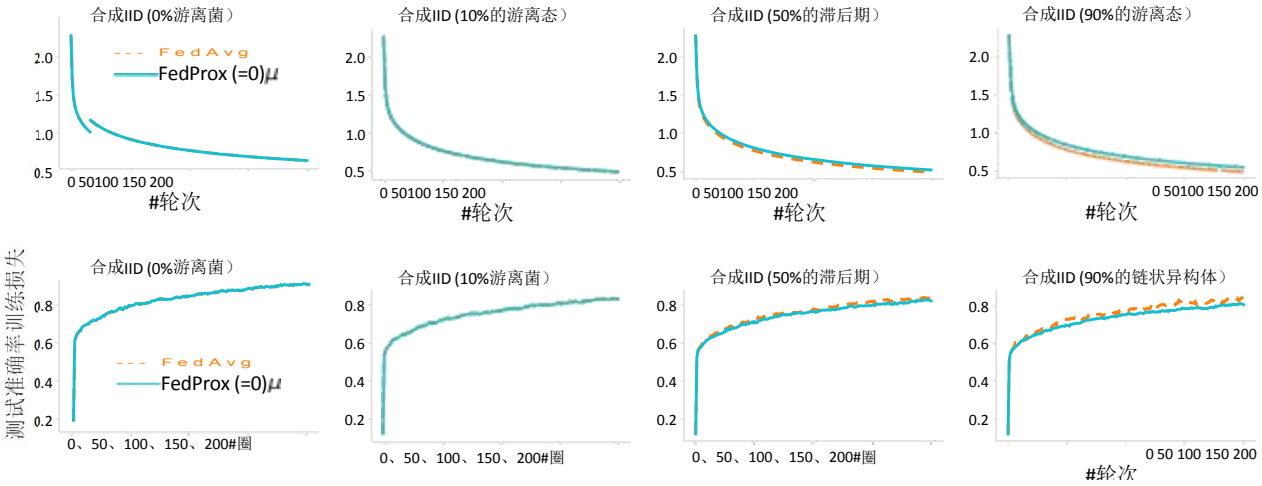

图5.IID数据表明，FedAvg对设备故障具有鲁棒性。在这种情况下，是否加入来自落后的节点的部分解决方案对收敛性影响不大。

C.3.2完整结果

见图6，我们在图2所示的实验相关的四个合成数据集上给出了测试精度。

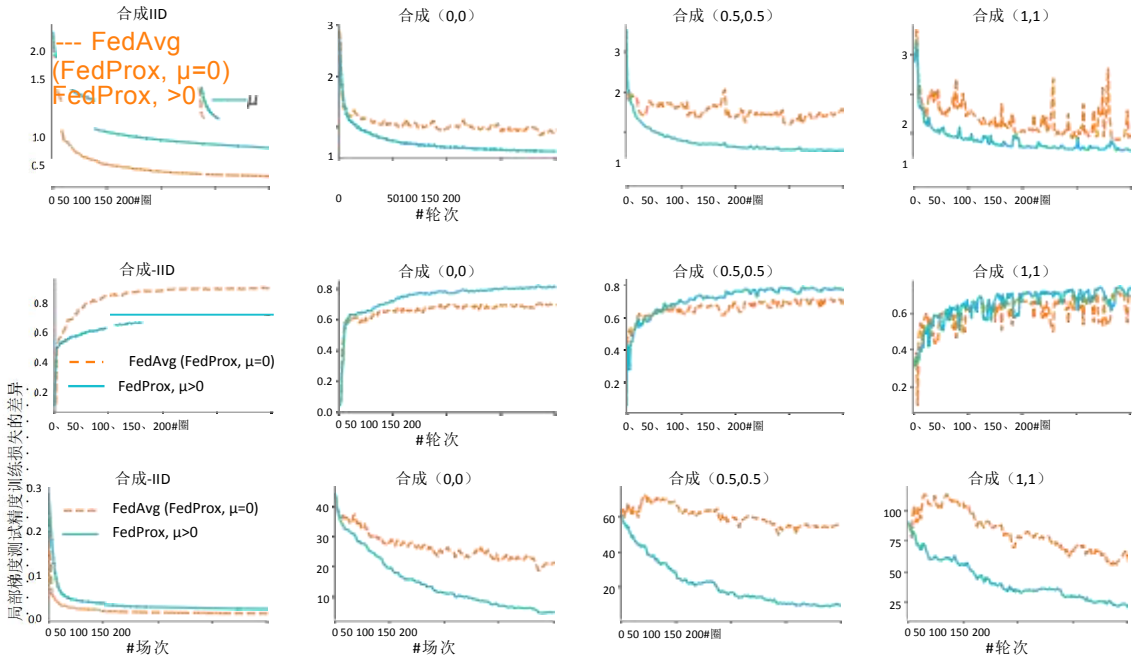

图6.图2中描述的实验的训练损失、测试准确率和不相似性测量。

见图7，我们展示了与图1中描述的实验相关的测试精度。我们通过识别FedProx和FedAvg在收敛、开始发散或运行足够多的轮次（例如1000轮）时的准确性来计算精度提升数值，以较早出现者为准。当两个连续轮次的损失差j f t- f t-1 j小于0.0001时，认为方法已收敛；而当我们看到f t- f t-10大于1时，则认为方法已发散。

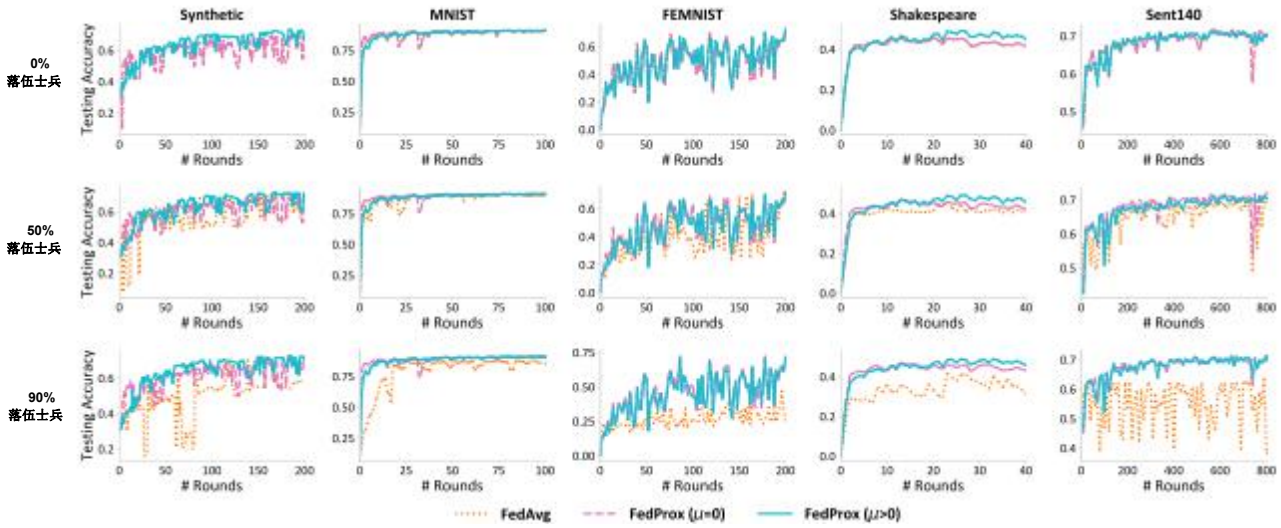

图7.图1中实验的测试精度。在高度异质化环境中（90%的落后的用户），FedProx在测试准确率方面平均提高了22%。

见图8，我们在图1中描述的5个数据集（包括4个真实数据集）上报告了不相似性测量。同样，不相似性特征与实际性能（损失）一致。

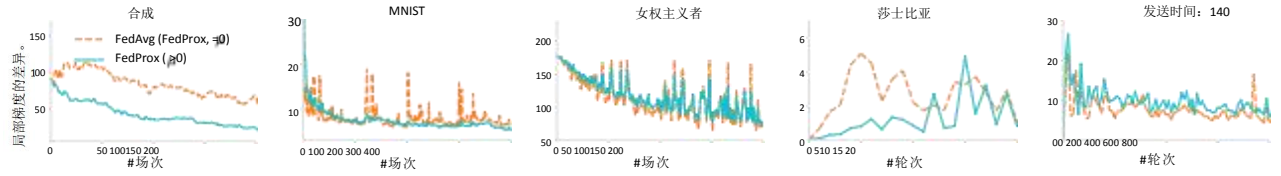

图8.图1中五个数据集的不相似度量。 我们仅考虑没有参与设备退出网络的情况，从而消除系统异质性。我们的不相似性假设捕捉了数据异质性，并且与实际性能一致（见图1中的训练损失）.

见图9图10，我们展示了允许部分解决方案在系统异质性下存在（即统计异质性不太可能对收敛产生负面影响）时（i.e.E=1）的影响（损失和测试准确性）。

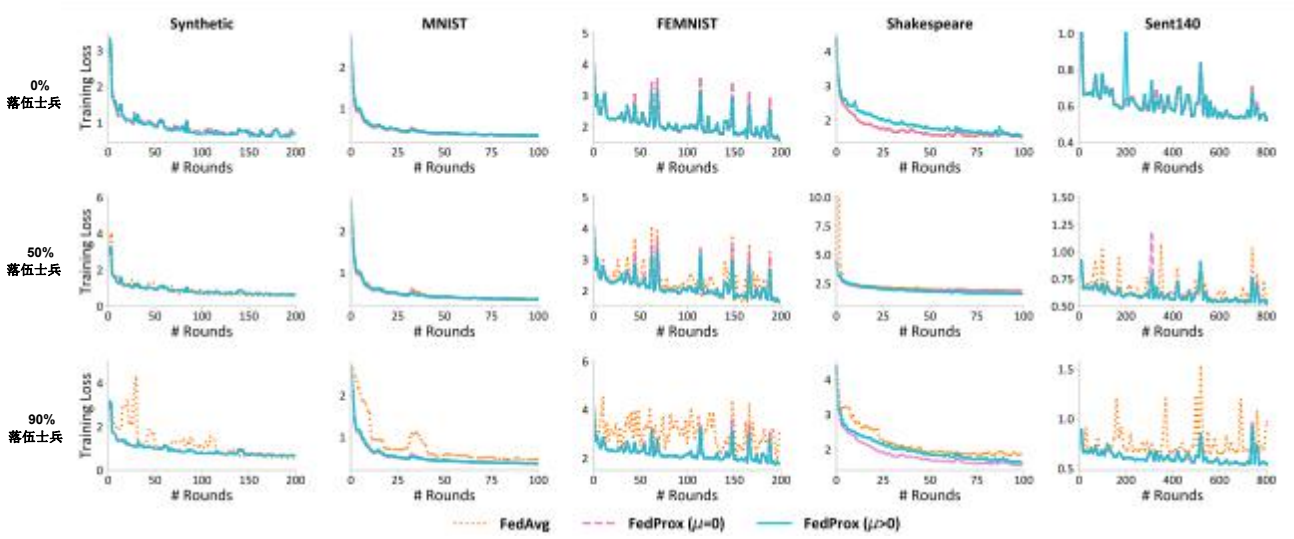

图9.在每个设备每次迭代最多运行1个周期（E= 1）的不同系统异质性设置下，FedAvg和FedProx的损失。由于局部更新与全局模型的偏差相比，在大E值下的偏差较小，因此统计异质性对收敛性产生负面影响的可能性较小。容忍部分解决方案发送到中央服务器（FedProx，μ = 0）的表现仍优于抛弃落后的节点（FedAvg）。

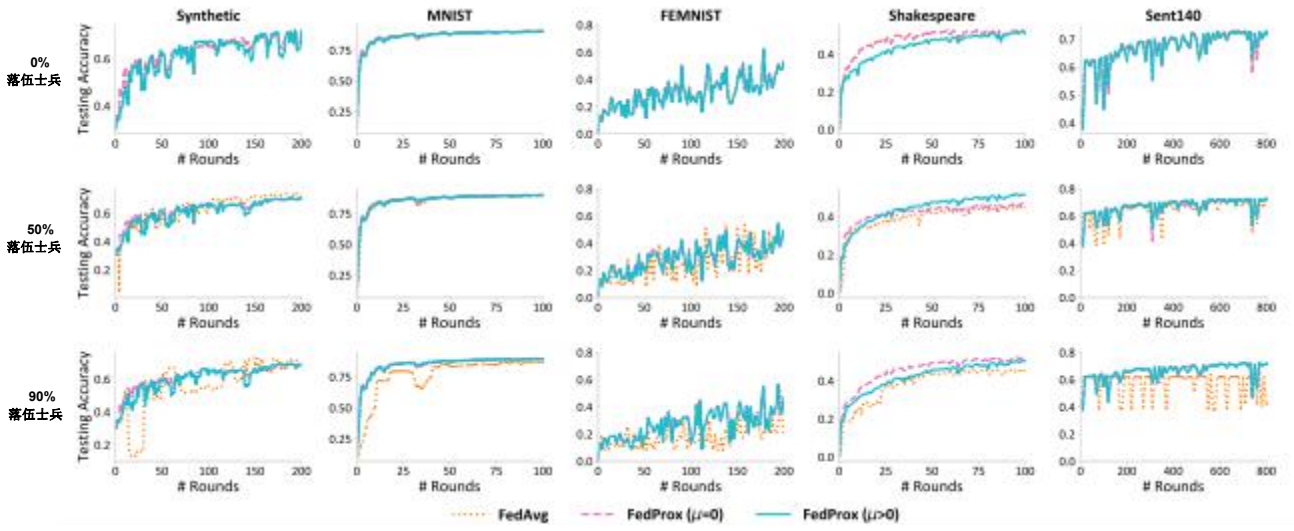

图10.图9所示实验的测试精度。

C.3.3 自适应settingμ

FedProx isμ的一个关键参数。我们在图11中提供了四个合成数据集上自适应settingμ简单启发式方法的完整结果。对于IID数据集（合成-IID），μ从1开始；而对于其他非IID数据集，μ从0开始。这种初始化方式对我们的方法不利。当损失继续下降5轮时，我们decreaseμ地将μ增加0.1；而当损失增加时，则将μ增加0.1。这一启发式方法能够保证性能竞争性。同时，它还可以缓解μ> 0可能在IID数据上减缓收敛速度的问题，这种情况在实际联邦设置中很少发生。

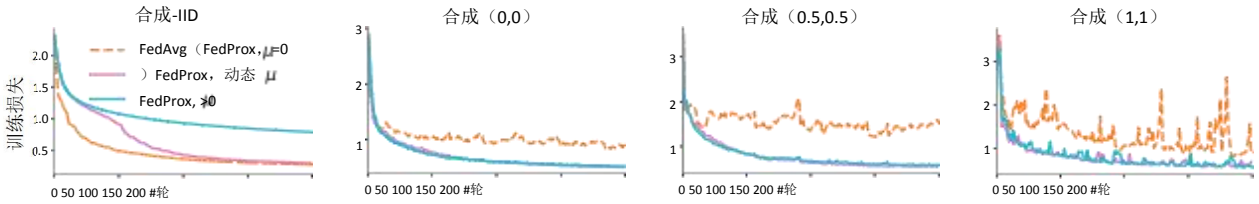

图11.choosingμ在所有合成数据集上的自适应完整结果。每当损失增加时，我们将其调整为0.1；每当损失减少时，我们将其调整为0.1，连续进行5次调整。对于IID数据（合成-IID），我们将initializeμ调整为1（以对抗我们的方法），而对于其他三个非IID数据集，则将其初始化为0。我们观察到，这种简单的启发式方法在实际应用中表现良好。

C.3.4比较两种器械取样方案

我们在图12中展示了FedProx在使用两种不同设备采样方案的合成数据集上的训练损失、测试准确率和不相似性测量。由于我们的目标是比较这两种采样方案，我们让每个设备对两种方法执行均匀的工作量（E= 20）。

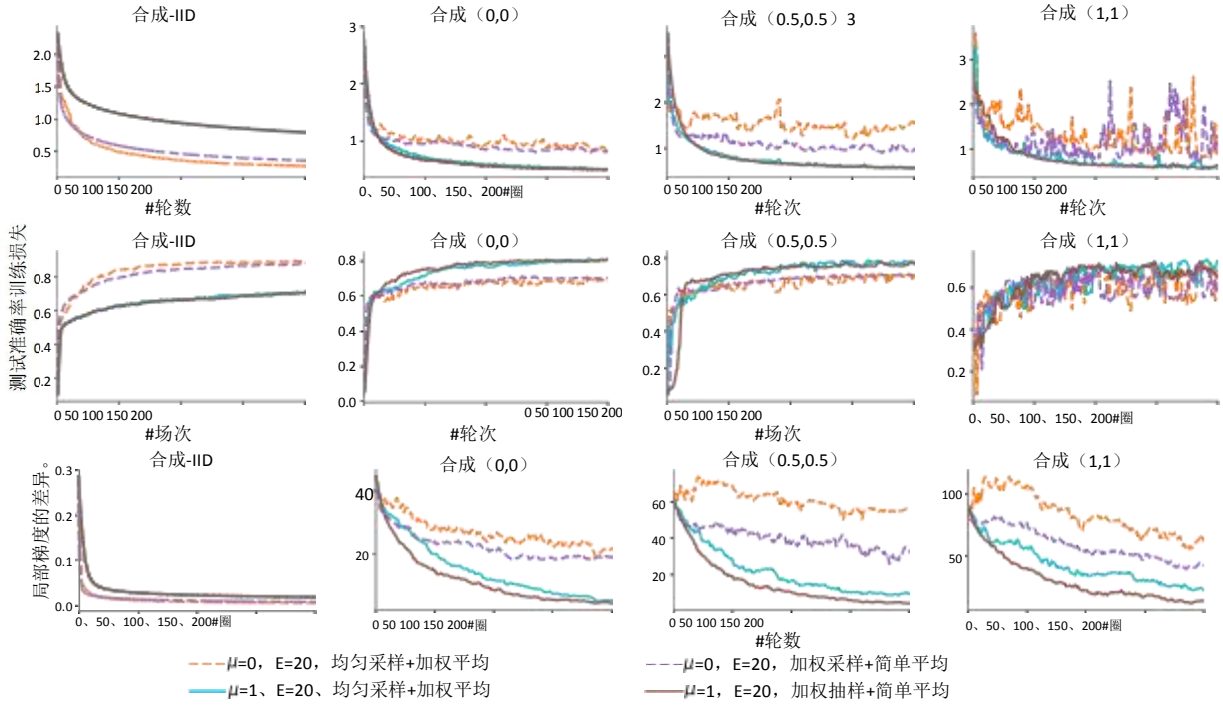

图12.两种采样方案在训练损失、测试准确率和不相似性度量方面的差异。采用概率与局部数据点数量成正比的采样设备，然后简单地平均局部模型的表现略优于均匀采样设备并按局部数据点数量成比例加权平均局部模型的情况。在任一采样方案下，μ=1的设置比withμ= 0的设置表现出更稳定的性能。