

非IID数据的联合学习



赵悦\*、李萌、赖良珍、娜文素达、达蒙·茨文、维卡斯·钱德拉

Arm，加利福尼亚州圣何塞

{岳.赵, 孟莉, 良珍.赖, 纳文苏达, 达蒙·西文, 钱达斯}@arm.com公司

# 摘要

联合学习使资源受限的边缘计算设备（如移动电话和物联网设备）能够学习用于预测的共享模型，同时保持训练数据的本地性。这种去中心化的培训模式提供了隐私、安全、监管和经济效益。在这项工作中，我们关注的是当局部数据是非IID时联合学习的统计挑战。我们首先证明了联合学习的准确度显著降低，对于为高度倾斜的非IID数据训练的神经网络，联邦学习的准确度降低了约55%，在这种情况下，每个客户端设备只对一类数据进行训练。我们进一步证明，这种精度降 低可以用权重偏差来解释，权重散度可以通过每个设备上的类分布和总体分布之间的土方移动距离（EMD）来量化。作为一种解决方案，我们提出了一种策略，通过创建在所有边缘设备之间全局共享的数据子集来改进非IID数据的训练。实验表明，对于只有5%的全局共享数据的CIFAR-10数据集，准确率可以提高30%。

# 1        介绍

移动设备已成为全球数十亿用户的主要计算资源，预计未来几年将有数十亿的物联网设备上线。这些设备产生大量有价值的数据，使用这些数据训练的机器学习模型有可能提高许多应用程序的智能性。但在移动设备上启用这些功能通常需要在服务器上全局共享数据，以便训练出令人满意的模型。从隐私、安全、监管或经济角度来看，这可能是不可能或不可取的。因此，在设备上保存数据并共享模型的方法变得越来越有吸引力。

近年来，关于该装置的推理有了很大的进展[1,2]。联合学习

[3，4，5]也提供了在当地进行培训的方法。McMahan等人。[3] 介绍了（FedAvg）算法，并证明了在基准图像分类数据集（如MNIST[6]和CIFAR-10[7]）上训练卷积神经网络（CNNs）和语言数据集LSTM的鲁棒性[8]。*联邦平均法费达夫*

近年来，人们对联合学习的通信挑战进行了大量的研究，如如何降低传送大量深层网络权值矩阵的通信成本，以及由于网络连通性、功耗和计算限制而导致的意外丢失或同步延迟。Bonawitz等人。[9] 为联合学习开发了一个有效的安全聚合协议，允许服务器从移动设备执行高维数据的计算。Konecnˇy`等人。[10] 提出结构化更新和草图更新，将通信成本降低两个数量级。Lin等人。[11] 提出了深度梯度压缩（DGC）将通信带宽降低两个数量级来训练高质量的模型。

预印本。正在进行中。

除了沟通的挑战，联合学习也面临统计上的挑战。联合学习广泛地依赖于经验下降（14，15），其中随机学习是广泛使用的。训练数据的IID抽样对于确保随机梯度是全梯度的无偏估计非常重要[15，16，18]。实际上，假设每个边缘设备上的本地数据总是IID是不现实的。为了应对非IID挑战，Smith等人。[28]提出了多任务学习（MTL）框架，并开发了MOCHA来解决MTL中的系统挑战。但是这种方法与以前的联合学习研究有很大的不同。McMahan等人。[3] 已经证明可以处理某些非IID数据。然而，正如我们将在第2节中展示的那样，使用算法训练的卷积神经网络的精确度可以显著降低，对于MNIST，CIFAR-10和关键字定位（KWS）数据集，使用高度倾斜的非IID数据集，精度可以显著降低，高达11%，CIFAR-10和55%。*费达夫费达夫*

为了解决联邦学习在统计上的挑战，我们在第3节中指出，精度的降低可以归因于权重发散，权重发散量化了来自两个不同训练过程的权重差异，权重初始值相同。然后我们证明了训练中的权重散度是由每个设备（或客户端）上的类的分布和总体分布之间的推土距离（EMD）所限定的。这个界限受学习速率、同步步数和梯度的影响。最后，在第4节中，我们提出了一种数据共享策略，通过分配少量的包含每个类的示例的全局共享数据来改进非IID数据。这就需要在准确性和集中性之间进行权衡。实验表明，在CIFAR-10上，如果我们愿意集中和分发5%的共享数据，则准确率可以提高30%。*费达夫*

# 2        非IID数据的FedAvg

在这一节中，我们通过在三个数据集上训练具有代表性的神经网络来演示非IID数据的精度降低。*费达夫*

2.1实验装置

在这项工作中，我们使用了训练在MNIST[6]、CIFAR-10[7]和语音命令数据集[29]上的卷积神经网络（CNNs）。MNIST和CIFAR-10是用于图像分类任务的数据集，具有10个输出类。语音命令数据集由35个单词组成，每个单词持续1秒。为了使其一致，我们使用一个包含10个关键字的数据子集作为关键字定位（KWS）数据集。对于每个音频片段，我们提取了10个MFCC特征，每帧30ms，步长为20ms，生成50x10个特征，用于神经网络训练。对于MNIST和CIFAR-10，我们使用与[3]相同的CNN架构，对于KWS，我们使用[1]中的CNN架构。

训练集被均匀地分为10个客户机。对于IID设置，每个客户端随机分配10个类的统一分布。对于非IID设置，数据按类排序并划分为两种极端情况：（a）1类非IID，每个客户端仅从单个类接收数据分区；（b）2类非IID，其中排序的数据被分成20个分区，每个客户端从2个类中随机分配2个分区。

我们对算法使用了与[3]相同的符号：B，批大小，E，局部时间的数量。以下参数用于：对于MNIST，B=10和100，E=1和5，=0.01，衰变率=0.995；对于CIFAR-10，B=10和100，E=1和5，=0.1，衰变率=0.992；对于KWS，B=10和50，E=1和5，=0.05，衰变率=0.992。每一个数据集的学习率都得到了优化，并且在通信轮次中呈指数衰减。对于SGD，学习率和衰减率是相同的，但B是10倍大。这是因为在每次同步时，在10个客户机之间平均使用来自的全局模型。使用IID的数据应该与使用洗牌数据的SGD进行比较，并且批处理大小要大几倍，其中是每次同步时包含的客户端数量。*费达夫费达夫ηηη费达夫费达夫KK费达夫*

2.2实验结果

对于IID实验，批量大小为的收敛曲线与SGD曲线基本重合，三个数据集的收敛曲线均为×10（图1）。仅观察到CIFAR-10的差异很小，B=10时收敛到82.62%，而B=100时的SGD*费达夫BB费达夫*

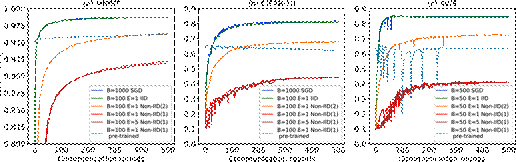


图1：使用（a）MNIST（b）CIFAR-10和（c）KWS数据集的IID和非IID数据集与SGD相比的通信轮的测试精度。非IID（2）表示2类非IID，非IID（1）表示1类非IID。*费达夫*

收敛到84.14%（图A.1）。从而实现了IID数据的SGD级测试精度，与文献[3]的结果一致。*费达夫*

与具有匹配批次大小的SGD相比，非IID数据的测试精度显著降低（图1和A.1）。表1总结了非IID数据的精度降低。对于最极端的1类非IID数据，最大精度降低。此外，更多的局部时间（E=5）并不能减少损失。E=1和E=5的收敛曲线大部分重叠。此外，SGD预先训练的CNN模型没有从非IID数据的训练中学习。对于CIFAR-10，当预先训练的CNN在非IID数据上训练时，精度会下降。因此，我们证明了非IID数据的测试精度降低。表A.2总结了所有实验的测试精度。请注意，本文中报告的标准差精度不是最先进的[6,30,31,1]，但我们训练的cnn足以实现我们对非IID数据的联邦学习进行评估的目标。*费达夫费达夫费达夫费达夫*

表1：非IID数据的测试精度降低。*费达夫*

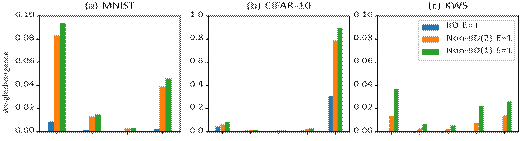
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 非IID | B | E | 最小值（%） | CIFAR-10（%） | 千瓦时（%） |
| 非IID（1） | 大的 | 1 | 6.52 | 37.66 | 43.64 |
| 非IID（1） | 大的 | 5 | 6.77 | 37.11 | 43.62 |
| 非IID（2） | 大的 | 1 | 2.4 | 14.51 | 12.16 |
| 非IID（1） | 小的 | 1 | 11.31 | 51.31 | 54.5 |
| 非IID（2） | 小的 | 1 | 1.77 | 15.61 | 15.07 |

# 3        非IID数据导致的重量偏差

在图1和A.1中，值得注意的是，2类非IID数据的缩减比1类非IID数据的减少少。结果表明，数据分布的精确性，即数据分布的偏斜性，可能会影响数据的准确性。由于测试精度是由训练的权重决定的，与SGD比较的另一种方法是在相同的权重初始化的情况下，查看相对于SGD的权重的差异。它被称为重量散度，可通过以下公式计算：*费达夫费达夫*

*重量发散*= ||*w费达夫*−*w新加坡元*||/||*w新加坡元*||（一）

如图2所示，随着数据变得越来越非IID，从IID到2级非IID再到1级非IID，所有层的权重发散都会增大。因此，预计权重散度和数据偏斜之间存在关联。第2节中发现的精度降低可以理解为权重发散，它量化了来自两个不同训练过程的权重差异，具有相同的权重初始化。在这一部分中，我们正式分析了权重差异的来源。在第3.1节中，我们提供了一个说明性的例子和一个形式化的命题来证明权重差异的根本原因是由于每个客户端上的数据分布与总体分布之间的距离。具体地说，我们发现这种距离可以用分布之间的土方移动距离（EMD）来评估。然后，在第节



conv1conv2fc1 fc2 Conv1Conv2FC2 fc2 fc3 Conv1Conv2FC2 fc3

                                  图层图层图层

图2:IID、2级非IID和1级非IID的CNN层的权重差异。

3.2，我们通过实验验证了EMD对权重散度和测试精度的影响。*费达夫*

3.1数学论证

我们对联合学习问题进行了形式化的定义，并分析了权重发散的根源。

我们考虑一个定义在紧空间X和标签空间Y=[C]上的类分类问题，其中[C]={1，…，C}。数据点{}沿着分布分布分布在X×Y上。函数：X→S将X映射到概率单纯形S，其中*C十，ypf*

1，zi≥0，∀i∈[C]}，表示第th类的概率。在假设类w上参数化，即神经网络的权重。我们将广泛使用的交叉熵损失定义为*金融机构我f`w*

*C C*

*`*（）=E∼p[1logfi（）]=（y=i）E | y[logfi（）]。*w十，y*十*是的*=我*十,w*十*p十*=我*十,w*

*我*=1 i=1

为了简化分析，我们忽略了泛化误差，假设种群损失是直接优化的。因此，学习问题成为

*C*

*.*

为了确定w，SGD迭代求解优化问题。让w表示之后的重量*t*-在集中设置中更新。然后，集中式SGD执行以下更新：

*C*

*w.*

=1

在联合学习中，我们假设有客户。表示数据量，表示客户机上的数据分布∈[K]。在每个客户机上，分别进行本地SGD。在客户端∈[K]上迭代时，局部SGD执行*Kn*（k）*p*（k）*ktk*

*C*

*w.*

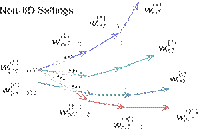
=1

然后，假设同步是每一步进行的，并让w表示第-次同步后计算的权重，则*T公吨*（六）*米*

*K*（k）

*w**.*

wand Wc之间的差异可以通过图3中的图示来理解。当数据是IID时，对于每个客户端，w*公吨*（六）*公吨*（三）*kt*（k）w是小的，在*米*-由于数据分布的不均匀性，使得数据在不同的数据源之间的分布也越来越大。为了正式地约束wand w之间的权重差异，我们有以下建议。*公吨*（六）*公吨*（三）*kt*（k）*t*（三）*公吨*（六）*公吨*（三）*公吨*（六）*公吨*（三）

IID设置：

客户1

（1） （1）客户K*w公吨*

新加坡元

费达夫

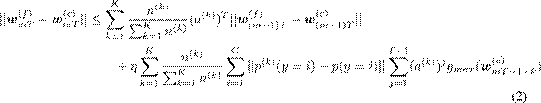
*w公吨*（K） 2（千）*w公吨*1 *w*

图3：使用IID和非IID数据的联邦学习的权重差异说明。

提案3.1。鉴于*K个客户，每个客户n*（k）*i、 分配后的i.d样品p*（k）*对于*

*每执行一次客户端同步k*∈[K]∇Ex | logfTi（）|∈[C]*. 如果w是的*=我*xsteps，那么，我们有下面的权重不等式,w是λ十是的*=我*-每堂课都有利普希茨我以及*

*散度第m次同步，*

*,*

*哪里**.*

命题3.1的详细证明见附录A.3。基于3.1号提案，我们有以下评论。

备注3.2。（m−1）后的重量散度*第m次同步主要来自两个部分，包括在同步后的权值发散-散度，即。，**T*||*以及客户机上数据分布的概率距离引起的权重发散k与整个人口的实际分布相比较，即：。，**.*

备注3.3。第次同步后的权重发散通过

*. 因为**. 因此，如果不同的客户机从*

*即使在初始权值不一致的情况下，IID的初始权值仍然会降低，这也会导致IID的初始精度下降。*

备注3.4。当所有客户机从与集中设置相同的初始化开始时，成为权重差异的根本原因。这个术语是客户机| | p（k）（y=i）−p（y=i）上的数据分布之间的EMD||*当距离测量定义为. EMD的效果受学习率的影响η、 同步之前的步骤数T、 梯度呢**.*

基于命题3.1，我们在第3.2节中验证了EMD是一个很好的度量来量化权重散度，从而验证了非IID数据的测试精度。*费达夫*

3.2实验验证

3.2.1实验装置

训练集被分类并划分为10个客户机，每个客户机都有示例。表2中列出了8个EMD值。由于一个EMD可能存在不同的分布，我们打算生成五个分布来计算权重散度和测试精度的平均值和变化。首先，对一个EMD生成一个超过10个类的概率分布。基于和，我们可以计算出一个客户端超过10个类的示例数。第二，通过将10个概率由1个元素移位生成一个新的分布。第二个客户机的示例数可以根据计算。对其他8个客户机重复此过程。因此，所有10个客户机都有一个分布在10个类上的M个示例，每个示例只使用一次。最后，将上述两个步骤重复5次，为每个EMD生成5个分布。CNN接受了500多轮关于上述程序处理数据的通信训练。有用于培训的关键参数：对于MNIST，B=100，E=*米P米PP*0 *PP*0*费达夫*

1, *η*=0.01，衰减率=0.995；对于CIFAR-10，B=100，E=1，*η*=0.1，衰减率=0.992；KWS，B=50，E=1，*η*=0.05，衰减率=0.992。在1次同步（即1次通信轮）后，根据公式（1）计算权重发散。

3.2.2重量偏差与EMD

对于每个EMD，计算5个分布的权重散度的平均值和标准差。对于这三个数据集，每一层的权重散度随着EMD的增加而增加，如图4所示。对于每个数据集上的所有SGD、IID和非IID实验，初始权重相同。因此，根据备注3.2，1次同步后的权重发散不受（m−1）的影响-散度，因为当*米*=1。因此，

图4中的结果支持命题3.1，即权重发散的界受EMD的影响。这种效应在第一个卷积层和最后一个完全连通层中更为显著。此外，由于问题本身和不同的CNN结构，CIFAR-10的最大权重散度显著高于MNIST和KWS，这受到公式（2）中梯度项的影响。与初始损失相同时，请注意，初始损失与客户的平均值相同[注3]。

3.2.3测试精度与EMD

对于每个EMD，在相同的5个分布上计算测试精度的平均值和标准差。结果汇总在表2中，并与EMD在图5中进行了对比。对于这三个数据集，EMD的测试精度都会下降。开始时下降率相对较小，随着数据变得更加非IID，下降率会变大。因此，在平衡非IID数据到IID数据和提高的准确性之间存在一种权衡。图中的误差条表示由于每个EMD分布的不同，测试精度的变化。为了更仔细地观察这些变化，当EMD=1.44时，箱线图显示了5次运行的测试精度。此外，表2显示，MNIST、CIFAR-10和KWS的最大变化分别小于0.086%、2%和1%。因此，精度受EMD的影响，而不是受潜在分布的影响。这表明EMD可以用来估计给定分布数据的精度。此外，MNIST、CIFAR-10和KWS的最大准确度分别降低了6.53%、37.03%和40.21%。这种差异可能会受到公式（2）中梯度的影响，这取决于问题本身。*费达夫费达夫*

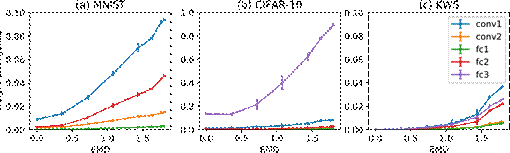


图4：（a）MNIST，（b）CIFAR-10和（c）KWS数据集上CNN层的权重差异与EMD。每个EMD的平均值和标准差在5个分布上计算。

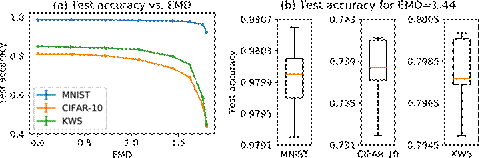


图5：（a）测试精度与EMD，以及（b）当EMD=1.44时，MNIST、CIFAR-10和KWS数据集的重量偏差箱线图。平均值和标准差是在5个分布上计算的。*费达夫*

表2：超过5个分布的检验精度的平均值和标准差。与平均值的刻度相比，标准差非常小。*费达夫*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 土方移动距离（EMD） | | 0 | 0.36 | 0.72 | 1.08 | 1.44 | 1.62 | 1.764 | 1.8 |
| MNIST公司 | 平均标准差（×10−4） | 0.9857  6.431 | 0.9860  2.939 | 0.9852  4.604 | 0.9835  4.308 | 0.9799  4.716 | 0.9756  8.085 | 0.962  8.232 | 0.922  1.939 |
| CIFAR10型 | 平均标准差（×10−3） | 0.8099  2.06 | 0.8090  2.694 | 0.8017  2.645 | 0.7817  3.622 | 0.7379  3.383 | 0.6905  2.048 | 0.5438  9.655 | 0.4396  1.068 |
| KWS公司 | 平均标准差（×10−3） | 0.8496  1.337 | 0.8461  3.930 | 0.8413  4.410 | 0.8331  5.387 | 0.7979  1.763 | 0.7565  3.329 | 0.5827  1.078 | 0.4475  4.464 |

# 4        建议解决方案

在本节中，我们提出了一种数据共享策略，通过创建在所有边缘设备之间全局共享的数据子集来改进非IID数据。实验表明，在CIFAR-10数据集上，只有5%的全局共享数据，测试精度可以提高30%左右。*费达夫*

4.1动机

如图5所示，当EMD超过一定阈值时，测试精度急剧下降。因此，对于高度倾斜的非IID数据，我们可以通过稍微减少EMD来显著提高测试精度。由于我们无法控制客户机的数据，所以我们可以将包含统一分布的一小部分全局数据从云端分发到客户机。这符合典型的联邦学习设置的初始化阶段。此外，预热模型可以根据全局共享的数据进行训练并分发给客户，而不是用随机权重分布模型。由于全球共享的数据可以减少客户的EMD，因此测试精度有望提高。

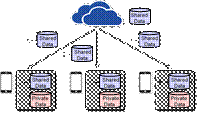
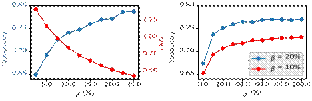
（a） （二）

图6：数据说明-图7：（a）测试精度和EMD vs.（b）测试精度共享策略。与分布分数*βα*

4.2数据共享策略

在此，我们提出了一个联邦学习环境中的数据共享策略，如图6所示。由类上的统一分布组成的全局共享数据集集中在云中。在的初始化阶段，训练的预热模型和的随机部分将分发给每个客户端。每个客户机的本地模型根据来自每个客户机的共享数据和私有数据进行训练。然后，云从客户端聚合本地模型，以训练全局模型。有两种权衡：（a）测试精度和*G费达夫GαGG费达夫G*，量化为100%，其中代表来自客户的总数据（b）测试准确度和。在CIFAR-10上进行了以下实验，以解决这两个问题。*Dα*

CIFAR-10训练集分为两部分，客户端部分有40000个示例，保留部分有10000个示例。用1类非IID数据划分为10个客户端。用于创建10个随机值，范围从2.5%到25%。首先，将每个完整的数据与每个客户机的数据进行合并，然后根据合并后的数据从头开始训练10个cnn，进行300多轮通信。测试精度如图7所示。其次，我们选择两个具体的，即当=10%和当=20%。对于每一个，（a）一个预热CNN模型被训练到大约60%的测试精度（b）只有一个随机部分被合并到每个客户的数据中，并且预热模型被训练在合并的数据上。测试精度如图7所示。第3节使用相同的训练参数。*DHDHG*0*sβG费达夫βG*0*sG*10%*βG*20%*βGGαα*

如图7（a）所示，随着测试精度的提高，测试精度提高到78.72%。即使使用较低的=10%，对于极端1类非IID数据，我们仍然可以达到74.12%的测试准确率，而没有数据共享策略的测试准确率为44%。此外，事实证明，没有必要将整个数据分发给客户，以达到类似的精度。相反，只需要将随机的一部分分发给每个客户机。如图7（b）所示，预热模型的测试精度缓慢提高，分别达到77.08%和73.12%。特别是，在最初的快速上升之后，当测试精度从50%变化到100%时，两者的测试精度变化都小于1%。因此，我们可以通过适当的选择来进一步减小每个客户端实际接收到的数据的大小。例如，如果我们愿意将10%的客户数据集中在云端，将50%的全球共享数据随机分发给客户，并在联合学习初始化阶段建立一个预热模型，对于极端1类非IID数据，其测试精度可提高30%，而每个客户端实际接收的数据仅占总数据的5%。*ββGGαG*20%*G*10%*G*20%*G*10%*αα费达夫*

总之，数据共享策略为非IID数据的联合学习提供了一个解决方案。全局共享数据集和随机分布分数（）的大小可以根据问题和应用程序进行调整。当联邦学习初始化时，该策略只需执行一次，因此通信开销不是主要问题。全局共享数据是一个独立于客户端数据的数据集，因此它对隐私不敏感。*α*

# 5        结论

联合学习将在分布式机器学习中发挥关键作用，其中数据隐私至关重要。不幸的是，如果每个边缘设备看到数据的唯一分布，模型训练的质量就会降低。在这项工作中，我们首先证明了联合学习的准确度显著降低，对于在高度倾斜的非IID数据上训练的神经网络，其准确度降低了约55%。我们进一步证明，这种精度下降可以用重量偏差来解释，重量偏差可以通过每个设备上的类分布和人口分布之间的土方移动距离（EMD）来量化。作为一种解决方案，我们提出了一种策略，通过创建在所有边缘设备之间全局共享的数据子集来改进非IID数据的训练。实验表明，对于只有5%的全局共享数据的CIFAR-10数据集，准确率可以提高30%。要使联合学习成为主流还有很多挑战，但是改进非IID数据的模型训练是在这方面取得进展的关键。

# 工具书类

[1] “你好，我是张先生，你好，我是张先生，我是张先生arXiv:1711.071282017年。

[2] L.Lai，N.Suda，V.Chandra，“Cmsis nn:arm cortex-m CPU的有效神经网络内核”，arXiv预印本AR601年第1801.14条2018年。

[3] H.B.McMahan、E.Moore、D.Ramage、S.Hampson和B.A.y Arcas，“从分散数据中高效学习深层网络”，载于《人工智能与统计国际会议》，2017年。

[4] J.Konecnˇy、B.McMahan和D.Ramage，“联邦优化：数据中心之外的分布式优化”，arXiv预印本arXiv:1511.03575，2015年。

[5] H.B.McMahan和D.Ramage，“联合学习：无集中培训数据的协作机器学习”，谷歌，2017年。

[6] Y.Lecun、L.Bottou、Y.Bengio和P.Haffner，“基于梯度的学习应用于文档识别”，发表于IEEE论文集，第2278-23241998年。

[7] A.Krizhevsky和G.Hinton，“从微小图像中学习多层特征”，多伦多大学技术报告，2009年。

[8] 莎士比亚，《莎士比亚全集》，年https://www.gutenberg.org/ebooks/100。

[9] K.Bonawitz、V.Ivanov、B.Kreuter、A.Marcedone、H.B.McMahan、S.Patel、D.Ramage、A.Segal和K.Seth，“保护隐私的机器学习的实用安全聚合”，载于2017年ACM SIGSAC计算机和通信安全会议论文集，第1175-1191页，ACM，2017年。

[10] J.Konecnˇy、H.B.McMahan、F.X.Yu、P.Richtárik、A.T.Suresh和D.Bacon，“联合学习：提高沟通效率的策略”，arXiv预印本arXiv:1610.054922016年。

[11] 林永林、韩世善、毛海华、王永杰、戴利，「深度梯度压缩：减少分布式训练的通讯频宽」，arXiv预印本arXiv:1712.018872017年。

[12] J.Duchi、E.Hazan和Y.Singer，“在线学习和随机优化的自适应次梯度方法”，机器学习研究杂志，第12卷，7月号，第2121–2159页，2011年。

[13] Tieleman和G.Hinton，“将梯度除以最近震级的平均值。《机器学习的神经网络》，技术代表，技术报告。在线提供：zh://https://。库塞拉。org/learn/neuralnetworks/tearch/YQHki/rmsprop-divide-the-gradient-by-a-running-average-ofits-recent-magnity（2017年4月21日访问）。

[14] D.Kinga和J.B.Adam，“随机优化方法”，国际学习表征会议（ICLR），2015年。

[15] L.Bottou，“随机梯度下降的大规模机器学习”，载《COMPSTAT学报》，第177-186页，Springer，2010。

[16] A.Rakhlin，O.Shamir，K.Sridharan，et al.，“使梯度下降为强凸随机优化的最优解”，载ICML，Citeseer，2012。

[17] J.Dean，G.Corrado，R.Monga，K.Chen，M.Devin，M.Mao，A.Senior，P.Tucker，K.Yang，Q.V.Le等，“大规模分布式深层网络”，神经信息处理系统进展，第1223-12312012页。

[18] S.Ghadimi和G.Lan，“非凸随机规划的随机一阶和零阶方法”，暹罗优化杂志，第23卷，第4期，第2341-23682013页。

[19] T.M.Chilimbi、Y.Suzue、J.Apacible和K.Kalyanaraman，“adam项目：构建高效和可扩展的深度学习培训系统”，OSDI，第14卷，第571-582页，2014年。

[20] B.Recht，C.Re，S.Wright，和F.Niu，“Hogwild:并行化随机梯度下降的无锁方法”，神经信息处理系统进展，第693–701页，2011年。

[21]J.Duchi、M.I.Jordan和B.McMahan，“数据稀疏时的估计、优化和并行性”，神经信息处理系统进展，第2832-2840页，2013年。

[22]R.Leblond、F.Pedregosa和S.Lacoste Julien，“Asaga:异步并行传奇”，arXiv预印本arXiv:1606.048092016年。

[23]S.J.Reddi、A.Hefny、S.Sra、B.Poczos和A.J.Smola，“随机梯度下降及其异步变量的方差减少”，神经信息处理系统进展，第2647–2655页，2015年。

[24]C.M.De Sa，C.Zhang，K.Olukotun和C.Ré，“驯服野生动物：霍格沃德式算法的统一分析”，神经信息处理系统进展，第2674–26822015页。

[25]H.Mania，X.Pan，D.Papailiopoulos，B.Recht，K.Ramchandran和M.I.Jordan，“异步随机优化的扰动迭代分析”，arXiv预印本arXiv:1507.06970，2015年。

[26]J.Chen、X.Pan、R.Monga、S.Bengio和R.Jozefowicz，“重温分布式同步sgd”，arXiv预印本arXiv:1604.009812016年。

[27]孟国伟，陈文华，王永明，马志明，刘国英，“带洗牌的分布随机梯度下降的收敛性分析”，arXiv预印本arXiv:1709.104322017年。

[28]V.Smith，C.-K.Chiang，M.Sanjabi，和A.S.Talwalkar，“联合多任务学习”，神经信息处理系统进展，第4427-4437页，2017年。

[29]P.Warden，“语音命令：有限词汇语音识别的数据集”，arXiv预印本arXiv:1804.032092018年。

[30]A.Krizhevsky，I.Sutskever，G.E.Hinton，“深度卷积神经网络的图像网络分类”，神经信息处理系统进展，第1097-11052012页。

[31]B.Graham，“分数最大池”，arXiv预印本arXiv:1412.60712014年。

[32]I.J.Goodfello、O.Vinyals和A.M.Saxe，“定性描述神经网络优化问题”，arXiv预印本arXiv:1412.65442014年。

# 附录

A、 1.针对较小的批量，在沟通回合中测试精确度

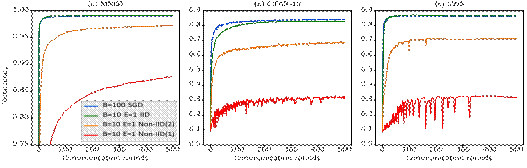


图8：使用（a）MNIST（b）CIFAR-10和（c）KWS数据集的IID和非IID数据集与SGD相比的通信轮的测试精度。非IID（2）表示2类非IID，非IID（1）表示1类非IID。*费达夫*

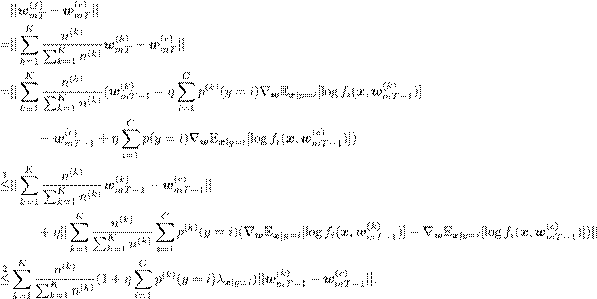
A、 2使用IID或非IID数据测试SGD的准确性*费达夫*

表3:SGD和IID或非IID数据的测试精度。*费达夫*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 培训 | B | E | 最小值（%） | CIFAR-10（%） | 千瓦时（%） |
| 费达夫IID | 大的 | 1 | 98.69 | 80.83 | 84.82 |
| FedAvg非IID（2） | 大的 | 1 | 96.29 | 67.00 | 72.30 |
| FedAvg非IID（1） | 大的 | 1 | 92.17 | 43.85 | 40.82 |
| FedAvg非IID（1） | 大的 | 5 | 91.92 | 44.40 | 40.84 |
| 预先培训的非IID（1） | 大的 | 1 | 96.19 | 61.72 | 63.58 |
| 费达夫IID | 小的 | 1 | 99.12 | 82.62 | 86.64 |
| FedAvg非IID（2） | 小的 | 1 | 97.24 | 68.53 | 71.21 |
| FedAvg非IID（1） | 小的 | 1 | 87.70 | 32.83 | 31.78 |

A、 3命题证明3.1

*证据。*基于w的定义*公吨*（六）我们有



这里，不等式1因为每堂课，

i、 e.所有客户的数据分布与整个人群的数据分布相同。不等式2成立是因为我们假设∇E | y[logfi（）]是| y-Lipschitz。*w十*=我*十,wλ十*=我

就客户而言*k*∈[K]，我们有

||（k） −1–（c）−1||*w公吨w公吨*

*C*

=| |（k）−2−ηXp（k）（y=i）∇E | y=i[logf（（k）−2）]*w公吨w十我十,w公吨*

*我*=1

*C*

−−2+η（y=i）∇E | y【logfi（−2）】||*w公吨*（三）十*pw十*=我*十,w公吨*（k）

*我*=1

*C*

≤| |（k）−2−（c）−2 | |+η| | Xp（k）（y=i）∇E | y=i[logf（（k）−2）]*w公吨w公吨w十我十,w公吨*

*我*=1

*C*

                    X（c）

              −p（y=i）∇E[logfi（−2）]||*w十*|*是的*=我*十,w公吨*

*我*=1

*C*

≤| | 3−2−2 |+η| |（y=i）（∇E | y[logfi（−2）]-∇E | y[logfi（−2）]）||*w公吨*（k）*w公吨*（三）十*p*（k）*w十*=我*十,w公吨*（k）*w十*=我*十,w公吨*（三）

*我*=1

*C*

+η|（p（k）（y=i）−p（y=i））∇E | y[logfi（−2）]||十*w十*=我*十,w公吨*（三）

*我*=1

*C C*

4（1+ηXp（k）（y=i）L | y=i）| |（k）−2−（c）−2 | |+ηg（（c）−2）X | | p（k）（y=i）−p（y=i）| |。*十w公吨w公吨最大值w公吨*

*我*=1 i=1

（三）

在这里，不等式3成立是因为

*C C*

||（y=i）∇E | y[logfi（−2）]—（y=i）∇E | y[对数坐标系（−2）]||十*p*（k）*w十*=我*十,w公吨*（k）十*pw十*=我*十,w公吨*（三）

*我*=1 i=1

*C C*

=| |（y=i）∇E | y[logfi（−2）]—（y=i）∇E | y[对数系数（−2）]+十*p*（k）*w十*=我*十,w公吨*（k）十*p*（k）*w十*=我*十,w公吨*（三）

*我*=1 i=1

*C C*

十*p*（k）（y=i）∇E | y[logfi（−2）]—（y=i）∇E | y[对数坐标系（−2）]||*w十*=我*十,w公吨*（三）十*pw十*=我*十,w公吨*（三）

*我*=1 i=1

*C*

≤|（y=i）（∇E | y[logfi（−2）]–∇E | y[logfi（−2）]）||十*p*（k）*w十*=我*十,w公吨*（k）*w十*=我*十,w公吨*（三）

*我*=1

*C*

+| |（p（k）（y=i）−p（y=i））∇E | y[对数系数（−2）]| |。十*w十*=我*十,w公吨*（三）

*我*=1

不等式4因为。

根据式（3），通过归纳，我们得到

||（k） −1–（c）−1||*w公吨w公吨*

*C*

≤a（k）| |（k）−2−（c）−2 |+ηg（（c）−2）X | | p（k）（y=i）−p（y=i）||*w公吨w公吨最大值w公吨*

*我*=1

*C*

≤（a（k））2 | |（k）−3−（c）−3 | |+ηX | | p（y=i）−p（y=i）|（g（（c）−2）+a（k）（（c）−3））*w公吨w公吨最大值w公吨g最大值w公吨*

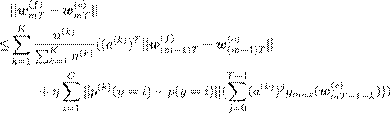
*我*=1

≤（a（k））−1 | |（（mk）−1）T−（（mc）−1）T | |+ηX=1C | | p（k）（y=i）−p（y=i）|（X=0−2（a（k））（（c）−2−j））*Tww我Tjjg最大值w公吨*

十

=（a（k））−1 | |（（mf）−1）T−（（mc）−1）T | |+ηX=1C | | p（k）（y=i）−p（y=i）|（=0−2（a（k））（（c）−2−j））*Tww我Tj公司jg最大值w公吨*

因此，

*.*

因此证明了这一点。

