使用近端随机方差减少梯度算法的联邦学习

会议论文·2020年8月
DOI: 10.1145/3404397.3404457

6位作者，包括:

Canh T.Dinh
悉尼大学
17篇出版物410次引用
SEE PROFILE

Nguyen H. Tran
悉尼大学
271篇出版物8,965次引用
SEE PROFILE

Tuan Dung Nguyen
澳大利亚国立大学
6篇论文34次引用
SEE PROFILE

魏保
南洋理工大学
93篇论文，6866次引用
SEE PROFILE

Federated Learning with Proximal Stochastic Variance Reduced Gradient算法

Canh T. Dinh
悉尼大学
澳大利亚NSW
tdin6081@uni.sydney.edu.au

guyen H. Tran
悉尼大学(University of Sydney)
澳大利亚NSW
nguyen.tran@sydney.edu.au

edu.auTuan阮勇*
墨尔本大学(University of Melbourne)
VIC，澳大利亚tuandung
n@unimelb.edu.au

魏保
悉尼大学(University of Sydney)
澳大利亚NSW wei.
bao@sydney.edu.auN

Albert Y. Zomaya
悉尼大学(University of Sydney)
澳大利亚NSW albert.zomay
a@sydney.

周冰b
悉尼大学
澳大利亚NSW bing.
zhou@sydney.edu.au

摘要

联邦学习(FL)是一种快速发展的分布式机器学习技术，涉及大量用户设备的参与。虽然FL具有数据隐私和丰富的用户生成数据的优点，但其在用户数据和设备之间的异构性挑战使算法设计和收敛分析复杂化。为了应对这些挑战，我们提出了一种利用非凸FL的近端随机方差减少梯度方法的算法。该算法由两个嵌套循环组成，允许用户设备在将这些本地模型发送到服务器进行全局模型更新(外循环)之前，将其本地模型更新到大约一个精度阈值(内循环)。我们描述了局部和全局模型更新的收敛条件，并通过算法的参数控制从这些条件中提取了各种见解。我们还提出了如何优化这些参数，使FL的训练时间最小化。实验结果不仅验证了理论收敛性，而且表明该算法在FL环境下的收敛速度优于现有的基于随机梯度下降的方法。

关键字

分布式机器学习，联邦学习，随机梯度下降

ACM参考格式:

丁灿彤，陈国强，陈国强，陈国强，周炳斌，包伟。2020。基于近端随机方差减少梯度算法的联邦学习。第49届并行处理国际会议- ICPP (ICPP ' 20)， 2020年8月17-20日，埃德蒙顿，AB,加拿大。ACM，美国纽约，11页。https://doi.org/10.1145/3404397.3404457

1 介绍

大规模分布式机器学习，主要是在数据中心设置中，已经极大地吸引了分布式优化算法的研究兴趣，以便在大数据规模上快速有效地训练深度学习模型[5,6,10,19,26,28]。在许多应用中，为了加快计算速度，大规模数据集被分布在多台机器上进行并行处理。这些研究中的大多数学习算法都是为具有平衡、独立和同分布(i.i.d)数据的机器设计的。

然而，维护客户数据的隐私要求从数据中心的计算到包括移动电话和传感器在内的许多设备上的集体计算进行根本性的转变，其中数据在本地存储和处理。现代智能设备可以被认为是一台具有强大处理器的体面计算机(例如，Snapdragon 835上带有高通Hexagon矢量扩展的Hexagon DSP[1])，以及用于收集大量数据的众多传感器(例如，摄像头，麦克风和GPS)，这使得本地训练可行。大多数用户的数据都是隐私敏感的，本质上很大，所以将数据记录到数据中心进行模型训练是有风险和密集的，这就需要这种边缘学习的发展。隐私保护学习技术的一个例子是最近提出的联邦学习(FL)范式[20]。这种学习技术允许用户设备通过仅将本地模型而不是大量原始数据发送到中央服务器来协作构建全局训练模型，从而不仅保护了设备上的数据隐私，还节省了通信带宽。由于本地数据源是自然分布的，FL也带来了设计学习算法的新挑战，考虑到设备数据可能不相同分布的统计异质性。

机器学习中事实上的优化算法，如梯度下降(GD)、随机梯度下降(SGD)和方差缩减SGD(例如SVRG[11])已被广泛应用于FL中实现设备的本地更新[12,20,31]。FL的先驱之一[20]提出了使用平均SGD本地更新的fedag，经验表明，该方法在非凸联邦设置中表现良好。作者在[12]中提出了FSVRG来经经验地提高fedag的性能

*Work done at The University of Sydney, Australia.

Nguyen H. Tran, Wei Bao, and Bing B. Zhou were supported by the Australian Research Council Discovery Project grant DP200103718. Albert Y. Zomaya and Wei Bao were supported by the Australian Research Council Discovery Project grant DP190103710.

ICPP ' 20, August 17–20, 2020, Edmonton, AB, Canada
 ACM ISBN 978-1-4503-8816-0/20/08...$15.00 https://doi.org/10.1145/3404397.3404457

SVRG。但这两部作品缺乏理论收敛分析。[30]中的作者使用SGD对他们提出的类似于fedag的方案进行了收敛分析，但对所有设备的数据进行了宽松的i.i.d假设。虽然[31]中提出的算法使用了收敛性可证明的GD，但他们的分析仅限于凸损失函数，这可能不适用于最近非常成功的深度学习应用，这些应用通常是非凸的[14]。此外，当训练样本数量很大时，与SGD或方差减少SGD(两者都适用于随机抽样的数据子集)相比，GD(适用于所有数据样本)更加耗时和计算密集。考虑到典型的智能设备电池容量和计算能力有限，为了FL，使用SGD及其变体进行更快的训练比GD更可行。另一方面，[16]提出了一个框架FedProx，通过近似解决用户局部问题来保证收敛性。

虽然SVRG[11]和SARAH[22]等方差减少SGD方法已被证明具有比SGD更快的收敛性，但它的代价是在一段迭代后进行全批梯度计算。值得注意的是，周期性全批评估策略完全符合FL固有的全局聚合性质(在每个设备上进行多次局部更新后)。因此，观察到文献中缺乏使用方差减少SGD具有可证明收敛性的完整非凸FL设计，我们通过以下贡献弥补了这一差距:

- 我们提出了一种基于近端随机方差减少梯度的算法，名为FedProxVR，它由两个嵌套的全局迭代和局部迭代循环组成。当服务器在每次全局迭代时执行全局模型更新时，设备使用近端SVRG或SARAH更新局部训练模型，直到精度阈值，这会影响局部迭代的次数。
- 我们首先为局部模型收敛到精度阈值提供了一个充分条件。这个条件反映了学习步长、局部迭代次数和精度阈值之间的关系。具体来说，我们定性地表明，通过控制参数化的学习步长，当局部迭代次数被限定在由步长参数和局部精度阈值决定的范围内时，局部模型更新的收敛性得到了保证。
- 我们接下来为全局模型更新提供了一个充分条件，它也捕获了设备和服务器之间的通信复杂性。我们证明了全局收敛是由两个重要的控制旋钮决定的:局部精度阈值和近端惩罚参数。这两个参数也表征了局部和全局模型收敛之间的权衡，这为算法设计提供了后见之明。
- 然后，我们提出了基于相应控制变量最小化总训练时间(分别是用户设备和服务器上的局部和全局模型更新延迟)的方法。我们还证明了该最小化问题的数值解与FedProxVR的收敛分析的理论结果相匹配。
- 我们最终通过使用Tensorflow的各种真实和合成数据集呈现经验收敛来验证理论发现。实验结果表明，与基于sgd的FL方法相比，FedProxVR可以提高收敛速度。

接下来，我们将分别在第2节和第3节中介绍相关工作和系统模型。FedProxVR的设计和分析将在第4节中展示。最后我们将在第5节中报告实验结果。所有的技术证明将在附录中提供。

2 相关工作

最近，人们对一种新的机器学习技术越来越感兴趣，这种技术利用了许多用户设备的参与，称为联邦学习[13,16,20,27,31]，其中每个设备生成自己的数据，因此跨设备的数据具有统计异质性。特别是，一些作品提出了边缘和/或无线网络背景下的FL算法[2,29,31]。

已经有一些尝试为非凸问题设计算法来解决FL的挑战，例如用户设备和数据的异构性。第一种方法侧重于在每个设备上为固定数量的本地迭代运行事实上的算法(SGD/SVRG)[12,20]。这种方法是实用的，因为它使用了机器学习中现有的核心优化包(例如SGD)。然而，这些研究大多缺乏收敛性分析，这主要是由于FL的异构性带来的挑战。[31]中的研究通过使用额外的假设，如凸性和L-Lipschitz函数，提供了收敛性分析，但核心算法是基于GD的，其计算复杂度相对于数据样本数量呈线性扩展。第二种方法允许设备近似地解决其原始问题，直至达到局部精度阈值[16,24]。近似地解决局部模型更新可以给出一个灵活的计算-通信权衡，即设备是否应该在其局部模型更新上运行更多的回合，还是与服务器进行更多的通信以进行全局模型更新。在这个方向上，尽管[27]的作者展示了他们的算法的收敛性保证，但他们采用的原对偶优化技术只适用于凸任务学习。我们观察到，近似方法的局部精度阈值与局部迭代次数之间存在联系，这在之前没有得到解决。我们将通过使用固定的参数化步长来展示这种联系，这对于分析FL的通信复杂性和算法设计至关重要。

在非fl环境下，为了构建比SGD更复杂、更精确的梯度估计器，人们提出了各种方法，如SVRG[11]和SARAH[22]，以减少SGD中梯度估计器的方差。基于这些众所周知的方差缩减技术，已经提出了诸如ProxSVRG[9,17]和ProxSARAH[23]之类的近端方法来处理非凸、非光滑问题。

3 系统模型

我们考虑一个由一个聚合服务器和N个设备的集合N组成的系统。每个设备n存储一个本地数据集Dn，用

其大小用Dn表示。然后，我们可以用D= ÍNn=1 Dn来定义总数据大小。数据可以通过设备的使用来生成，例如，通过与移动应用程序的交互。

设备n的损失函数定义为

$$F_n(w) := \frac{1}{D_n}\sum_{i\in\mathcal{D}_n} f_i(w). \tag{1}$$

那么，学习模型就是下面全局损失函数最小化问题的最小化器

$$\min_{w\in\mathbb{R}^d} \bar{F}(w) := \sum_{n=1}^{N} \frac{D_n}{D} F_n(w). \tag{2}$$

假设1。fi(·)是l -光滑的，Fn(·)分别是(−λ)-强凸和σn -散度，∀n (λ， σn > 0)，如下所示

$$\left\|\nabla f_i(w) - \nabla f_i(w')\right\| \le L\left\|w - w'\right\| \tag{3}$$

$$F_n(w) + \left\langle \nabla F_n(w), w' - w\right\rangle \le F_n(w') + \frac{\lambda}{2}\left\|w - w'\right\|^2 \tag{4}$$

$$\left\|\nabla F_n(w) - \nabla \bar{F}(w)\right\| \le \sigma_n\left\|\nabla \bar{F}(w)\right\|. \tag{5}$$

这里的⊠w,w '⊠是向量w和w '的内积。所有的范数都是欧几里德范数。虽然l -光滑假设在文献中很常见[16,31]，但强凸性参数−λ不仅允许Fn是非凸的(特别是对于深度学习应用)，而且还限制了适用于近端方法的非凸程度[4,16]。这将我们的分析与具有凸性假设的传统方法区分开来。σn -散度假设(5)表征了设备间数据的异质性，其中较大的σn表示设备n与其他设备之间的不相似性。与[31]中的梯度散度假设(基本假设∇F n (x)−∇F(x)¯≤σn)不同，如果Fn()是凸二次函数[3]，则σn -散度允许在任意方向上二次增长。这个假设(5)也包括了[16]中的b -局部不相似度作为特例。虽然我们只通过(5)中的σn来指定器件的数据异质性，但为了简化符号，我们对所有n使用相同的L和λ。但是我们注意到，即使我们允许Ln和λn的异质性值，其中L， λ可以用引理1中的Ln， λn和定理1中的L，¯λ¯来代替，这项工作中的所有结果都是不变的。我们还使用符号σ¯2:= ÍN Dn n= 1 D σn2，类似于其他情况。

4 算法设计与分析

在本节中，我们介绍算法，提供其收敛性分析，并优化算法参数。

4.1 算法设计

在本节中，我们提出了一个使用近端随机方差减少梯度算法的FL框架，称为FedProxVR，作为

Algorithm 1 FedProxVR

1: **input:** $\bar{w}_0$, $\eta = \frac{1}{\beta L}$.
2: **for** $s = 1, \ldots, T$ **do** *Global iterations*
3: **for** $n = 1, \ldots, N$ **do in parallel**
4: $w_{n,s}^{(0)} = \bar{w}^{(s-1)}$
$v_{n,s}^{(0)} = \nabla F_n(w_{n,s}^{(0)})$
$w_{n,s}^{(1)} = \mathrm{prox}_{\eta h_s}(w_{n,s}^{(0)} - \eta v_{n,s}^{(0)})$,
5: **for** $t = 1, \ldots, \tau$ **do** *Local iterations by devices*
6: Uniformly randomly pick $(x_{i_t}, y_{i_t}) \in \mathcal{D}_n$.
7: Update $v_{n,s}^{(t)}$ according to (8a) or (8b)
8: $w_{n,s}^{(t+1)} = \mathrm{prox}_{\eta h_s}(w_{n,s}^{(t)} - \eta v_{n,s}^{(t)})$,
9: **end for**
10: Set $w_n^{(s)} = w_{n,s}^{(t')}$ where t' is chosen uniformly at random from $\{0, \ldots, \tau\}$
11: **end for**
12: $\bar{w}^{(s)} = \sum_{n=1}^{N} \frac{D_n}{D} w_n^{(s)}$ *Global model update by server*
13: **end for**
14: **output:** $\bar{w}^{(T)}$

如图1所示。为了解决问题(2)，FedProxVR需要全局迭代数w¯(s)来更新全局模型，通过聚合所有设备的所有本地模型w (s)。

局部模型更新:为了获得局部模型w¯(s)，每个设备n将求解其下面的代理函数(第3至10行)

$$\min_{w\in\mathbb{R}^d}\left\{J_n(w) := F_n(w) + h_s(w)\right\}, \tag{6}$$

$$\text{where } h_s(w) := \frac{\mu}{2}\left\|w - \bar{w}^{(s-1)}\right\|^2, \tag{7}$$

可以将正则化的hs视为“软”共识约束，以惩罚与当前全局模型w¯(s−1)的任何局部偏差。弹性平均SGD和FedProx在[34]和[16]中也分别使用了该函数。由于hs(.)是μ~强凸，我们可以通过选择μ满足μ~:=μ−λ> 0来允许Jn (w)是μ~强凸。FedProxVR使用近端更新规则(第8行)与使用(8a)或(8b)的方差减少随机梯度估计器解决(6)

$$v_{n,s}^{(t)} = \begin{cases} \nabla f_{i_t}(w_{n,s}^{(t)}) - \nabla f_{i_t}(w_{n,s}^{(t-1)}) + v_{n,s}^{(t-1)}, \text{ (SARAH)} & \text{(8a)} \\ \nabla f_{i_t}(w_{n,s}^{(t)}) - \nabla f_{i_t}(w_{n,s}^{(0)}) + v_{n,s}^{(0)}. \text{ (SVRG)} & \text{(8b)} \end{cases}$$

如果v(t)n,s =∇ fi(t)(w(t)n,s)我们就得到了SGD。SARAH[22]和SVRG[11]都使用外部循环进行一次全梯度评估和“锚定”模型更新(第2行至第4行)。然而，在内部循环(第5行至第9行)中，SARAH与SVRG不同的是，它基于先前局部迭代(t−1)和当前SGD的随机分量递归地更新随机方向vn,s (t)。proximal算子的定义如下

$$\mathrm{prox}_{\eta h_s}(x) := \arg\min_{w\in\mathbb{R}^d}\left(h_s(w) + \frac{1}{2\eta}\left\|w - x\right\|^2\right) \tag{9}$$

$$= \frac{\eta}{1+\eta\mu}\left(\mu\,\bar{w}^{(s-1)} + \frac{1}{\eta}x\right). \tag{10}$$

局部问题(6)(在给定s处)的收敛准则定义如下

$$\mathbb{E}\left[\|\nabla J_n(w_n^{(s)})\| \mid \tilde{w}^{(s-1)}\right] \le \theta\left\|\nabla F_n(\tilde{w}^{(s-1)})\right\|, \tag{11}$$

它由局部精度θ∈(0,1)参数化，因此由局部迭代总数τ参数化。这个局部精度概念类似于[28]和[16,24]中的近似因子和不精确因子。这里θ= 0表示局部问题(6)需要最优解，θ= 1表示局部问题没有进展，即通过设置τ= 0。由于所有设备都具有相同的τ，因此局部模型更新是同步的。E[·]是对FedProxVR中所有随机性的期望。

全局模型更新:在接收到设备发送的所有本地模型后，服务器将根据第12行更新全局模型，并将其反馈给所有设备进行下一次全局迭代更新。我们还使用梯度的期望平方范数作为非凸问题的收敛指标(即平稳间隙)[3]，全局问题(2)实现⊠-accurate解如果

$$\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\left\|\nabla\bar{F}(\tilde{w}^{(s)})\right\|^2 \le \epsilon. \tag{12}$$

4.2 FedProxVR的收敛性分析

在FedProxVR中，我们选择一个固定的步长η1，用β参数化η= β 1l。局部模型更新的收敛性如下所示。

引理1。如果β和τ满足以下条件，设备n获得θ-精确解(11)

a)当使用SARAH更新(8a)时:

$$0 \le \frac{3(\beta^2L^2+\mu^2)}{\theta^2\tilde{\mu}L(\beta-3)} \le \tau \le \frac{5\beta^2-4\beta}{8} \tag{13}$$

b)当使用SVRG更新(8b)时:

$$0 \le \frac{3(\beta^2L^2+\mu^2)}{\theta^2\tilde{\mu}L(\beta-3)} \le \tau \le \frac{5\beta^2-4\beta}{8\,a} - 2 \tag{14}$$

当存在a > 0使得a−4⩾4pa(τ+ 1)。

下面的注释是关于局部精度θ、局部迭代次数τ和步长参数β之间的关系。

备注1。

(1)对于任意θ∈(0,1)，我们总是可以选择一个足够大的β来满足(13)和(14)，其中τ的下界和上界分别为Ω(β)和O(β2)。这意味着在足够小的(固定的)步长η下，可以保证局部收敛。

(2)我们看到τ= Ω(θ - 1-2)。因此，如果θ较小，τ必须较大才能满足下界条件。很简单，θ值越小，我们得到(6)的解就越接近最优，这就需要运行更多的局部迭代。

(3)在实践中，由于大步长η和快速收敛(小τ)是首选，我们通过求解(例如，在SARAH的情况下)来选择满足引理1条件的最小βmin。

$$\frac{3(\beta^2L^2+\mu^2)}{\theta^2\tilde{\mu}L(\beta-3)} = \frac{5\beta^2-4\beta}{8}, \beta > 3, \tag{15}$$

并相应得到(最小)τ:

$$\tau = \frac{5\beta_{min}^2 - 4\beta_{min}}{8}. \tag{16}$$

(4)观察到τ的下界为Ω(μ)，从而增大μ(例如;(例如，当λ较大时，使μ≈⩾0)将增加τ。这是因为较大的μ将使局部更新更接近于每个s中的“锚定”点w¯(s−1)，从而使收敛到θ-精确解的速度更慢。

(5)与SARAH相比，SVRG具有更严格的上界条件(由于a⩾4)。因此，SVRG需要更大的βmin来满足条件(14)，因此需要更大的τ(由于下界)。这可以解释为SARAH使用了比SVRG更稳定的随机梯度估计，这在[22]中也得到了样本数据集的验证。我们注意到，SARAH和SVRG之间的具体理论比较之前还没有进行过探讨。

hDefining将任意点w¯(0)的costi间隙通过Δ(w¯(0)):= E F(¯w¯(0))− F(¯w¯*)，给出FedProxVR全局模型更新的收敛条件。

定理1。考虑所有设备满足引理1条件的FedProxVR，我们有

$$\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\left\|\nabla\bar{F}(\tilde{w}^{(s)})\right\|^2 \le \frac{\Delta(\bar{w}^{(0)})}{\Theta\,T}. \tag{17}$$

在哪里

$$\Theta = \frac{1}{\mu}\left(1-\theta\sqrt{2(1+\bar{\sigma}^2)} - \frac{2L}{\tilde{\mu}}\sqrt{(1+\theta^2)(1+\bar{\sigma}^2)} - \frac{2L\mu}{\tilde{\mu}^2}(1+\theta^2)(1+\bar{\sigma}^2)\right) > 0.$$

推论1。实现(2)⊠-accurate解所需的全局迭代次数为

$$T \ge \frac{\Delta(\bar{w}^{(0)})}{\Theta\,\epsilon}, \tag{18}$$

备注2。

(1)我们看到θ和μ是FedProxVR收敛的重要控制“旋钮”。具体来说，为了使Θ> 0，我们必须选择足够大的μ和Θ < (2 (1 + σ¯2))−1/2，这显示了数据异质性如何影响局部和全局收敛。具体来说，较大的σ¯2，即较小的θ，意味着设备将运行更多的局部迭代。

(2)这两个参数也表征了局部收敛和全局收敛之间的权衡。虽然全局收敛要求θ足够小，但为了更快的局部收敛，设备更喜欢较大的θ(见备注1)。另一方面，虽然μ必须足够大以确保全局收敛，但它不应该太大，以免对局部收敛(即使τ变大)和全局收敛(即使Θsmall变大，因此T变大)产生负面影响。

(3)为了使Θ> 0，大的L和λ需要大的μ。

(4)与传统近端SVRG[17]或SARAH[32](具有非凸和固定步长但没有FL设置)的O(⊠1)迭代相比，我们看到T = O(Θ1⊠)的FedProxVR被联邦因子Θ缩放。接下来，我们将

1使用固定的步长比减小步长更实用[3]。

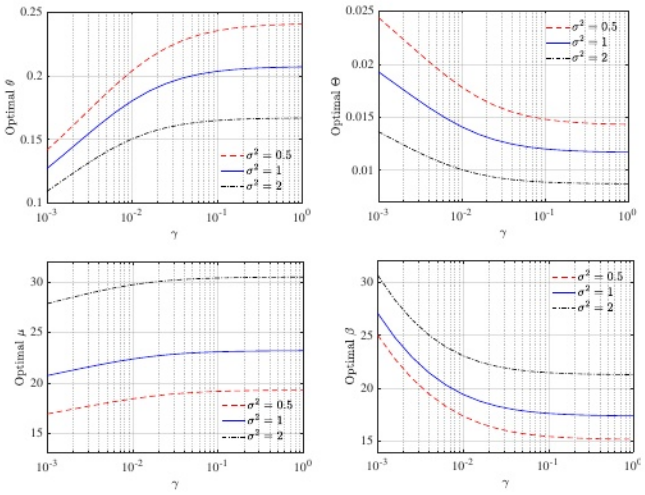

图1:权重因子γ对解的影响

问题(23)，其中L = 1， λ= 0.5(这两个值可以通过抽样真实数据集来估计。)

优化FedProxVR的参数，包括联邦因子。

4.3 FedProxVR参数优化

表示设备通过dcmp和dcom向服务器发送本地模型更新的计算(即算法1中的第7步和第8步)和通信延迟，则FedProxVR的总训练时间为

$$\mathcal{T} := T(d_{com} + d_{cmp}\tau). \tag{19}$$

定义一个权重因子γ:= ddcmp和T = Δ(w¯(0))

Θ，我们最小化com

T以收敛条件为约束:

$$\underset{\mu,\theta,\beta,\tau}{\text{minimize}} \quad \frac{1}{\Theta}\left(1+\gamma\,\tau\right) \tag{20}$$

$$\text{subject to} \quad (15), (16), \text{ and } \Theta > 0. \tag{21}$$

通过去掉约束(15)、(16)，用SARAH代入

$$\theta^2 = \frac{24(\beta^2 L^2 + \mu^2)}{\tilde{\mu} L(5\beta^2 - 4\beta)(\beta - 3)} \tag{22}$$

代入Θ，我们进一步将这个优化问题简化为

$$\underset{\mu,\beta}{\text{minimize}} \quad \frac{1}{\Theta}\left(1+\gamma\,\frac{5\beta^2-4\beta}{8}\right) \tag{23}$$

$$\text{subject to} \quad \beta > 3 \text{ and } \Theta > 0, \tag{24}$$

它比原始形式(20)拥有更少的变量和约束。问题(23)不幸是非凸的。然而，由于只有两个变量需要优化，我们可以采用数值方法来寻找全局最优解。我们在图1中数值说明了权重因子γ对最优参数的影响。当γ非常小时，这意味着通信延迟比局部计算延迟昂贵得多，我们看到最优β(因此τ)非常大，即设备具有更多的局部计算比通信轮数更好。当γ增大时，虽然最优β减小，使得局部模型更新可以用较小的τ近似求解，但由于θ的相应增大值，最优μ增大以确保Ω> 0。我们还观察到，较大的σ¯2增加了最优μ和β，但减少了Θ和Θ。图1中的所有数值观测都与引理1和定理1的理论注释完全匹配。

5 实验

在本节中，我们将通过现实世界的实验，比较FedProxVR与基于sgd的fedag[20]的有效性。我们还展示了FedProxVR的经验收敛如何通过改变其控制超参数与其理论结果相关。所有代码和数据都准备在GitHub上发布[7]。

实验设置:为了评估Fed-ProxVR在各种任务和学习模型上的性能，我们将在实验中使用不同类型的数据集。除了在[16,26]中捕获统计异质性的“合成”数据集外，我们还考虑使用凸和非凸模型进行图像分类任务的真实数据集，如“MNIST”[15]和“FASHION-MNIST”[33]。所有数据集被随机分割，75%用于训练，25%用于测试。

为了生成模拟FL异构特性的设备的数据集，我们为凸模型(多项逻辑回归)模拟了100个设备，为非凸卷积神经网络(CNN)模型模拟了10个设备(因为运行100个设备的CNN需要花费更长的时间);每个设备都有不同的样本量，根据幂律生成，如[16]所示。此外，每个设备在10个标签上只包含两个不同的标签。对于“Synthetic”、“MNIST”和“FASHION-MNIST”，每个设备的数据样本数量在[37,3277]、[454,3939]和[37,1350]的范围内。我们使用Tensorflow框架实现FedProxVR和基于sgd的fedag。

为了进行公平的比较，所有算法在实验过程中使用相同的参数β，τ， N,T。在最后的实验中，使用每个算法的最优超参数进行性能比较。我们将使用多项逻辑回归模型对凸任务进行图像分类，使用两层CNN模型对非凸任务进行分类。对于CNN模型，我们采用与[20]类似的结构，两个5x5的卷积层(第一层和第二层分别为32和64通道，每层后使用最大池大小2x2)，ReLu激活，并在CNN的末端使用softmax层。虽然在Alg. 1中没有提到mini-batch，但实验使用mini-batch来减少在大量数据点的情况下寻找最优局部点的解决挑战。

步长参数β和局部迭代τ的影响:我们首先比较了图2和3中FedProxVR和fedag在不同超参数设置下的收敛性。在这两个图中，我们首先选择β的值，然后根据其在引理1中的上界确定τ，从而使算法经验收敛。虽然τ的上界仅取决于β，但它的下界由L和μ ~等参数决定，这些参数更难从数据集和学习任务中估计出来。我们从较小的β和τ值开始，然后逐渐增大，观察FedProxVR的收敛行为以及权向量γ对最优参数β和τ的影响。

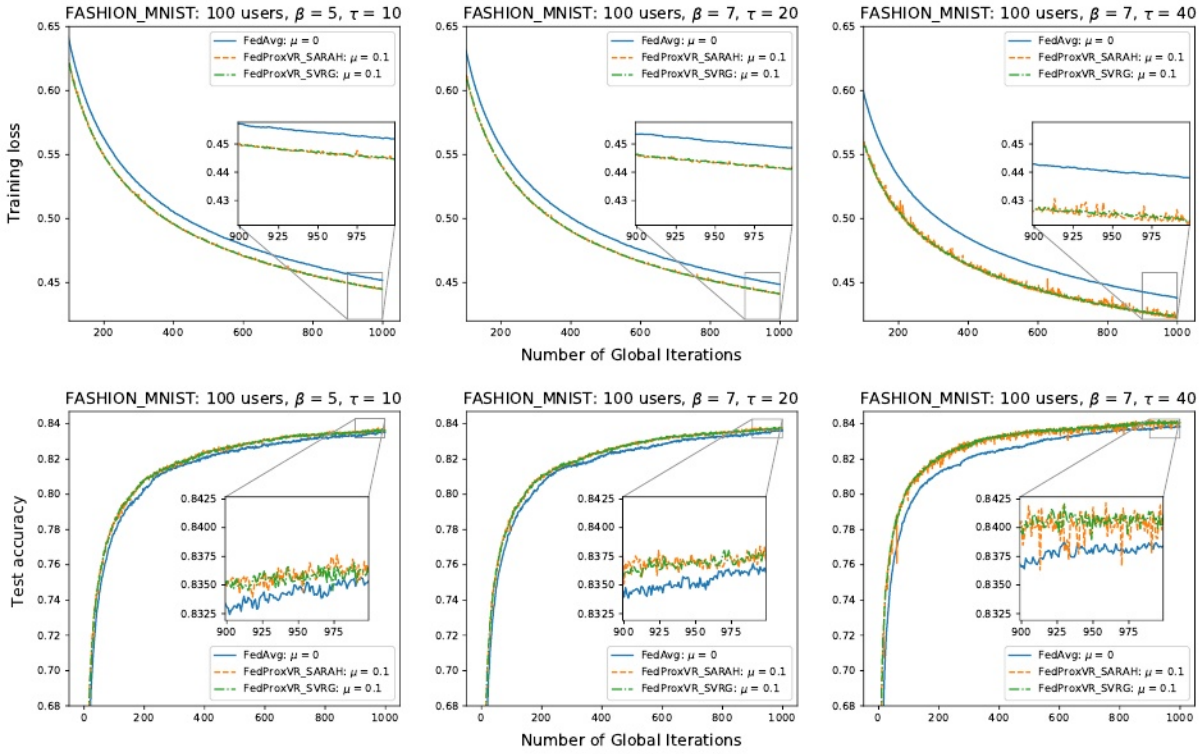

图2:FedProxVR与凸任务在Fashion-MNIST数据集上的收敛性。所有算法都使用大小为B = 32的批处理。

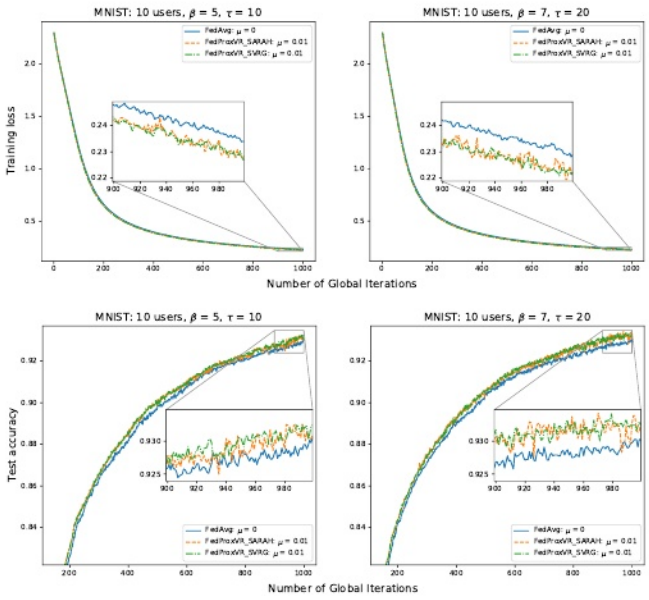

图3:FedProxVR与非凸任务(CNN)在MNIST数据集(B = 64)上的收敛性。

在图2中，我们观察到β和τ对所有算法在同一凸任务上的收敛性的相似影响。首先,当

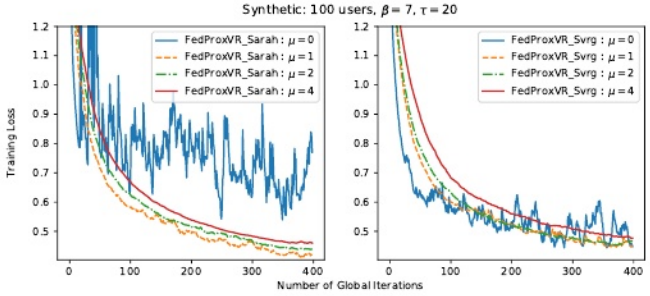

图4:具有凸任务的近端惩罚μ到FedProxVR收敛对合成数据集的影响。

β和τ更小(分别为5和10)，相当于更大的学习率η和γ，FedProxVR的性能比fedag有显著提高。其次，当β和τ分别增加到7和20时，我们不仅可以看到性能差距的增加，而且所有算法的训练损失和测试精度也收敛得更快。这验证了第4.3节中的理论发现，在设备-服务器通信延迟是更昂贵的本地通信延迟(小γ)的场景中，将τ和β的最优值增加到一定值将降低通信成本，同时保证算法的收敛性。最后，如果τ的值被设置为高于由β指定的上界(对于FedProxVR使用

表1:在凸任务上使用最佳超参数比较模型的测试精度。

Algorithm	τ	β	μ	B	T	Accuracy
FedAvg	10	10	0	16	983	**84.02%**
FedProxVR (SVRG)	20	10	0.1	32	895	**84.12%**
FedProxVR (SARAH)	20	5	0.1	32	965	**84.21%**

表2:在非凸任务上使用最佳超参数比较模型的测试精度。

Algorithm	τ	β	μ	B	T	Accuracy
FedAvg	20	10	0	16	995	**93.52%**
FedProxVR (SVRG)	20	10	0.01	16	970	**94.06%**
FedProxVR (SARAH)	20	9	0.01	32	958	**93.75%**

SVRG和SARAH)和(对于使用SVRG的FedProxVR)，从而违反引理1,FedProxVR的学习曲线波动更为明显，尽管FedProxVR和fedag的性能仍然有所提高和区分。因此，当选择τ满足其下界和上界条件时，FedProxVR有望比fedag收敛得更好。

FedProxVR和fedag在非凸任务上的性能如图3所示。在这里，我们在凸设置中观察到与我们的实验相似的结果，并且FedProxVR和fedag之间的性能差距略大。

近端惩罚μ对全局迭代的影响T:我们在图4中评估了近端惩罚μ对FedProxVR收敛的影响。在合成数据集上使用FedProxVR，我们观察到FedProxVR的训练损失在μ= 0时发散，增加μ> 0使损失趋于稳定，使其收敛。然而，同样值得注意的是，较大的μ值将使FedProxVR的收敛速度变慢。因此，μ也反映了学习曲线的平滑性和FedProxVR的收敛速度之间的权衡。

使用优化参数的性能比较:由于算法在相同的超参数上表现不同(例如，在我们的实验中，μ，τ和η)，我们对精心选择的超参数范围进行随机搜索，以确定它们的哪种组合相对于每种算法会产生最高的测试精度。结果如表1和表2所示。可以看出，当使用他们优化的超参数时，FedProxVR在凸和非凸任务上都设法提高了fedag的准确性。此外，虽然fedag在凸任务上较小的批处理规模上表现更好，但FedProxVR从较大的批处理规模中受益。最后，在这两个任务上，FedProxVR比fedag更早开始收敛。

6 结论

在本文中，我们提出了一种使用近端随机方差减少梯度方法的FL算法，该算法可以解决由于具有非相同分布数据源的大规模参与设备而导致的FL异构挑战。在提出的算法中，允许每个用户设备独立地近似地解决其学习问题，进行多次局部迭代以进行局部模型更新，这些迭代将被发送到服务器以进行全局模型更新。我们对局部和全局模型更新的收敛分析进行了表征，这为算法设计提供了一些富有成效的见解。我们还提出了如何找到最优的算法参数来最小化FL训练时间。使用Tensorflow，我们通过在各种真实和合成数据集上展示所提出算法的经验收敛性来验证理论发现，表明与基于sgd的FL方法相比，我们的算法可以提高收敛速度。

参考文献

[1] 2017. We Are Making On-Device AI Ubiquitous. https://www.qualcomm.com/news/onq/2017/08/16/we-are-making-device-ai-ubiquitous.

[2] Mohammad Mohammadi Amiri and Deniz Gunduz. 2019. Machine Learning at the Wireless Edge: Distributed Stochastic Gradient Descent Over-the-Air. arXiv:1901.00844 [cs, math] (Jan. 2019). arXiv:1901.00844 [cs, math]

[3] L. Bottou, F. Curtis, and J. Nocedal. 2018. Optimization Methods for Large-Scale Machine Learning. SIAM Rev. 60, 2 (Jan. 2018), 223–311. https://doi.org/10.1137/16M1080173

[4] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. 2018. Accelerated Methods for NonConvex Optimization. SIAM Journal on Optimization 28, 2 (Jan. 2018), 1751–1772. https://doi.org/10.1137/17M1114296

[5] Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc’Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, and Andrew Y. Ng. 2012. Large Scale Distributed Deep Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems -Volume 1 (NIPS’12). Curran Associates Inc., Lake Tahoe, Nevada, 1223–1231.

[6] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. 2012. Optimal Distributed Online Prediction Using Mini-Batches. J. Mach. Learn. Res. 13 (Jan. 2012), 165–202.

[7] Charlie Dinh. 2020. CharlieDinh/FederatedLearningWithSVRG. https://github. com/CharlieDinh/FederatedLearningWithSVRG

[8] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. 2016. Mini-Batch Stochastic Approximation Methods for Nonconvex Stochastic Composite Optimization. Mathematical Programming 155, 1-2 (Jan. 2016), 267–305. https://doi.org/10. 1007/s10107-014-0846-1

[9] Sashank J. Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. 2016. Proximal Stochastic Methods for Nonsmooth Nonconvex Finite-Sum Optimization. In Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 1145–1153.

[10] Martin Jaggi, Virginia Smith, Martin Taká, Jonathan Terhorst, Sanjay Krishnan, Thomas Hofmann, and Michael I. Jordan. 2014. Communication-Efficient Distributed Dual Coordinate Ascent. arXiv:1409.1458 [cs, math, stat] (Sept. 2014). arXiv:1409.1458 [cs, math, stat]

[11] Rie Johnson and Tong Zhang. 2013. Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1 (NIPS’13). Curran Associates Inc., Lake Tahoe, Nevada, 315–323.

[12] Jakub Kone ný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. arXiv:1610.02527 [cs] (Oct. 2016). arXiv:1610.02527 [cs]

[13] Jakub Kone ný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated Learning: Strategies for Improving Communication Efficiency. http://arxiv.org/abs/1610.05492 (Oct. 2016). arXiv:1610.05492

[14] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. Nature 521, 7553 (May 2015), 436–444. https://doi.org/10.1038/nature14539

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-Based Learning Applied to Document Recognition. Proc. IEEE 86, 11 (Nov. 1998), 2278–2324. https://doi.org/10.1109/5.726791

[16] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2019. Federated Optimization for Heterogeneous Networks. In Proceedings of the 1st Adaptive & Multitask Learning, ICML Workshop, 2019. Long Beach, CA, 16.

[17] Zhize Li and Jian Li. 2018. A Simple Proximal Stochastic Gradient Method for Nonsmooth Nonconvex Optimization. In Proceedings of the 32Nd International Conference on Neural Information Processing Systems (NIPS’18). Curran Associates Inc., Montréal, Canada, 5569–5579.

[18] Zhize Li and Jian Li. 2018. A Simple Proximal Stochastic Gradient Method for Nonsmooth Nonconvex Optimization. arXiv:1802.04477 [cs, math, stat] (Feb. 2018). arXiv:1802.04477 [cs, math, stat]

[19] Chenxin Ma, Jakub Kone ný, Martin Jaggi, Virginia Smith, Michael I. Jordan, Peter Richtárik, and Martin Taká. 2017. Distributed Optimization with Arbitrary

Local Solvers. Optimization Methods and Software 32, 4 (July 2017), 813–848. https://doi.org/10.1080/10556788.2016.1278445

[20] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Net-works from Decentralized Data. In Artificial Intelligence and Statistics. 1273–1282.

[21] Yurii Nesterov. 2018. Lectures on Convex Optimization. Vol. 137. Springer International Publishing. https://doi.org/10.1007/978-3-319-91578-4

[22] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Taká. 2017. SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient. In Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70), Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 2613– 2621.

[23] Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. [n. d.]. ProxSARAH: An Efficient Algorithmic Framework for Stochastic Composite Nonconvex Optimization. ([n. d.]), 45.

[24] Sashank J. Reddi, Jakub Kone ný, Peter Richtárik, Barnabás Póczós, and Alex Smola. 2016. AIDE: Fast and Communication Efficient Distributed Optimization. arXiv:1608.06879 [cs, math, stat] (Aug. 2016). arXiv:1608.06879 [cs, math, stat]

[25] Sashank J. Reddi, Suvrit Sra, Barnabas Poczos, and Alex Smola. 2016. Fast Sto-chastic Methods for Nonsmooth Nonconvex Optimization. arXiv:1605.06900 [cs, math, stat] (May 2016). arXiv:1605.06900 [cs, math, stat]

[26] Ohad Shamir, Nathan Srebro, and Tong Zhang. 2014. Communication-Efficient Distributed Optimization Using an Approximate Newton-Type Method. In ICML. Beijing, China, II–1000–II–1008.

[27] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. 2017. Federated Multi-Task Learning. In Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4424–4434.

[28] Virginia Smith, Simone Forte, Chenxin Ma, Martin Taká, Michael I. Jordan, and Martin Jaggi. 2018. CoCoA: A General Framework for Communication-Efficient Distributed Optimization. Journal of Machine Learning Research 18, 230 (2018), 1–49.

[29] Nguyen H. Tran, Wei Bao, Albert Zomaya, Minh N.H. Nguyen, and Choong Seon Hong. 2019. Federated Learning over Wireless Networks: Optimization Model Design and Analysis. In IEEE INFOCOM 2019. Paris, France.

[30] Jianyu Wang and Gauri Joshi. 2018. Cooperative SGD: A Unified Frame-work for the Design and Analysis of Communication-Efficient SGD Algorithms. arXiv:1808.07576 [cs, stat] (Aug. 2018). arXiv:1808.07576 [cs, stat]

[31] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan. 2019. Adaptive Federated Learning in Resource Constrained Edge Computing Systems. IEEE Journal on Selected Areas in Communications 37, 6 (June 2019), 1205–1221. https://doi.org/10.1109/JSAC.2019.2904348

[32] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. 2018. SpiderBoost: A Class of Faster Variance-Reduced Algorithms for Nonconvex Optimization. arXiv:1810.10690 [cs, math, stat] (Oct. 2018). arXiv:1810.10690 [cs, math, stat]

[33] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv:1708.07747 [cs, stat] (Aug. 2017). arXiv:1708.07747 [cs, stat]

[34] Sixin Zhang, Anna E Choromanska, and Yann LeCun. 2015. Deep Learning with Elastic Averaging SGD. In Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 685–693.

对现有有用成果的回顾

A.1 强凸性

假设Jn(·)为μ ~ -强凸，μ ~ > 0，根据[21][定理2.1.10，等式。(2.1.24)和(2.1.26)]，我们有

$$2\tilde{\mu}(J_n(w) - J_n(w^*)) \le \left\|\nabla J(w)\right\|^2, \forall w. \tag{25}$$

$$\tilde{\mu}\left\|w - w^*\right\| \le \left\|\nabla J(w)\right\|, \forall w. \tag{26}$$

其中w*是问题(6)的解，即∇J(w*)= 0。

A.2 近端随机梯度方法

在Alg. 1的上下文中，假设问题(6)是由每个设备的局部迭代来解决的(在一个固定的全局迭代s中)

$$w_{n,s}^{(t+1)} = \text{prox}_{\eta h_s}\left[w_{n,s}^{(t)} - \eta v_{n,s}^{(t)}\right] \tag{27}$$

其中，s (t)为随机(方差约简或递归)梯度下降(即SVRG或SARAH)向量。定义

$$\hat{w}_{n,s}^{(t+1)} := \text{prox}_{\eta h_s}\left[w_{n,s}^{(t)} - \eta \nabla F_n(w_{n,s}^{(t)})\right], \tag{28}$$

为近端(全)梯度向量，则根据[18]的Eq.(25)，可得

$$\begin{aligned}&\mathbb{E}\left[J_n(w_{n,s}^{(t+1)}) - J_n(w_{n,s}^{(t)})\right]\\ &\le \mathbb{E}\Big[-\Big(\frac{1}{3\eta} - L\Big)\left\|w_{n,s}^{(t)} - \hat{w}_{n,s}^{(t+1)}\right\|^2 + \eta\left\|\nabla F_n(w_{n,s}^{(t)}) - v_{n,s}^{(t)}\right\|^2\\ &\quad - \Big(\frac{5}{8\eta} - \frac{L}{2}\Big)\left\|w_{n,s}^{(t+1)} - w_{n,s}^{(t)}\right\|^2\Big]\end{aligned} \tag{29}$$

我们接下来定义梯度映射[8]

$$G_n(w) := \frac{1}{\eta}\left(w - \text{prox}_{\eta h_s}(w - \eta \nabla F_n(w))\right) \tag{30}$$

然后我们看到

$$\left\|w_{n,s}^{(t)} - \hat{w}_{n,s}^{(t+1)}\right\|^2 = \eta^2\left\|G_n(w_{n,s}^{(t)})\right\|^2. \tag{31}$$

A.3 SARAH

我们重写了SARAH算法的核心[22]

$$v_{n,s}^{(t)} = \nabla f_{i_t}(w_{n,s}^{(t)}) - \nabla f_{i_t}(w_{n,s}^{(t-1)}) + v_{n,s}^{(t-1)} \tag{32}$$

根据[22]中的引理2，在假设1中的l -光滑性和vn下，s (t)在(32)中被更新，我们有

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla F_n(w_{n,s}^{(t)}) - v_{n,s}^{(t)}\right\|^2\right] \le \sum_{j=1}^{t}\mathbb{E}\left[\left\|v_{n,s}^{(j)} - v_{n,s}^{(j-1)}\right\|^2\right]\\ &= \sum_{j=1}^{t}\left\|\nabla f_{i_t}(w_{n,s}^{(j)}) - \nabla f_{i_t}(w_{n,s}^{(j-1)})\right\|^2 \le \sum_{j=1}^{t} L^2\left\|w_{n,s}^{(j)} - w_{n,s}^{(j-1)}\right\|^2.\end{aligned} \tag{33}$$

A.4 SVRG

我们重写了SVRG算法的核心[11]

$$v_{n,s}^{(t)} = \nabla f_{i_t}(w_{n,s}^{(t)}) - \nabla f_{i_t}(\bar{w}^{(s-1)}) + \nabla F_n(\bar{w}^{(s-1)}) \tag{34}$$

根据[22]中的引理3和[18]中的Eq.(29)，我们有

$$\mathbb{E}\left[\left\|\nabla F_n(w_{n,s}^{(t)}) - v_{n,s}^{(t)}\right\|^2\right] \le L^2\,\mathbb{E}\left[\left\|w_{n,s}^{(t)} - \bar{w}^{(s-1)}\right\|^2\right]. \tag{35}$$

B对引理1的证明

下面的结果以后会有用。

引理2。

$$\left\|\nabla J_n(w)\right\|^2 \le 2(\mu^2\eta^2 + 1)\left\|G_n(w)\right\|^2, \forall w. \tag{36}$$

证明:为了简单，通过定义w+ = proxηh s (w−η∇F n (w))，从(30)，我们有

$$G_n(w) = \frac{1}{\eta}(w - w^+) \tag{37}$$

$$= \nabla F_n(w) + \nabla h_s(w^+). \tag{38}$$

然后我们得到

$$\left\|\nabla J_n(w)\right\|^2 \le 2\left\|\nabla J_n(w) - G_n(w)\right\|^2 + 2\left\|G_n(w)\right\|^2 \tag{39}$$

$$= 2\left\|\nabla h_s(w) - \nabla h_s(w^+)\right\|^2 + 2\left\|G_n(w)\right\|^2 \tag{40}$$

$$= 2\mu^2\left\|w - w^+\right\|^2 + 2\left\|G_n(w)\right\|^2 \tag{41}$$

$$= 2(\mu^2\eta^2 + 1)\left\|G_n(w)\right\|^2, \tag{42}$$

其中(39)由于Young不等式(α= 1)而成立

$$\|x+y\|^2 \le (1+\alpha)\|x\|^2 + (1+1/\alpha)\|y\|^2, \forall x, y \in \mathbb{R}^d, \quad (43)$$

(40)适用于(38)和Jn(·)的定义，(41)使用h(·)的定义，(42)使用(37)。

B.0.1 FedProxVR与SARAH。通过考虑一个全局迭代s，我们看到Alg. 1 with v(t)n,s update(32)是解决问题(6)的近端SARAH。因此，使用现有的结果(29)，(30)和(33)，我们有

$$\begin{aligned}&\mathbb{E}\left[J_n(w_{n,s}^{(t+1)}) - J_n(w_{n,s}^{(t)})\right]\\ &\le \mathbb{E}\Big[-\Big(\frac{1}{3\eta} - L\Big)\eta^2\|G_n(w_{n,s}^{(t)})\|^2 + \eta L^2\sum_{j=1}^{t}\|w_{n,s}^{(j)} - w_{n,s}^{(j-1)}\|^2\\ &\quad - \Big(\frac{5}{8\eta} - \frac{L}{2}\Big)\|w_{n,s}^{(t+1)} - w_{n,s}^{(t)}\|^2\Big].\end{aligned}$$

通过设置η= β1 L，并将上述不等式在0⩽t⩽τ上相加，我们有:

$$\begin{aligned}&\mathbb{E}\left[J_n(w_{n,s}^{(\tau+1)}) - \overline{J_n(w_{n,s}^{(0)})}\right]\\ &\le \mathbb{E}\Big[-\Big(\frac{\beta-3}{3L\beta^2}\Big)\sum_{t=0}^{\tau}\|G_n(w_{n,s}^{(t)})\|^2 + \frac{L}{\beta}\sum_{t=1}^{\tau}\sum_{j=1}^{t}\|w_{n,s}^{(j)} - w_{n,s}^{(j-1)}\|^2 \quad (44)\\ &\quad - L\Big(\frac{5\beta-4}{8}\Big)\sum_{t=0}^{\tau}\|w_{n,s}^{(t+1)} - w_{n,s}^{(t)}\|^2\Big]\\ &\le \mathbb{E}\Big[-\Big(\frac{\beta-3}{3L\beta^2}\Big)\sum_{t=0}^{\tau}\|G_n(w_{n,s}^{(t)})\|^2 + \frac{L}{\beta}\sum_{t=0}^{\tau-1}\tau\|w_{n,s}^{(t+1)} - w_{n,s}^{(t)}\|^2\\ &\quad - L\Big(\frac{5\beta-4}{8}\Big)\sum_{t=0}^{\tau-1}\|w_{n,s}^{(t+1)} - w_{n,s}^{(t)}\|^2\Big], \quad (45)\\ &\le \mathbb{E}\Big[-\Big(\frac{\beta-3}{3L\beta^2}\Big)\sum_{t=0}^{\tau}\|G_n(w_{n,s}^{(t)})\|^2\Big], \quad (46)\end{aligned}$$

h (0) (0) 2i其中(44)是由于E ∇F n (wn,s)−vn,s = 0和(45)在β⩾4/5和以下事实时成立

$$\begin{aligned}&\sum_{t=1}^{\tau}\sum_{j=1}^{t}\|w_{n,s}^{(j)} - w_{n,s}^{(j-1)}\|^2 \le \sum_{t=1}^{\tau}\sum_{j=1}^{\tau}\|w_{n,s}^{(j)} - w_{n,s}^{(j-1)}\|^2\\ &= \sum_{t=1}^{\tau}\tau\|w_{n,s}^{(t)} - w_{n,s}^{(t-1)}\|^2 = \sum_{t=0}^{\tau-1}\tau\|w_{n,s}^{(t+1)} - w_{n,s}^{(t)}\|^2 \quad (47)\end{aligned}$$

和(46)在我们选择时成立

$$0 \le \tau \le \frac{5\beta^2 - 4\beta}{8}. \quad (48)$$

从(46)中，我们有:

$$\begin{aligned}\frac{\beta-3}{3L\beta^2}\sum_{t=0}^{\tau}\mathbb{E}\left[\|G_n(w_{n,s}^{(t)})\|^2\right] &\le \mathbb{E}\left[J_n(w_{n,s}^{(0)}) - J_n(w_{n,s}^{(\tau+1)})\right]\\ &\le \mathbb{E}\left[J_n(w_{n,s}^{(0)}) - J_n(w_{n,s}^{*})\right] \quad (49)\\ &\le \frac{\mathbb{E}\left[\|\nabla J_n(w_{n,s}^{(0)})\|^2\right]}{2\tilde{\mu}} \quad (50)\end{aligned}$$

其中(49)使用Jn (wn,s *)⩽Jn (wn,s(τ +1))，(50)使用(25)。

最后，我们有

$$\begin{aligned}&\mathbb{E}\left[\|\nabla J_n(w_n^{(s)})\|^2 \mid \tilde{w}^{(s-1)}\right]\\ &\le 2\Big(\frac{\mu^2}{\beta^2L^2} + 1\Big)\mathbb{E}\left[\|G_n(w_n^{(s)})\|^2 \mid \tilde{w}^{(s-1)}\right] \quad (51)\\ &= 2\Big(\frac{\mu^2}{\beta^2L^2} + 1\Big)\sum_{t=0}^{\tau}\frac{1}{\tau+1}\mathbb{E}\left[\|G_n(w_{n,s}^{(t)})\|^2 \mid \tilde{w}^{(s-1)}\right] \quad (52)\\ &\le 2\Big(\frac{\mu^2}{\beta^2L^2} + 1\Big)\frac{1}{2\tilde{\mu}}\Big(\frac{3L\beta^2}{\beta-3}\Big)\frac{1}{\tau+1}\mathbb{E}\left[\|\nabla J_n(w_{n,s}^{(0)})\|^2 \mid \tilde{w}^{(s-1)}\right] \quad (53)\\ &\le \frac{3(\beta^2L^2+\mu^2)}{\tilde{\mu}L(\beta-3)}\frac{1}{\tau}\|\nabla F_n(\tilde{w}^{(s-1)})\|^2 \quad (54)\end{aligned}$$

其中(51)由于引理2，(52)成立，因为wn (s)是从{wn,s (t)}中均匀随机选择的，t=0，…τ(53)使用(50)和(54)是由于w (0) n, s = w¯(s−1)。From(54)，为了有

$$\begin{aligned}\mathbb{E}\left[\|\nabla J_n(w_n^{(s)})\| \mid \tilde{w}^{(s-1)}\right] &\le \mathbb{E}\left[\|\nabla J_n(w_n^{(s)})\|^2 \mid \tilde{w}^{(s-1)}\right]^{\frac{1}{2}}\\ &\le \theta\|\nabla F_n(\tilde{w}^{(s-1)})\|\end{aligned}$$

我们获得

$$\tau \ge \frac{3(\beta^2L^2+\mu^2)}{\theta^2\tilde{\mu}L(\beta-3)} \ge 0. \quad (55)$$

B.0.2 FedProxVR与SVRG。同样，使用(29)和(35)，设置η= β1 L，我们得到

$$\begin{aligned}&\mathbb{E}\left[J_n(w_{n,s}^{(t+1)}) - \underline{J_n(w_{n,s}^{(t)})}\right]\\ &\le \mathbb{E}\Big[-\Big(\frac{\beta}{3} - 1\Big)\frac{1}{L\beta^2}\|G_n(w_{n,s}^{(t)})\|^2 + \frac{L}{\beta}\|w_{n,s}^{(t)} - \tilde{w}^{(s-1)}\|^2\\ &\quad - L\Big(\frac{5\beta-4}{8}\Big)\|w_{n,s}^{(t+1)} - w_{n,s}^{(t)}\|^2\Big]\\ &\le \mathbb{E}\Big[-\Big(\frac{\beta}{3} - 1\Big)\frac{1}{L\beta^2}\|G_n(w_{n,s}^{(t)})\|^2\\ &\quad + L\Big(\frac{1}{\beta} + \frac{1}{\sqrt{at}}\cdot\frac{5\beta-4}{8}\Big)\|w_{n,s}^{(t)} - \tilde{w}^{(s-1)}\|^2\\ &\quad - \Big(\frac{L}{\sqrt{at}+1}\cdot\frac{5\beta-4}{8}\Big)\|w_{n,s}^{(t+1)} - \tilde{w}^{(s-1)}\|^2\Big], \quad (56)\end{aligned}$$

其中(56)通过使用杨氏不等式成立

$$\begin{aligned}\|w_{n,s}^{(t+1)} - \tilde{w}^{(s-1)}\|^2 &\le (1+\frac{1}{\alpha})\|w_{n,s}^{(t)} - \tilde{w}^{(s-1)}\|^2 \quad (57)\\ &\quad + (1+\alpha)\|w_{n,s}^{(t+1)} - w_{n,s}^{(t)}\|^2,\end{aligned}$$

特别是通过选择α=pa(t+ 1)，a > 0，我们有:

$$\begin{aligned}&-\|w_{n,s}^{(t+1)} - w_{n,s}^{(t)}\|^2\\ &\le -\frac{1}{1+\alpha}\|w_{n,s}^{(t+1)} - \tilde{w}^{(s-1)}\|^2 + (\frac{1+1/\alpha}{1+\alpha})\|w_{n,s}^{(t)} - \tilde{w}^{(s-1)}\|^2\\ &= -\frac{1}{1+\sqrt{a(t+1)}}\|w_{n,s}^{(t+1)} - \tilde{w}^{(s-1)}\|^2 + \frac{1}{\sqrt{a(t+1)}}\|w_{n,s}^{(t)} - \tilde{w}^{(s-1)}\|^2.\end{aligned} \quad (58)$$

伸缩(56)大于0≤t≤τ，我们有:

$$\begin{aligned}
&\mathbb{E}\Big[J_n(w_{n,s}^{(\tau+1)}) - J_n(w_{n,s}^{(0)})\Big] \\
&\le \mathbb{E}\Big[-\Big(\frac{\beta-3}{3L\beta^2}\Big)\sum_{t=0}^{\tau}\big\|G_n(w_{n,s}^{(t)})\big\|^2 \\
&\quad - \sum_{t=0}^{\tau}\frac{L}{1+\sqrt{a(t+1)}}\Big(\frac{5\beta-4}{8}\Big)\big\|w_{n,s}^{(t+1)} - \tilde{w}^{(s-1)}\big\|^2 \\
&\quad + \sum_{t=1}^{\tau} L\Big(\frac{1}{\sqrt{a(t+1)}}\cdot\frac{5\beta-4}{8} + \frac{1}{\beta}\Big)\big\|w_{n,s}^{(t)} - \tilde{w}^{(s-1)}\big\|^2\Big] \qquad (59)\\
&\leqslant \mathrm{E}-\frac{\mathrm{H}\beta-3\,\tilde{\mathrm{O}}\tau}{3\,1\beta 2-}\ \mathrm{Gn}\,(\mathrm{wn,s\,(t)})\ 2 \\
&\quad -\overset{\tau}{\underset{t=0}{\mathrm{O}}}\ \mathrm{L}\,\delta\mathrm{t}\cdot\ \frac{5\beta-4}{8}-\frac{1}{\beta}\ \ \mathrm{wn\ s}\ \ -\mathrm{w}^{-}(\mathrm{s}-1)\ \ {}^{2}\text{我} \qquad (60)
\end{aligned}$$

$$\text{with } \delta_t := \frac{1}{1+\sqrt{a(t+1)}} - \frac{1}{\sqrt{a(t+2)}}$$

$$\begin{aligned}
&\le \mathbb{E}\Big[-\Big(\frac{\beta-3}{3\beta^2 L}\Big)\sum_{t=0}^{\tau}\big\|G_n(w_{n,s}^{(t)})\big\|^2 \\
&\quad - \sum_{t=0}^{\tau} L\Big(\frac{1}{a(t+2)}\cdot\frac{5\beta-4}{8} - \frac{1}{\beta}\Big)\big\|w_{n,s}^{(t+1)} - \tilde{w}^{(s-1)}\big\|^2\Big] \qquad (61)\\
&\le \mathbb{E}\Big[-\Big(\frac{\beta-3}{\quad}\Big)\sum^{\tau}\big\|G_n(w_{n\ \ s}^{(t)})\big\|^2\Big], \qquad (62)
\end{aligned}$$

其中(59)是由于wn,s (t) - w¯(s−1)= 0，当t = 0时，(61)成立，因为

$$\delta_t = \frac{\sqrt{a(t+2)} - (1+\sqrt{a(t+1)})}{\sqrt{a(t+2)}(1+\sqrt{a(t+1)})} \ge \frac{1}{a(t+2)},\ \forall 1 \le t \le \tau, \qquad (63)$$

当

$$\sqrt{a(t+2)} - (1+\sqrt{a(t+1)}) \ge 1, \forall t \le \tau, \qquad (64)$$

$$\Leftrightarrow a - 4 \ge 4\sqrt{a(\tau+1)} \qquad (65)$$

和(62)在条件下成立

$$\tau \le \frac{5\beta^2 - 4\beta}{8\,a} - 2. \qquad (66)$$

由于(62)与(46)相同，因此证明的其余部分类似于从(50)到(54)的步骤，从而获得与(55)相同的τ下界。

定理1的C证明

我们首先定义分布pn = Dn /D，∀n，其期望符号为En(为了避免与只处理FedProxVR中所有随机性的E[·]混淆)，我们有

$$\mathbb{E}_n\big[w_n^{(s)}\big] = \sum\nolimits_{n=1}^{N} p_n w_n^{(s)} = \tilde{w}^{(s)}. \qquad (67)$$

通过定义Jn(.)和h(.)

$$\nabla J_n(w_n^{(s)}) = \nabla F_n(w_n^{(s)}) + \mu(w_n^{(s)} - \tilde{w}^{(s-1)}). \qquad (68)$$

因此，我们有

$$\mathbb{E}_n\big[\nabla J_n(w_n^{(s)}) - \nabla F_n(w_n^{(s)})\big] = \mu(\tilde{w}^{(s)} - \tilde{w}^{(s-1)}). \qquad (69)$$

我们首先提供了以下有用的结果:对w¯(s−1)进行条件反射，我们有

$$\mathbb{E}\Big[\big\|\tilde{w}^{(s)} - \tilde{w}^{(s-1)}\big\|^2\Big] \le \frac{4(1+\theta^2)(1+\bar{\sigma}^2)}{\tilde{\mu}^2}\big\|\nabla\bar{F}(\tilde{w}^{(s-1)})\big\|^2.$$

证明:定义wsn*= arg minwns Jn (wns)，我们得到

$$\big\|w_s^{n*} - \tilde{w}^{(s-1)}\big\|^2 \overset{(26)}{\le} \frac{1}{\tilde{\mu}^2}\big\|\nabla J_n(\tilde{w}^{(s-1)})\big\|^2 \overset{(68)}{=} \frac{1}{\tilde{\mu}^2}\big\|\nabla F_n(\tilde{w}^{(s-1)})\big\|^2.$$

类似地，我们有

$$\mathbb{E}\Big[\big\|w_s^{n*} - w_n^{(s)}\big\|^2\Big] \overset{(26)}{\le} \frac{1}{\tilde{\mu}^2}\mathbb{E}\Big[\big\|\nabla J_n(w_n^{(s)})\big\|^2\Big] \le \frac{\theta^2}{\tilde{\mu}^2}\big\|\nabla F_n(\tilde{w}^{(s-1)})\big\|^2, \qquad (70)$$

其中(70)使用引理1。然后，使用Young不等式(43)(α= 1)，我们有

$$\begin{aligned}
\mathbb{E}\Big[\big\|w_n^{(s)} - \tilde{w}^{(s-1)}\big\|^2\Big] &\le \mathbb{E}\Big[2\big\|w_s^{n*} - w_n^{(s)}\big\|^2 + 2\big\|w_s^{n*} - \tilde{w}^{(s-1)}\big\|^2\Big] \\
&\le \frac{2(1+\theta^2)}{\tilde{\mu}^2}\big\|\nabla F_n(\tilde{w}^{(s-1)})\big\|^2. \qquad (71)
\end{aligned}$$

我们得到了下一个有用的结果

$$\begin{aligned}
&\mathbb{E}_n\big[\big\|\nabla F_n(\tilde{w}^{(s-1)})\big\|^2\big] \\
&\overset{(43)}{\le} \mathbb{E}_n\big[2\big\|\nabla F_n(\tilde{w}^{(s-1)}) - \nabla\bar{F}(\tilde{w}^{(s-1)})\big\|^2\big] + 2\big\|\nabla\bar{F}(\tilde{w}^{(s-1)})\big\|^2 \\
&\le 2(1+\bar{\sigma}^2)\big\|\nabla\bar{F}(\tilde{w}^{(s-1)})\big\|^2, \qquad (72)
\end{aligned}$$

其中(72)由于假设1而成立。最后，我们有

$$\begin{aligned}
\mathbb{E}\Big[\big\|\tilde{w}^{(s)} - \tilde{w}^{(s-1)}\big\|^2\Big] &= \mathbb{E}\Big[\big\|\mathbb{E}_n[w_n^{(s)} - \tilde{w}^{(s-1)}]\big\|^2\Big] \\
&\le \mathbb{E}\Big[\mathbb{E}_n\big[\big\|w_n^{(s)} - \tilde{w}^{(s-1)}\big\|\big]^2\Big] \qquad (73)\\
&\le \frac{2(1+\theta^2)}{\tilde{\mu}^2}\mathbb{E}_n\big[\big\|\nabla F_n(\tilde{w}^{(s-1)})\big\|^2\big] \qquad (74)\\
&\le \frac{4(1+\theta^2)(1+\bar{\sigma}^2)}{\tilde{\mu}^2}\big\|\nabla\bar{F}(\tilde{w}^{(s-1)})\big\|^2, \qquad (75)
\end{aligned}$$

其中(73)使用Jensen不等式，(74)由于期望和使用(71)的交换而成立，(75)使用(72)。

由于F(.)¯也是L- lipschitz光滑(即∇F(w)¯−∇F(w¯')≤En∇F n (w)−∇F n (w ')≤L w−w '，通过两次使用Jensen不等式和假设1)，我们有

$$\begin{aligned}
&\bar{F}(\tilde{w}^{(s)}) - \bar{F}(\tilde{w}^{(s-1)}) \le \langle\nabla\bar{F}(\tilde{w}^{(s-1)}), \tilde{w}^{(s)} - \tilde{w}^{(s-1)}\rangle + \frac{L}{2}\big\|\tilde{w}^{(s)} - \tilde{w}^{(s-1)}\big\|^2 \\
&= \frac{1}{\mu}(\tilde{w}^{(s-1)}), \mathbb{E}_n\big[\nabla J_n(w_n^{(s)}) - \nabla F_n(w_n^{(s)})\big] + \frac{L}{2}\big\|\tilde{w}^{(s)} - \tilde{w}^{(s-1)}\big\|^2 \qquad (76)\\
&= \frac{1}{\mu}\big\langle\nabla\bar{F}(\tilde{w}^{(s-1)}), \mathbb{E}_n\big[\nabla J_n(w_n^{(s)}) - \nabla F_n(w_n^{(s)})\big] + \nabla\bar{F}(\tilde{w}^{(s-1)})\big\rangle \\
&\quad - \frac{1}{\mu}\big\|\nabla\bar{F}(\tilde{w}^{(s-1)})\big\|^2 + \frac{L}{2}\big\|\tilde{w}^{(s)} - \tilde{w}^{(s-1)}\big\|^2 \qquad (77)\\
&= \frac{1}{\mu}\big\langle\nabla\bar{F}(\tilde{w}^{(s-1)}), \mathbb{E}_n\big[\nabla F_n(\tilde{w}^{(s-1)}) - \nabla F_n(w_n^{(s)})\big]\big\rangle \\
&\quad + \frac{1}{\mu}\big\langle\nabla\bar{F}(\tilde{w}^{(s-1)}), \mathbb{E}_n\big[\nabla J_n(w_n^{(s)})\big]\big\rangle \\
&\quad - \frac{1}{\mu}\big\|\nabla\bar{F}(\tilde{w}^{(s-1)})\big\|^2 + \frac{L}{2}\big\|\tilde{w}^{(s)} - \tilde{w}^{(s-1)}\big\|^2 \qquad (78)\\
&\le \frac{1}{\mu}\big\|\nabla\bar{F}(\tilde{w}^{(s-1)})\big\| \cdot \mathbb{E}_n\big[\big\|\nabla F_n(\tilde{w}^{(s-1)}) - \nabla F_n(w_n^{(s)})\big\|\big] \\
&\quad + \frac{1}{\mu}\big\|\nabla\bar{F}(\tilde{w}^{(s-1)})\big\| \cdot \mathbb{E}_n\big[\big\|\nabla J_n(w_n^{(s)})\big\|\big] \\
&\quad - \frac{1}{\mu}\big\|\nabla\bar{F}(\tilde{w}^{(s-1)})\big\|^2 + \frac{L}{2}\big\|\tilde{w}^{(s)} - \tilde{w}^{(s-1)}\big\|^2, \qquad (79)
\end{aligned}$$

式中(76)使用式(69)，式(77)通过加减fn (wn (s))成立，式(79)使用Cauchy-Schwarz和Jensen不等式。

取(79)的条件期望，我们有

$$\mathbb{E}\left[\bar{F}(\bar{w}^{(s)}) - \bar{F}(\bar{w}^{(s-1)}) \mid \bar{w}^{(s-1)}\right]$$

$$\leqslant \frac{L}{\mu}\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\| \cdot \mathbb{E}_n\left[\mathbb{E}\left[\left\|\bar{w}^{(s-1)} - w_n^{(s)}\right\|\right]\right] \tag{80}$$

$$+ \frac{1}{\mu}\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\| \cdot \mathbb{E}_n\left[\mathbb{E}\left[\left\|\nabla J_n(w_n^{(s)})\right\|\right]\right] \tag{81}$$

$$- \frac{1}{\mu}\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\|^2 + \frac{L}{2}\mathbb{E}\left[\left\|\bar{w}^{(s)} - \bar{w}^{(s-1)}\right\|^2\right] \tag{82}$$

$$\leq \frac{L}{\mu}\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\| \cdot \frac{2}{\tilde{\mu}}\sqrt{(1+\theta^2)(1+\bar{\sigma}^2)}\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\| \tag{83}$$

$$+ \frac{1}{\mu}\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\| \cdot \theta\sqrt{2(1+\bar{\sigma}^2)}\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\| \tag{84}$$

$$- \frac{1}{\mu}\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\|^2 + \frac{2L(1+\theta^2)(1+\bar{\sigma}^2)}{\tilde{\mu}^2}\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\|^2 \tag{85}$$

$$\leq -\frac{1}{\mu}\left(1 - \frac{2L}{\tilde{\mu}}\sqrt{(1+\theta^2)(1+\bar{\sigma}^2)} - \theta\sqrt{2(1+\bar{\sigma}^2)}\right.$$

$$\left. - \frac{2L\mu}{\tilde{\mu}^2}(1+\theta^2)(1+\bar{\sigma}^2)\right)\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\|^2, \tag{86}$$

$$= -\Theta\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\|^2. \tag{87}$$

其中(80)是由于Fn(.)的l -平滑性，我们在(80)和(81)中交换了期望。(83)成立，因为w¯(s−1)的条件成立，我们有

$$\mathbb{E}_n\left[\mathbb{E}\left[\left\|\bar{w}^{(s-1)} - w_n^{(s)}\right\|\right] \mid \bar{w}^{(s-1)}\right] \leq \mathbb{E}_n\left[\mathbb{E}\left[\left\|\bar{w}^{(s-1)} - w_n^{(s)}\right\|^2\right]^{\frac{1}{2}}\right]$$

$$\overset{(71)}{\leq} \frac{\sqrt{2(1+\theta^2)}}{\tilde{\mu}}\mathbb{E}_n\left[\left\|\nabla F_n(\bar{w}^{(s-1)})\right\|\right] \leq \frac{\sqrt{2(1+\theta^2)}}{\tilde{\mu}}\left(\mathbb{E}_n\left[\left\|\nabla F_n(\bar{w}^{(s-1)})\right\|^2\right]\right)^{\frac{1}{2}}$$

$$\overset{(72)}{\leq} \frac{2}{\tilde{\mu}}\sqrt{(1+\theta^2)(1+\bar{\sigma}^2)}\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\|$$

类似地，(84)成立because

$$\mathbb{E}_n\left[\mathbb{E}\left[\left\|\nabla J_n(w_n^{(s)})\right\|\right]\right] \overset{\text{Lemma 1}}{\leq} \theta\,\mathbb{E}_n\left[\left\|\nabla F_n(\bar{w}^{(s-1)})\right\|\right]$$

$$\overset{(72)}{\leq} \theta\sqrt{2(1+\bar{\sigma}^2)}\left\|\nabla\bar{F}(\bar{w}^{(s-1)})\right\|.$$

(85)用到了引理3。

取(86)对整个历史的期望从1伸缩到T，我们有

$$\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\left[\left\|\bar{F}(\bar{w}^{(s)})\right\|^2\right] \leq \frac{\mathbb{E}\left[\bar{F}(\bar{w}^{(0)}) - \bar{F}(\bar{w}^{(T)})\right]}{\Theta T} \leq \frac{\mathbb{E}\left[\bar{F}(\bar{w}^{(0)}) - \bar{F}(\bar{w}^{*})\right]}{\Theta T}.$$