联合学习：

挑战、方法和未来方向

                                                               田莉阿尼特·库马尔·萨胡

                                              卡内基梅隆大学博世人工智能中心

                                                       tianli@cmu.edu阿尼特·萨胡@gmail.com网站

                                                      弗吉尼亚史密斯

                                卡内基梅隆大学&卡内基梅隆大学

                                                   talwalkar@cmu.edusmithv@cmu.edu

# 摘要

联合学习包括在远程设备或孤立的数据中心（如移动电话或医院）上训练统计模型，同时保持数据本地化。异构和潜在的大规模网络中的培训带来了新的挑战，这些挑战要求从根本上背离大规模机器学习、分布式优化和隐私保护数据分析的标准方法。在本文中，我们将讨论联合学习的独特特征和挑战，提供当前方法的广泛概述，并概述与广泛研究社区相关的未来工作的几个方向。

# 1         介绍

移动电话、可穿戴设备和自动驾驶车辆只是现代分布式网络中每天产生大量数据的一小部分。由于这些设备的计算能力不断增长，再加上对传输私人信息的担忧，在本地存储数据并将网络计算推向边缘越来越有吸引力。

边缘计算并不是一个新概念。事实上，在分布式、低功耗设备上计算简单查询是一个长达数十年的研究领域，在传感器网络的查询处理、边缘计算和雾计算[12,29,40,49,74]的范围内进行了探索。最近的研究还考虑集中训练机器学习模型，但在本地服务和存储它们；例如，这是移动用户建模和个性化的常用方法[60，90]。

然而，随着分布式网络中设备的存储和计算能力的增长，可以在每个设备上利用增强的本地资源。这引起了人们对联合学习的兴趣[75]，它探索直接在远程设备上训练统计模型[1]。正如我们在本文中讨论的，在这样的环境中学习与传统的分布式环境有很大的不同

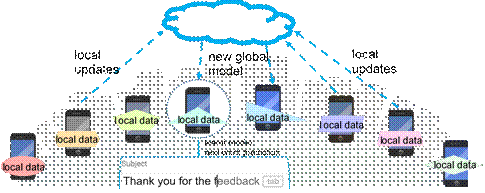


图1：联合学习在手机下一个单词预测任务中的应用示例。为了保护文本数据的隐私性并减轻网络压力，我们寻求以分布式方式训练预测器，而不是将原始数据发送到中央服务器。在此设置中，远程设备定期与中央服务器通信以了解全局模型。在每一轮通信中，所选电话的一个子集对其非相同分布的用户数据进行本地训练，并将这些本地更新发送到服务器。合并更新后，服务器将新的全局模型发送回另一个设备子集。这个迭代训练过程在整个网络中继续进行，直到达到收敛或满足某个停止准则为止。

在隐私、大规模机器学习和分布式优化等领域取得了根本性进展，并在不同领域的交叉点提出了新问题，如机器学习和系统[91]。

联合学习方法已被主要服务提供商部署[11，124]，并在支持隐私敏感应用程序方面发挥着关键作用，其中培训数据分布在边缘[例如，5,46,51,89,105,127,139]。潜在应用的例子包括：学习情绪、语义位置或移动电话用户的活动；适应自动驾驶车辆中的行人行为；预测可穿戴设备引发的心脏病等健康事件[6,52,84]。下面我们将讨论联合学习的几种典型应用：

•通过在大量手机中联合学习用户行为，统计模型可以为下一个单词预测、人脸检测和语音识别等应用提供动力[46，89]。然而，为了保护个人隐私或节省有限的带宽/电池电量，用户可能不愿意共享数据。联合学习有可能在智能手机上启用预测功能，而不会减少用户体验或泄露私人信息。图1描述了一个这样的应用程序，其中我们的目标是根据用户的历史文本数据在大规模移动电话网络中学习下一个单词预测器[46]。*智能手机。*

•在联合学习的背景下，组织或机构也可以被视为“设备”。例如，医院是一个组织，它包含大量用于预测医疗保健的患者数据。然而，当地医院可能会面临严格的隐私限制。对于这些应用程序来说，联合学习是一个很有前途的解决方案[52]，因为它可以减轻网络压力，并在各种设备/组织之间实现私有学习。*组织。*

•现代物联网网络，如可穿戴设备、自主车辆或智能家居，可能包含大量传感器，允许它们实时收集、响应和适应传入数据。例如，自动驾驶车辆的车队可能需要交通、建筑或行人行为的最新模型才能安全运行。然而，由于数据的私有性质和每个设备的有限连接性，在这些场景中构建聚合模型可能很困难。联合学习方法可以帮助训练模型，有效地适应这些系统中的变化，同时维护用户隐私[84，98]。*物联网。*

## 1.1        问题表述

规范的联合学习问题涉及到从存储在数千万甚至数百万个远程设备上的数据学习单个全局统计模型。我们的目标是在设备生成的数据被本地存储和处理的约束下学习这个模型，而只有中间更新被周期性地与中央服务器通信。特别是，目标通常是最小化以下目标函数：

*米*

最小F（），其中F（）：=∑（）。（一）*wwpkFk公司ww*

*k*=1

这里，m是设备总数，pk≥0和∑k pk=1，Fk是kth装置的局部目标函数。局部目标函数通常被定义为对局部数据的经验风险，即Fk fjk（；xjk，yjk），其中nk是本地可用的样本数。用户定义的术语pk指定每个设备的相对影响，有两个自然设置为pk或pk，其中n=∑k nk是样本总数。我们将在整篇文章中引用问题（1），但是，正如下面讨论的，我们注意到其他目标或建模方法可能是适当的，这取决于感兴趣的应用。*w*=*n*1 =*nnk*

## 1.2        核心挑战

接下来，我们将描述与解决（1）中提出的分布式优化问题相关的四个核心挑战。这些挑战使得联邦环境有别于其他经典问题，如数据中心环境中的分布式学习或传统的私有数据分析。

**沟通挑战：1。**通信是联邦网络中的一个关键瓶颈，再加上发送原始数据的隐私问题，使得在每个设备上生成的数据保持本地。事实上，联邦网络可能由大量的设备组成，例如数百万的智能手机，网络中的通信速度可能比本地计算慢许多数量级[50115]。为了使模型与联邦网络中的设备生成的数据相匹配，因此有必要开发通信效率高的方法，在训练过程中反复发送小消息或模型更新，而不是通过网络发送整个数据集。为了进一步减少这种情况下的通信，需要考虑两个关键方面：（i）减少通信轮的总数，或者（ii）减少每轮发送的消息的大小。

**挑战2：系统异构性。**由于硬件（CPU、内存）、网络连接（3G、4G、5G、wifi）和电源（电池电量）的变化，联邦网络中每个设备的存储、计算和通信能力可能有所不同。另外，每个设备上的网络大小和与系统相关的约束通常只导致一小部分设备同时处于活动状态，例如，百万个设备网络中有数百个活动设备[11]。每个设备也可能是不可靠的，并且由于连接或能量限制，活动设备在给定迭代中退出的情况并不少见。

这些系统级特性极大地加剧了诸如分散器缓解和容错等挑战。因此，开发和分析的联邦学习方法必须：（i）预期低参与量，（ii）容忍异构硬件，以及（iii）对网络中丢弃的设备具有鲁棒性。

**挑战3：统计异质性。**设备经常在网络上以非同一分布的方式生成和收集数据，例如，移动电话用户在下一个单词预测任务的上下文中使用了不同的语言。此外，跨设备的数据点的数量可能显著变化，并且可能存在一个底层结构，其捕获设备之间的关系及其相关分布。这种独立的概率分析方法，可能会增加分布式模型的复杂度。事实上，尽管（1）的规范化联合学习问题旨在学习单个全局模型，但也存在其他选择，例如通过多任务学习框架同时学习不同的局部模型[cf.106]。在这方面，元学习和联合学习有着密切的联系。多任务和元学习视角都支持个性化或特定于设备的建模，这通常是处理数据统计异构性的更自然的方法。

**隐私挑战：4。**最后，隐私通常是联合学习应用程序中的一个主要问题。联邦学习通过共享模型更新（例如梯度信息）而不是原始数据，朝着保护每个设备上生成的数据迈出了一步[17，31，33]。然而，在整个训练过程中进行模型更新的通信仍然会向第三方或中央服务器泄露敏感信息[76]。虽然最近的方法旨在使用诸如安全多方计算或差分隐私等工具来增强联合学习的隐私性，但这些方法通常以降低模型性能或系统效率为代价来提供隐私。在理论上和实践上，理解和平衡这些权衡是实现私有联合学习系统的一个相当大的挑战。

本文的其余部分组织如下。在第2节中，我们介绍了以前和现在的工作，这些工作旨在解决联合学习的四个已讨论的挑战。在第三部分，我们概述了未来研究的几个有希望的方向。

# 2         现状调查及相关工作

乍一看，联合学习的挑战类似于隐私、大规模机器学习和分布式优化等领域的经典问题。例如，在机器学习、优化和信号处理领域，已经提出了许多方法来解决昂贵的通信问题。然而，这些方法通常无法完全处理联邦网络的规模，更不用说系统和统计异构性的挑战了。类似地，虽然隐私是许多机器学习应用程序的一个重要方面，但是由于数据的统计变化，联邦学习的隐私保护方法可能很难严格断言，由于每个设备和潜在的大规模网络上的系统限制，实现起来可能更加困难。在本节中，我们将更详细地探讨第1节中提出的挑战，包括对经典结果的讨论，以及更近期专门关注联合学习的工作。

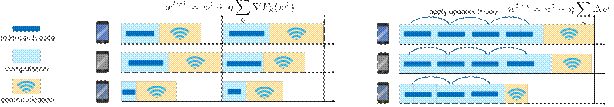


图2：左：分布式（小批量）SGD。每个设备k本地计算从一个小批量数据点到大约∇（）的梯度，并在服务器上应用聚合的小批量更新。右：本地更新方案。每个设备在计算完本地更新（例如，渐变）后立即应用它们，并且服务器在可变数量的本地更新之后执行全局聚合。*Fk公司w*

本地更新方案可以通过在本地执行额外的工作来减少通信。

## 2.1        沟通效率

通信是开发联邦网络方法时要考虑的一个关键瓶颈。虽然本文不在本文的讨论范围之内，但我们指出了几个基本方向，我们将这些方法分为（1）局部更新方法，（2）压缩方案，和（3）分散训练。

### 2.1.1       本地更新

小批量优化方法涉及到扩展经典随机方法来同时处理多个数据点，已成为数据中心环境中分布式机器学习的流行范例[28,88,96,102,103]。然而，在实践中，它们已经被证明在适应通信计算权衡方面的灵活性有限，这将最大限度地利用分布式数据处理[107108]。作为回应，最近提出了几种方法来提高分布式环境中的通信效率，允许在每个通信轮并行地在每台机器上应用可变数量的本地更新，使得计算量与通信量大体上更加灵活。对于凸目标，分布式局部更新原对偶方法已成为解决此类问题的一种流行方法[54,62,72,107,128]。这些方法利用对偶结构有效地将全局目标分解为子问题，这些子问题可以在每个通信轮并行解决。此外，还提出了几种分布式局部更新原始方法，这些方法具有适用于非凸目标的额外好处[93，136]。在实际应用中，这些方法极大地提高了性能，并且在实际数据中心环境中，与传统的小批量方法或分布式方法（如ADMM[14]）相比，可以实现数量级的加速。我们在图2中提供了本地更新方法的直观说明。

在联邦设置中，允许灵活的本地更新和低客户端参与的优化方法已成为事实上的解决方案[65,75,106]。联邦学习最常用的方法是联邦平均（FedAvg）[75]，这是一种基于原始问题的局部随机梯度下降（SGD）更新的平均方法。FedAvg已经被证明在经验上工作得很好，特别是对于非凸问题，但是没有收敛保证，并且在实际环境中，当数据是异构的时，可能会出现分歧[65]。我们将在第2.3.2节中更详细地讨论处理这种统计异质性的方法。

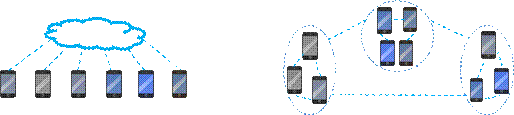


图3：集中式与分散式拓扑。在典型的联邦学习设置中，作为本文的重点，我们假设一个星形网络（左），其中服务器与所有远程设备连接。当与服务器的通信成为瓶颈时，分散拓扑（右）是一种潜在的替代方案。

### 2.1.2       压缩方案

虽然局部更新方法可以减少通信轮的总数，但是稀疏化、子抽样和量化等模型压缩方案可以显著地减少每轮通信的消息大小。在以前的数据中心环境中的分布式培训文献中，这些方法已经在经验和理论上得到了广泛的研究；我们将读者推迟到[119135]进行更全面的回顾。在联邦环境中，设备的低参与度、非均匀分布的本地数据和本地更新方案对这些模型压缩方法提出了新的挑战。例如，经典分布式学习[101]中常用的错误补偿技术不能直接扩展到联邦设置，因为如果不经常对设备进行采样，本地累积的错误可能是过时的。然而，有几项工作在联邦环境下提供了实用的策略，例如强制更新模型为稀疏低秩；使用结构化随机旋转执行量化[59]；使用有损压缩和丢失来减少服务器到设备的通信[15]；以及应用Golomb无损编码[99]。从理论的角度来看，虽然先前的工作已经探索了在存在非均匀分布数据的情况下，通过低精度训练的收敛保证[例如，111]，但所做的假设没有考虑到联邦环境的共同特征，如低设备参与度或局部更新优化方法。

### 2.1.3       分散式培训

在联邦学习中，星形网络（中央服务器连接到设备网络，如图3的左面板所示）是主要的通信拓扑；因此，本文将重点讨论星形网络设置。然而，我们简要地讨论了作为一种潜在的替代方案的分散拓扑（其中设备只与相邻设备通信，如图3的右面板）。在数据中心环境中，在低带宽或高延迟的网络上运行时，分散式培训比集中式培训快；我们让读者参考[47，67]进行更全面的回顾。类似地，在联邦学习中，分散算法理论上可以降低中央服务器上的高通信成本。最近的一些工作[47，61]研究了具有局部更新方案的异构数据的分散训练。然而，它们要么局限于线性模型[47]，要么假设设备完全参与[61]。最后，还提出了分层通信模式[68，70]，以进一步减轻中央服务器的负担，首先利用边缘服务器聚合来自边缘设备的更新，然后依赖云服务器来聚合来自边缘服务器的更新。虽然这是减少通信的一种有希望的方法，但它并不适用于所有网络，因为这种类型的物理层次结构可能不存在或事先已知。

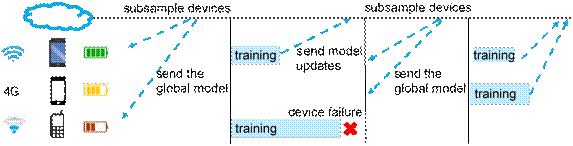


图4：联合学习中的系统异构性。设备在网络连接、电源和硬件方面可能有所不同。此外，在训练过程中，有些设备可能随时掉落。因此，联合训练方法必须能够容忍异构系统环境和设备的低参与度，也就是说，它们必须只允许一小部分设备在每轮中处于活动状态。

## 2.2        系统异构性

在联邦设置中，由于设备在硬件、网络连接性和电池电量方面可能有所不同，所以整个网络的系统特性存在显著差异。如图4所示，这些系统特性使得诸如离散器之类的问题比在典型的数据中心环境中更为普遍。我们大致将处理系统异构性的几个关键方向分为：

（i） 异步通信，（ii）有源器件采样，（ii）容错。如第2.1.3节所述，我们在下面的讨论中假设一个星形拓扑。

### 2.2.1       异步通信

在传统的数据中心环境中，同步和异步方案都被用来并行化迭代优化算法，每种方法各有利弊。同步方案很简单，可以保证一个串行等效的计算模型，但是在器件变化的情况下，它们也更容易受到离散器的影响。在异构环境中，特别是在共享内存系统中，异步方案是一种很有吸引力的方法来减少离散器[27,30,48,92,141]。然而，它们通常依赖于有界延迟假设来控制过时程度，对于设备k，这取决于自从设备k从中央服务器拉出来之后已经更新的其他设备的数量。虽然异步参数服务器在分布式数据中心（例如27、48、141）中已经取得了成功，但在联邦设置中，传统的有界延迟假设可能是不现实的，在联邦设置中，延迟可能是小时到几天，或者完全没有限制。

### 2.2.2       主动采样

在联合网络中，通常只有一小部分设备参加每一轮训练。然而，绝大多数联邦方法，例如在[11，47，65，75，106]中描述的方法是被动的，因为它们不旨在影响哪些设备参与。另一种方法是在每一轮中积极选择参与的设备。例如，Nishio和Yonetani[83]基于系统资源探索新的设备采样策略，目的是让服务器在预定义的时间窗口内尽可能多地聚合设备更新。同样，Kang等人。[57]在设计激励机制以鼓励具有高质量数据的设备参与学习过程时，考虑到每个设备上产生的系统开销。然而，这些方法假设了网络系统特性的静态模型；如何扩展这些方法来处理计算和通信延迟中的实时、设备特定的波动仍然是个未知数。此外，虽然这些方法主要关注系统的可变性以执行主动采样，但我们注意到，基于

基本统计结构。

### 2.2.3       容错

容错已经在系统界得到了广泛的研究，并且是经典分布式系统的一个基本考虑因素[19，71，110]。最近的研究还专门针对数据中心环境中的机器学习工作负载研究了容错性[例如，87112]。然而，当通过远程设备学习时，容错变得更为关键，因为在给定的训练迭代完成之前，一些参与设备在某个点退出是很常见的[11]。一种实际的策略是简单地忽略此类设备故障[11]，如果故障设备具有特定的数据特征，则可能会在设备采样方案中引入偏差。例如，由于网络连接不良，来自远程区域的设备可能更容易掉线，因此经过训练的联邦模型将偏向于具有良好网络条件的设备。理论上，虽然最近的一些研究工作已经研究了联邦学习方法变体的收敛保证[56，123，131，132]，但很少有分析允许低参与度[例如，65，106]，或直接研究掉落设备的影响。

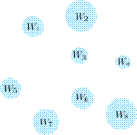
*编码计算*是通过引入算法冗余来容忍设备故障的另一种选择。最近的研究探索了使用代码来加速分布式机器学习训练[例如，20,21,63,94,109]。例如，在存在离散器的情况下，梯度编码及其变体[20，21，109]在计算节点之间仔细地复制数据块（以及对这些数据块的梯度计算），以获得准确或不精确的真实梯度恢复。虽然这对于联邦环境来说似乎是一种很有前途的方法，但是这些方法在联邦网络中面临着根本的挑战，因为由于隐私限制和网络规模，跨设备共享数据/复制通常是不可行的。

## 2.3        统计异质性

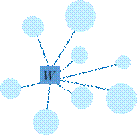
当从不同设备间不完全相同分布的数据训练联邦模型时，无论是在数据建模方面（如图5所示），还是在分析相关训练过程的收敛行为方面，都会出现挑战。我们将在下面这些方向讨论相关工作。

### 2.3.1       异构数据建模

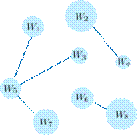
在机器学习方面，有大量的文献通过元学习[114]和多任务学习[18，37]等方法对统计异质性进行了建模；这些思想最近被扩展到联邦环境[24，26，35，58，106，138]。例如，MOCHA[106]，一个为联邦环境设计的优化框架，可以通过学习每个设备的独立但相关的模型，同时通过多任务学习利用共享表示来实现个性化。该方法对所考虑的目标具有可证明的理论收敛性保证，但其能力有限



（a） 学习每个设备的个性化模型；不要向同行学习。



（b） 学习全球模式；向同行学习。



（c） 学习每个设备的个性化模型；向同行学习。

图5：联邦网络中的不同建模方法。根据所关注的数据、网络和应用程序的属性，可以选择（a）为每个设备学习单独的模型，（b）将单个全局模型应用于所有设备，或（c）学习网络中相关但不同的模型。

可扩展到大规模网络，且仅限于凸目标。另一种方法[26]将星型拓扑建模为贝叶斯网络，并在学习过程中进行变分推理。虽然这种方法可以处理非凸模型，但推广到大型联邦网络的代价很高。Khodak等人。[58]可证明的元学习使用多任务信息（每个任务对应于一个设备）的任务内学习率，并证明了比普通FedAvg更好的经验性能。Eichner等人。[35]研究多元解决方案（在全局模型和设备特定模型之间进行自适应选择），以解决联合训练期间数据样本中的循环模式。赵等。[138]在集中训练一些共享代理数据的全局模型后，运行FedAvg，探索个性化传输学习。尽管最近取得了这些进展，关键的挑战仍然是如何在联邦环境中为异构建模提供健壮、可伸缩和自动化的方法。

在为联邦数据建模时，考虑超出准确性的问题也很重要，比如公平性。具体而言，天真地求解如（1）中的聚合损失函数可能隐含地对一些设备有利或不利，因为所学习的模型可能偏向于具有较大数据量的设备，或者（如果对设备进行相等的加权）偏向于常见的设备组。最近的工作提出了改进的建模方法，旨在减少模型性能在设备间的差异。一些启发式算法只是根据本地丢失执行不同数量的本地更新[52]。其他更具原则性的方法包括不可知论联合学习[80]，它通过极大极小优化方案，为客户分布的混合形成的任何目标分布优化集中式模型。Li等人采用了另一种更为普遍的方法。[66]提出了一个称为q-FFL的目标，在该目标中，具有较高损耗的设备被赋予更高的相对权重，以鼓励在最终精度分布中减少方差。除了公平问题之外，我们还注意到，诸如联合学习中的责任性和可解释性等方面也值得探讨，但由于网络的规模和异构性，可能会面临挑战。

### 2.3.2       非IID数据的收敛保证

统计异质性在分析联邦环境下的收敛行为方面也提出了新的挑战，即使在学习单个全局模型时也是如此。事实上，当数据在网络中不同的设备之间分布不同时，FedAvg等方法在实践中已经被证明是不同的[65，75]。并行SGD和相关变体使本地更新类似于FedAvg，已在I.I.D.设置中进行了分析[68、93、104、108、120、121、122、125、136、140]。然而，结果依赖于这样一个前提：每个局部解算器都是同一随机过程的副本（由于I.I.D.假设），

在典型的联邦设置中则不是这样。为了了解FedAvg在统计异构环境中的性能，FedProx[65]最近被提出。FedProx对FedAvg方法做了一个小的修改，以确保收敛性，无论是在理论上还是在实践中。DAVOX也可以被解释为实际系统的一个泛化版本。其他几项研究[56，123，131，132]也探讨了在存在具有不同假设的异构数据时的收敛保证，例如凸性[123]或一致有界梯度[131]。也有一些启发式方法旨在解决统计上的异构性，通过共享本地设备数据或一些服务器端代理数据[52，55，138]。然而，这些方法可能是不切实际的：除了对网络带宽施加负担之外，向服务器发送本地数据[55]违反了联邦学习的关键隐私假设，并且向所有设备发送全局共享的代理数据[52，138]需要努力仔细生成或收集此类辅助数据。

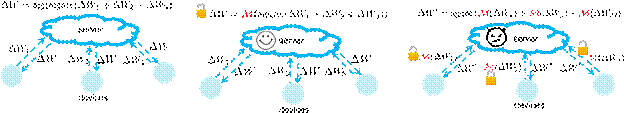
## 2.4        隐私

隐私问题常常促使人们需要将原始数据保存在联邦设置中的本地设备上。然而，在培训过程中共享其他信息（如模型更新）也可能泄露敏感的用户信息[8,17,39,78]。例如，卡里尼等人。[17] 证明可以从一个训练了用户语言数据的递归神经网络中提取敏感的文本模式，例如特定的信用卡号码。鉴于人们对隐私保护学习方法的兴趣越来越大，在第2.4.1节中，我们首先简要回顾了在一般（分布式）机器学习环境中增强隐私性的先前工作。然后，我们将在第2.4.2节中回顾最近专门为联邦设置设计的隐私保护方法。

### 2.4.1       机器学习中的隐私

隐私保护学习已经被机器学习[例如，76]、系统[例如，4,11]和理论[例如，38,69]社区广泛研究。我们将简要回顾三种主要策略，包括用差分隐私通信噪声数据草图、同态加密操作加密数据、安全函数评估或多方计算。

在这些不同的隐私方法中，差分隐私[32，33，34]由于其强大的信息理论保证、算法简单性和相对较小的系统开销而得到最广泛的应用。简单地说，如果一个输入元素的改变不会导致输出分布的太大差异，那么随机机制是有差别的私有的；这意味着我们无法得出在学习过程中是否使用特定样本的结论。这种样本级隐私可以在许多学习任务中实现[2,7,22,53,85,86]。对于基于梯度的学习方法，一种流行的方法是通过在每次迭代中随机扰动中间输出来应用差分隐私权[例如，2,7,126]。在应用扰动之前，例如，通过高斯噪声[2]、拉普拉斯噪声[77]或二项式噪声[3]，通常需要剪裁梯度，以限制每个示例对整体更新的影响。在差异隐私和模型精度之间存在一种内在的权衡，因为添加更多的噪声会导致更大的隐私，但可能会严重损害准确性。尽管差分隐私是机器学习中隐私的事实度量标准，但还有许多其他隐私定义，例如k-匿名性[36]、δ-存在性[81]和距离相关性[117]，它们可能适用于不同的学习问题[118]。



（a） 联合学习无需额外的隐私保护（b）全局隐私，其中可信（c）本地隐私，其中中央服务器机制。假定为服务器。可能是恶意的。

图6：一轮联合学习中不同隐私增强机制的说明。M表示用于私有化数据的随机机制。使用全局隐私（b），模型更新对除单个受信任方（中央服务器）以外的所有第三方都是私有的。使用本地隐私（c），单个模型更新也对服务器是私有的。

除了差分隐私外，同态加密还可以通过对加密数据的计算来保护学习过程，尽管它目前已应用于有限的设置中，例如训练线性模型[82]或仅涉及少数实体[133]。当敏感数据集分布在不同的数据所有者之间时，另一个自然的选择是通过安全函数评估（SFE）或安全多方计算（SMC）来执行隐私保护学习。由此产生的协议可以使多方协作计算商定的函数，而不泄漏任何一方的输入信息，除了可以从输出推断出的信息之外[例如，23、43、95]。因此，虽然SMC不能保证防止信息泄漏，但它可以与差分隐私相结合，以实现更强的隐私保障。然而，这些方法可能不适用于大规模的机器学习场景，因为它们会增加大量的通信和计算成本。此外，SMC协议需要为目标学习算法中的每个操作仔细设计和实现[25，79]。我们将感兴趣的读者推迟到[13，97]来更全面地回顾基于同态加密和SMC的方法。

### 2.4.2       联合学习中的隐私

联邦设置对现有的隐私保护算法提出了新的挑战。除了提供严格的隐私保障外，还需要开发计算成本低、通信效率高、能够容忍掉落设备的方法，而不必过度损害精度。虽然联邦学习中有各种隐私定义[8,17,41,64,76,113]，但通常可以将它们分为两类：全局隐私和本地隐私。如图6所示，全局隐私要求每轮生成的模型更新对除中央服务器之外的所有不受信任的第三方是私有的，而本地隐私进一步要求更新也是对服务器私有的。

当前旨在提高联合学习的隐私性的工作通常建立在以前的经典密码协议之上，如SMC[10,42]和差分隐私[3,8,41,76]。Bonawitz等人。[10] 引入SMC协议来保护单个模型更新。中央服务器无法看到任何本地更新，但仍然可以观察到每轮的确切聚合结果。SMC是一种无损的方法，可以在很高的隐私保证下保持原始的准确性。然而，由此产生的方法产生了显著的额外通信开销。其他著作[41，76]将差异隐私应用于联合学习，并提供全局差异隐私。尽管如此，仍需谨慎选择一些有助于降低超调精度的自适应策略。在需要更强隐私保障的情况下，Bhowmick等人。[8] 通过限制潜在对手的力量，引入一种放松的本地隐私。它提供了比全局隐私更强的隐私保证，并且比严格的本地隐私具有更好的模型性能。Li等人。[64]在元学习的背景下提出了局部差异私有算法，该算法可以应用于个性化的联合学习，同时在凸环境下也提供了可证明的学习保证。此外，差分隐私可以与模型压缩技术相结合，减少通信，同时获得隐私利益[3]。

# 3         未来方向

联合学习是一个活跃和持续的研究领域。虽然最近的工作已经开始解决第2节中讨论的挑战，但仍有一些关键的开放方向有待探讨。在本节中，我们简要概述了一些围绕前面讨论的挑战（昂贵的通信、系统异构性、统计异构性和隐私问题）的有前途的研究方向，并介绍了与联邦环境下的产品化和基准测试等问题有关的其他挑战。

•在联合学习中，需要多少沟通还有待观察。事实上，众所周知，机器学习的优化方法可以容忍缺乏精度；这个错误事实上有助于泛化[129]。虽然人们在传统的数据中心环境中探索了一次或分而治之的通信方案[73，137]，但在大规模或统计异构网络中，这些方法的行为并没有得到很好的理解。类似地，最近有人提出了一次/几次试探法[44，45，134]，但还没有从理论上进行分析或大规模评估。**极端通信方案。**

•我们讨论了在联合训练中减少通信的几种方法，例如本地更新和模型压缩。为了创建一个现实的联合学习系统，了解这些技术是如何相互结合的，并且系统地分析每种方法在准确性和交流性之间的权衡是很重要的。特别是，最有用的技术将证明在帕累托前沿的改进，在相同的通信预算下，实现比任何其他方法都更高的精度，并且理想情况下，在广泛的通信/精度配置文件中。为了有效的神经网络推理[例如，9]，已经进行了类似的综合分析，为了以有意义的方式比较用于联邦学习的通信简化技术，有必要进行这种分析。**通讯减少和帕累托前沿。**

• As discussed in Section 2.2.1, two communication schemes most commonly studied in distributed optimization are bulk synchronous approaches and asynchronous approaches (where it is assumed that the delay is bounded). These schemes are more realistic in data center settings—where worker nodes are typically dedicated to the workload, i.e., they are ready to &apos;pull&apos; their next job from the central node immediately after they &apos;push&apos; the results of their previous job. In contrast, in federated networks, each device is often undedicated to the task at hand and most devices are not active on any given iteration. Therefore, it is worth studying the effects of this more realistic device-centric communication scheme—in which each device can decide when to &apos;wake up&apos; and interact with the central server in an event-triggered manner.**Novel models of asynchrony.**

• Recent works have aimed to quantify statistical heterogeneity through metrics such as local dissimilarity (as defined in the context of federated learning in [65] and used for other purposes in works such as [100, 116, 130]) and earth mover&apos;s distance [138]. However, these metrics cannot be easily calculated over the federated network before training occurs. The importance of these metrics motivates the following open questions: (i) Do simple diagnostics exist to quickly determine the level of heterogeneity in federated networks a priori? (ii) Can analogous diagnostics be developed to quantify the amount of systems-related heterogeneity? (iii) Can current or new definitions of heterogeneity be exploited to further improve the convergence of federated optimization methods?**Heterogeneity diagnostics.**

• The definitions of privacy outlined in Section 2.4.2 cover privacy at a local or global level with respect to all devices in the network. However, in practice, it may be necessary to define privacy on a more granular level, as privacy constraints may differ across devices or even across data points on a single device. For instance, Li et al. [64] recently proposed sample-specific (as opposed to user-specific) privacy guarantees, thus providing a weaker form of privacy in exchange for more accurate models. Developing methods to handle mixed (device-specific or sample-specific) privacy restrictions is an interesting and ongoing direction of future work.**Granular privacy constraints.**

• It is important to note that the methods discussed thus far have been developed with the task of supervised learning in mind, i.e., they assume that labels exist for all of the data in the federated network. In practice, much of the data generated in realistic federated networks may be unlabeled or weakly labeled. Furthermore, the problem at hand may not be to fit a model to data as presented in (1), but instead to perform some exploratory data analysis, determine aggregate statistics, or run a more complex task such as reinforcement learning. Tackling problems beyond supervised learning in federated networks will likely require addressing similar challenges of scalability, heterogeneity, and privacy.**Beyond supervised learning.**

• Beyond the major challenges discussed in this article, there are a number of practical concerns that arise when running federated learning in production. In particular, issues such as concept drift (when the underlying data-generation model changes over time); diurnal variations (when the devices exhibit different behavior at different times of the day or**Productionizing federated learning.**

week) [35]; and cold start problems (when new devices enter the network) must be handled with care. We defer the readers to [11], which discusses some of the practical systems-related issues that exist in production federated learning systems.

• Finally, as federated learning is a nascent field, we are at a pivotal time to shape the developments made in this area and ensure that they are grounded in real-world settings, assumptions, and datasets. It is critical for the broader research communities to further build upon existing implementations and benchmarking tools, such as LEAF [16] and TensorFlow Federated [1], to facilitate both the reproducibility of empirical results and the dissemination of new solutions for federated learning.**Benchmarks.**

# 4         Conclusion

In this article, we have provided an overview of federated learning, a learning paradigm where statistical models are trained at the edge in distributed networks. We have discussed the unique properties and associated challenges of federated learning compared with traditional distributed data center computing and classical privacy-preserving learning. We provided an extensive survey on classical results as well as more recent work specifically focused on federated settings. Finally, we have outlined out a handful of open problems worth future research effort. Providing solutions to these problems will require interdisciplinary effort from a broad set of research communities.

**Acknowledgement.** We thank Jeffrey Li and Mikhail Khodak for helpful discussions and comments. This work was supported in part by DARPA FA875017C0141, the National Science Foundation grants IIS1705121 and IIS1838017, an Okawa Grant, a Google Faculty Award, an Amazon Web Services Award, a JP Morgan A.I. Research Faculty Award, a Carnegie Bosch Institute Research Award and the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA, the National Science Foundation, or any other funding agency.

# References

[1] Tensorflow federated: Machine learning on decentralized data. URL https://www.tensorflow.org/federated.

[2] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In Conference on Computer and Communications Security, 2016.

[3] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan. cpSGD: Communication-efficient and differentially-private distributed sgd. In Advances in Neural Information Processing Systems, 2018.

[4] R. Agrawal and R. Srikant. Privacy-preserving data mining. In International Conference on Management of Data, 2000.

[5] M. Ammad-ud din, E. Ivannikova, S. A. Khan, W. Oyomno, Q. Fu, K. E. Tan, and A. Flanagan. Federated collaborative filtering for privacy-preserving personalized recommendation system. arXiv preprint arXiv:1901.09888, 2019.

[6] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2013.

[7] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In Foundations of Computer Science, 2014.

[8] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers. Protection against reconstruction and its applications in private federated learning. arXiv preprint arXiv:1812.00984, 2018.

[9] T. Bolukbasi, J. Wang, O. Dekel, and V. Saligrama. Adaptive neural networks for efficient inference. In International Conference on Machine Learning, 2017.

[10] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth. Practical secure aggregation for privacy-preserving machine learning. In Conference on Computer and Communications Security, 2017.

[11] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecny, S. Mazzocchi, H. B. McMahan, T. V. Overveldt, D. Petrou, D. Ramage, and J. Roselander. Towards federated learning at scale: system design. In Conference on Systems and Machine Learning, 2019.

[12] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli. Fog computing and its role in the internet of things. In SIGCOMM Workshop on Mobile Cloud Computing, 2012.

[13] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser. Machine learning classification over encrypted data. In Network and Distributed System Security Symposium, 2015.

[14] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends R , 3:1–122, 2011.*in Machine Learning*

[15] S. Caldas, J. Konecˇny, H. B. McMahan, and A. Talwalkar. Expanding the reach of federated learning by reducing client resource requirements. arXiv preprint arXiv:1812.07210, 2018.

[16] S. Caldas, P. Wu, T. Li, J. Konecˇny, H. B. McMahan, V. Smith, and A. Talwalkar. Leaf: A benchmark for federated` settings. arXiv preprint arXiv:1812.01097, 2018.

[17] N. Carlini, C. Liu, J. Kos, Ú. Erlingsson, and D. Song. The secret sharer: Measuring unintended neural network memorization & extracting secrets. arXiv preprint arXiv:1802.08232, 2018.

[18] R. Caruana. Multitask learning. Machine Learning, 28:41–75, 1997.

[19] M. Castro, B. Liskov, et al. Practical byzantine fault tolerance. In Operating Systems Design and Implementation, 1999.

[20] Z. Charles and D. Papailiopoulos. Gradient coding using the stochastic block model. In International Symposium on Information Theory, 2018.

[21] Z. B. Charles, D. S. Papailiopoulos, and J. Ellenberg. Approximate gradient coding via sparse random graphs. arXiv preprint arXiv:1711.0677, 2017.

[22] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. Journal of Machine Learning Research, 12:1069–1109, 2011.

[23] D. Chaum. The dining cryptographers problem: Unconditional sender and recipient untraceability. Journal of Cryptology, 1:65–75, 1988.

[24] F. Chen, Z. Dong, Z. Li, and X. He. Federated meta-learning for recommendation. arXiv preprint arXiv:1802.07876, 2018.

[25] V. Chen, V. Pastro, and M. Raykova. Secure computation for machine learning with spdz. arXiv preprint arXiv:1901.00329, 2019.

[26] L. Corinzia and J. M. Buhmann. Variational federated multi-task learning. arXiv preprint arXiv:1906.06268, 2019.

[27] W. Dai, A. Kumar, J. Wei, Q. Ho, G. Gibson, and E. P. Xing. High-performance distributed ML at scale through parameter server consistency models. In AAAI Conference on Artificial Intelligence, 2015.

[28] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. Journal of Machine Learning Research, 13:165–202, 2012.

[29] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong. Model-based approximate querying in sensor networks. The VLDB Journal, 14:417–443, 2005.

[30] J. Duchi, M. I. Jordan, and B. McMahan. Estimation, optimization, and parallelism when data is sparse. In Advances in Neural Information Processing Systems, 2013.

[31] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. In Advances in Neural Information Processing Systems, 2012.

[32] C. Dwork. A firm foundation for private data analysis. Communications of the ACM, 54:86–95, 2011.

[33] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9:211–407, 2014.

[34] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography Conference, 2006.

[35] H. Eichner, T. Koren, H. B. McMahan, N. Srebro, and K. Talwar. Semi-cyclic stochastic gradient descent. In International Conference on Machine Learning, 2019.

[36] K. El Emam and F. K. Dankar. Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association, 15:627–637, 2008.

[37] T. Evgeniou and M. Pontil. Regularized multi–task learning. In Conference on Knowledge Discovery and Data Mining, 2004.

[38] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta. Privacy amplification by iteration. In Foundations of Computer Science, 2018.

[39] M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In Conference on Computer and Communications Security, 2015.

[40] P. Garcia Lopez, A. Montresor, D. Epema, A. Datta, T. Higashino, A. Iamnitchi, M. Barcellos, P. Felber, and E. Riviere. Edge-centric computing: Vision and challenges. SIGCOMM Computer Communication Review, 45:37–42, 2015.

[41] R. C. Geyer, T. Klein, and M. Nabi. Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557, 2017.

[42] B. Ghazi, R. Pagh, and A. Velingker. Scalable and differentially private distributed aggregation in the shuffled model. arXiv preprint arXiv:1906.08320, 2019.

[43] S. Goryczka and L. Xiong. A comprehensive comparison of multiparty secure additions with differential privacy. IEEE Transactions on Dependable and Secure Computing, 14:463–477, 2015.

[44] N. Guha and V. Smith. Model aggregation via good-enough model spaces. arXiv preprint arXiv:1805.07782, 2018. [45] N. Guha, A. Talwalkar, and V. Smith. One-shot federated learning. arXiv preprint arXiv:1902.11175, 2019.

[46] A. Hard, K. Rao, R. Mathews, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage. Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604, 2018.

[47] L. He, A. Bian, and M. Jaggi. Cola: Decentralized linear learning. In Advances in Neural Information Processing Systems, 2018.

[48] Q. Ho, J. Cipar, H. Cui, S. Lee, J. K. Kim, P. B. Gibbons, G. A. Gibson, G. Ganger, and E. P. Xing. More effective distributed ML via a stale synchronous parallel parameter server. In Advances in Neural Information Processing Systems, 2013.

[49] K. Hong, D. Lillethun, U. Ramachandran, B. Ottenwälder, and B. Koldehofe. Mobile fog: A programming model for large-scale applications on the internet of things. In SIGCOMM Workshop on Mobile Cloud Computing, 2013.

[50] J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, and O. Spatscheck. An in-depth study of lte: effect of network protocol and application behavior on performance. SIGCOMM Computer Communication Review, 43: 363–374, 2013.

[51] L. Huang and D. Liu. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. arXiv preprint arXiv:1903.09296, 2019.

[52] L. Huang, Y. Yin, Z. Fu, S. Zhang, H. Deng, and D. Liu. Loadaboost: Loss-based adaboost federated machine learning on medical data. arXiv preprint arXiv:1811.12629, 2018.

[53] R. Iyengar, J. P. Near, D. Song, O. Thakkar, A. Thakurta, and L. Wang. Towards practical differentially private convex optimization. In Conference on Computer and Communications Security, 2019.

[54] M. Jaggi, V. Smith, M. Takác, J. Terhorst, S. Krishnan, T. Hofmann, and M. I. Jordan. Communication-efficient distributed dual coordinate ascent. In Advances in Neural Information Processing Systems, 2014.

[55] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. arXiv preprint arXiv:1811.11479, 2018.

[56] P. Jiang and G. Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. In Advances in Neural Information Processing Systems, 2018.

[57] J. Kang, Z. Xiong, D. Niyato, H. Yu, Y.-C. Liang, and D. I. Kim. Incentive design for efficient federated learning in mobile networks: A contract theory approach. arXiv preprint arXiv:1905.07479, 2019.

[58] M. Khodak, M.-F. Balcan, and A. Talwalkar. Adaptive gradient-based meta-learning methods. arXiv preprint arXiv:1906.02717, 2019.

[59] J. Konecˇny, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: strategies for` improving communication efficiency. arXiv preprint arXiv:1610.05492, 2016.

[60] T. Kuflik, J. Kay, and B. Kummerfeld. Challenges and solutions of ubiquitous user modeling. In Ubiquitous Display Environments. 2012.

[61] A. Lalitha, X. Wang, O. Kilinc, Y. Lu, T. Javidi, and F. Koushanfar. Decentralized bayesian learning over graphs. arXiv preprint arXiv:1905.10466, 2019.

[62] C.-P. Lee and D. Roth. Distributed box-constrained quadratic optimization for dual linear svm. In International Conference on Machine Learning, 2015.

[63] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran. Speeding up distributed machine learning using codes. IEEE Transactions on Information Theory, 64:1514–1529, 2017.

[64] J. Li, M. Khodak, S. Caldas, and A. Talwalkar. Differentially-private gradient-based meta-learning. Technical Report, 2019.

[65] T. Li, A. K. Sahu, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith. Federated optimization for heterogeneous networks. arXiv preprint arXiv:1812.06127, 2018.

[66] T. Li, M. Sanjabi, and V. Smith. Fair resource allocation in federated learning. arXiv preprint arXiv:1905.10497, 2019.

[67] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In Advances in Neural Information Processing Systems, 2017.

[68] T. Lin, S. U. Stich, and M. Jaggi. Don&apos;t use large mini-batches, use local sgd. arXiv preprint arXiv:1808.07217, 2018. [69] Y. Lindell and B. Pinkas. Privacy preserving data mining. In Advances in Cryptology, 2000.

[70] L. Liu, J. Zhang, S. Song, and K. B. Letaief. Edge-assisted hierarchical federated learning with non-iid data. arXiv preprint arXiv:1905.06641, 2019.

[71] Y. Liu, J. K. Muppala, M. Veeraraghavan, D. Lin, and M. Hamdi. Data center networks: Topologies, architectures and fault-tolerance characteristics. Springer Science & Business Media, 2013.

[72] C. Ma, V. Smith, M. Jaggi, M. I. Jordan, P. Richtárik, and M. Takácˇ. Adding vs. averaging in distributed primal-dual optimization. In International Conference on Machine Learning, 2015.

[73] L. W. Mackey, M. I. Jordan, and A. Talwalkar. Divide-and-conquer matrix factorization. In Advances in Neural Information Processing Systems, 2011.

[74] S. R. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong. Tinydb: an acquisitional query processing system for sensor networks. Transactions on Database Systems, 30:122–173, 2005.

[75] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In Conference on Artificial Intelligence and Statistics, 2017.

[76] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private recurrent language models. In International Conference on Learning Representations, 2018.

[77] L. Melis, G. Danezis, and E. D. Cristofaro. Efficient private statistics with succinct sketches. In Network and Distributed System Security Symposium, 2016.

[78] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov. Exploiting unintended feature leakage in collaborative learning. In IEEE Symposium on Security & Privacy, 2019.

[79] P. Mohassel and P. Rindal. Aby 3: a mixed protocol framework for machine learning. In Conference on Computer and Communications Security, 2018.

[80] M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In International Conference on Machine Learning, 2019.

[81] M. E. Nergiz and C. Clifton. -presence without complete world knowledge. IEEE Transactions on Knowledge and Data Engineering, 22:868–883, 2010.*δ*

[82] V. Nikolaenko, U. Weinsberg, S. Ioannidis, M. Joye, D. Boneh, and N. Taft. Privacy-preserving ridge regression on hundreds of millions of records. In Symposium on Security and Privacy, 2013.

[83] T. Nishio and R. Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In International Conference on Communications, 2019.

[84] A. Pantelopoulos and N. G. Bourbakis. A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Transactions on Systems, Man, and Cybernetics, 40:1–12, 2010.

[85] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In International Conference on Learning Representations, 2017.

[86] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson. Scalable private learning with pate. In International Conference on Learning Representations, 2018.

[87] A. Qiao, B. Aragam, B. Zhang, and E. Xing. Fault tolerance in iterative-convergent machine learning. In International Conference on Machine Learning, 2019.

[88] Z. Qu, P. Richtárik, and T. Zhang. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In Advances in Neural Information Processing Systems, 2015.

[89] S. Ramaswamy, R. Mathews, K. Rao, and F. Beaufays. Federated learning for emoji prediction in a mobile keyboard. arXiv preprint arXiv:1906.04329, 2019.

[90] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In European Conference on Computer Vision, 2016.

[91] A. Ratner et al. SysML: The new frontier of machine learning systems. arXiv preprint arXiv:1904.03257, 2019.

[92] B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In Advances in Neural Information Processing Systems, 2011.

[93] S. J. Reddi, J. Konecˇny, P. Richtárik, B. Póczós, and A. Smola. Aide: Fast and communication efficient distributed` optimization. arXiv preprint arXiv:1608.06879, 2016.

[94] A. Reisizadeh, S. Prakash, R. Pedarsani, and A. S. Avestimehr. Coded computation over heterogeneous clusters. IEEE Transactions on Information Theory, 65:4227–4242, 2019.

[95] M. S. Riazi, C. Weinert, O. Tkachenko, E. M. Songhori, T. Schneider, and F. Koushanfar. Chameleon: A hybrid secure computation framework for machine learning applications. In Asia Conference on Computer and Communications Security, 2018.

[96] P. Richtárik and M. Takácˇ. Distributed coordinate descent method for learning with big data. Journal of Machine Learning Research, 17:2657–2681, 2016.

[97] B. D. Rouhani, M. S. Riazi, and F. Koushanfar. Deepsecure: Scalable provably-secure deep learning. In Design Automation Conference, 2018.

[98] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah. Federated learning for ultra-reliable low-latency v2v communications. In Global Communications Conference, 2018.

[99] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek. Robust and communication-efficient federated learning from non-iid data. arXiv preprint arXiv:1903.02891, 2019.

[100] M. Schmidt and N. L. Roux. Fast convergence of stochastic gradient descent under a strong growth condition. arXiv preprint arXiv:1308.6370, 2013.

[101] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In International Speech Communication Association, 2014.

[102] S. Shalev-Shwartz and T. Zhang. Accelerated mini-batch stochastic dual coordinate ascent. In Advances in Neural Information Processing Systems, 2013.

[103] O. Shamir and N. Srebro. Distributed stochastic optimization and learning. In Allerton Conference on Communication, Control, and Computing, 2014.

[104] O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In International Conference on Machine Learning, 2014.

[105] S. Silva, B. Gutman, E. Romero, P. M. Thompson, A. Altmann, and M. Lorenzi. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. arXiv preprint arXiv:1810.08553, 2018.

[106] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar. Federated multi-task learning. In Advances in Neural Information Processing Systems, 2017.

[107] V. Smith, S. Forte, C. Ma, M. Takac, M. I. Jordan, and M. Jaggi. Cocoa: a general framework for communicationefficient distributed optimization. Journal of Machine Learning Research, 18:1–47, 2018.

[108] S. U. Stich. Local sgd converges fast and communicates little. In International Conference on Learning Representations, 2019.

[109] R. Tandon, Q. Lei, A. G. Dimakis, and N. Karampatziakis. Gradient coding: Avoiding stragglers in distributed learning. In International Conference on Machine Learning, 2017.

[110] A. S. Tanenbaum and M. Van Steen. Distributed systems: principles and paradigms. Prentice-Hall, 2007.

[111] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu. Communication compression for decentralized training. In Advances in Neural Information Processing Systems, 2018.

[112] H. Tang, C. Yu, C. Renggli, S. Kassing, A. Singla, D. Alistarh, J. Liu, and C. Zhang. Distributed learning over unreliable networks. In International Conference on Machine Learning, 2019.

[113] O. Thakkar, G. Andrew, and H. B. McMahan. Differentially private learning with adaptive clipping. arXiv preprint arXiv:1905.03871, 2019.

[114] S. Thrun and L. Pratt. Learning to learn. Springer Science & Business Media, 2012.

[115] C. Van Berkel. Multi-core for mobile phones. In Conference on Design, Automation and Test in Europe, 2009.

[116] S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models (and an accelerated perceptron). In Conference on Artificial Intelligence and Statistics, 2019.

[117] P. Vepakomma, O. Gupta, A. Dubey, and R. Raskar. Reducing leakage in distributed deep learning for sensitive health data. arXiv preprint arXiv:1812.00564, 2019.

[118] I. Wagner and D. Eckhoff. Technical privacy metrics: a systematic survey. ACM Computing Surveys, 51:57, 2018.

[119] H. Wang, S. Sievert, S. Liu, Z. Charles, D. Papailiopoulos, and S. Wright. Atomo: Communication-efficient learning via atomic sparsification. In Advances in Neural Information Processing Systems, 2018.

[120] J. Wang and G. Joshi. Cooperative sgd: A unified framework for the design and analysis of communicationefficient sgd algorithms. arXiv preprint arXiv:1808.07576, 2018.

[121] J. Wang and G. Joshi. Adaptive communication strategies to achieve the best error-runtime trade-off in localupdate sgd. In Conference on Systems and Machine Learning, 2019.

[122] S. Wang, F. Roosta-Khorasani, P. Xu, and M. W. Mahoney. Giant: Globally improved approximate newton method for distributed optimization. In Advances in Neural Information Processing Systems, 2018.

[123] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan. Adaptive federated learning in resource constrained edge computing systems. Journal on Selected Areas in Communications, 37:1205–1221, 2019.

[124] WeBank AI Group. Federated learning white paper v1.0. 2018.

[125] B. Woodworth, J. Wang, A. Smith, B. McMahan, and N. Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. In Advances in Neural Information Processing Systems, 2018.

[126] X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In International Conference on Management of Data, 2017.

[127] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10:12, 2019.

[128] T. Yang. Trading computation for communication: Distributed stochastic dual coordinate ascent. In Advances in Neural Information Processing Systems, 2013.

[129] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. Constructive Approximation, 26:289–315, 2007.

[130] D. Yin, A. Pananjady, M. Lam, D. Papailiopoulos, K. Ramchandran, and P. Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In Conference on Artificial Intelligence and Statistics, pages 1998–2007, 2018.

[131] H. Yu, S. Yang, and S. Zhu. Parallel restarted sgd for non-convex optimization with faster convergence and less communication. In AAAI Conference on Artificial Intelligence, 2018.

[132] H. Yu, R. Jin, and S. Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In International Conference on Machine Learning, 2019.

[133] J. Yuan and S. Yu. Privacy preserving back-propagation neural network learning made practical with cloud computing. IEEE Transactions on Parallel and Distributed Systems, 25:212–221, 2013.

[134] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, T. N. Hoang, and Y. Khazaeni. Bayesian nonparametric federated learning of neural networks. In International Conference on Machine Learning, 2019.

[135] H. Zhang, J. Li, K. Kara, D. Alistarh, J. Liu, and C. Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In International Conference on Machine Learning, 2017.

[136] S. Zhang, A. E. Choromanska, and Y. LeCun. Deep learning with elastic averaging sgd. In Advances in Neural Information Processing Systems, 2015.

[137] Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. Journal of Machine Learning Research, 16:3299–3340, 2015.

[138] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra. Federated learning with non-iid data. arXiv preprint arXiv:1806.00582, 2018.

[139] Y. Zhao, J. Zhao, L. Jiang, R. Tan, and D. Niyato. Mobile edge computing, blockchain and reputation-based crowdsourcing iot federated learning: A secure, decentralized and privacy-preserving system. arXiv preprint arXiv:1906.10893, 2019.

[140] F. Zhou and G. Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. In International Joint Conference on Artificial Intelligence, 2018.

[141] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. In Advances in Neural Information Processing Systems, 2010.

[[1]](" \l "_ftnref1" \o ") We use the term &apos;device&apos; throughout the article to describe entities in the network, such as nodes, clients, sensors, or organizations.