

一种自适应系统规模的联邦深度学习方法

吴宾宾, 杨桂松*

上海理工大学光电信息与计算机工程学院, 上海

收稿日期: 2024年6月17日; 录用日期: 2024年7月10日; 发布日期: 2024年7月19日

摘要

在面向复杂任务协作的环境中, 通信带宽和计算资源的限制以及对隐私保护的需求共同构成了该研究领域的主要挑战。为了解决这些问题, 研究者提出了联邦学习(Federated Learning, FL)框架作为一种解决方案。FL允许多个设备在不直接交换原始数据的情况下进行协同模型训练, 从而降低了通信需求并保护了数据隐私。然而, 一些FL方法采用了全客户端参与的策略, 即所有客户端在每一轮中更新其本地模型。这种方法不仅增加了通信次数, 而且随着客户端规模的增大, 也会导致系统性能下降、响应延迟等问题。因此, 本文介绍了一种基于原始-对偶优化的新FL协议(Federated Deep Learning Alternating Direction Method of Multipliers, FDLADMM)。FDLADMM算法利用双变量来引导客户端进行本地训练, 减少了设备间的通信次数, 优化了模型训练速度, 并且随着系统规模增大, 无需进行超参数调整即可有效适应。通过实验, 本文展示了所提出的方法在通信效率和训练速度方面的优势, 并且当系统规模不断调整时, 无需进行超参数调整即可有效适应。这一创新的方法为应对复杂任务协作中的挑战提供了一种可行且高效的解决方案, 并有望在未来的研究和实践中得到广泛应用。

关键词

联邦学习(FL), 系统自适应, 原始-对偶优化, FDLADMM

A Federated Deep Learning Method for Adaptive System Scale

Binbin Wu, Guisong Yang*

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science & Technology, Shanghai

Received: Jun. 17th, 2024; accepted: Jul. 10th, 2024; published: Jul. 19th, 2024

Abstract

In the context of complex task-oriented collaboration, the limitations of communication band-

*通讯作者。

文章引用: 吴宾宾, 杨桂松. 一种自适应系统规模的联邦深度学习方法[J]. 建模与仿真, 2024, 13(4): 4507-4514.
DOI: 10.12677/mos.2024.134407

width and computing resources, as well as the need for privacy protection, together constitute major challenges in this research area. To solve these problems, researchers propose a framework of Federated Learning (FL) as a solution. FL allows multiple devices to perform collaborative model training without directly exchanging raw data, thereby reducing communication requirements and protecting data privacy. However, some FL methods adopt a full-client-participation strategy, where all clients update their local model in each round. This method not only increases the number of communication, but also leads to performance degradation and response delay with the increase of client size. Therefore, this paper introduces a new FL protocol (Federated Deep Learning Alternating Direction Method of Multipliers, FDLADMM) based on primitive-dual optimization. FDLADMM algorithm uses bivariate to guide the client to carry out local training, reduces the communication times between devices, optimizes the training speed of the model, and with the increase of the system size, it can adapt effectively without super parameter adjustment. Through experiments, the advantages of the proposed method in communication efficiency and training speed are demonstrated, and it can be adapted effectively without hyperparameter adjustment when the system size is adjusted continuously. This innovative approach provides a viable and efficient solution to the challenges of collaboration in complex missions and is expected to be widely applied in future research and practice.

Keywords

Federated Learning, System Adaptive, Primal-Dual Optimization, FDLADMM

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着数据在生活各个领域的大量积累, 机器学习(Machine Learning, ML)技术得到了迅速发展和广泛应用。在众多应用场景中, 数据以分布式的方式被大量生成, 例如通过众包、移动电话、自动驾驶汽车、医疗中心和智能电网中的分布式传感器等手段。大量的数据, 加之对个人隐私保护的迫切需求, 使得传统的分布式机器学习实践难以直接应用。数据中心分布式学习框架[1] [2], 常常涉及将原始数据在机器间转移, 这不仅违反了隐私限制, 而且在网络资源有限时处理大规模数据变得不可行。此外, 个人设备如智能手机的计算能力的迅速提升, 也促使计算从云端向边缘转移。

在这样的背景下, 联邦学习作为一个新范式被提出, 用以解决与数据中心环境相比具有显著不同特点的大规模分布式机器学习问题。FL 是一种机器学习的分布式学习方法, 旨在保护数据隐私的同时进行模型训练。在传统的集中式机器学习中, 所有数据被收集到一个中心服务器进行训练, 这可能会涉及大量用户数据的集中存储, 存在隐私泄露的风险。而在联邦学习中, 模型的训练过程分布在多个设备或智能体中进行, 每个设备或智能体只需在本地存储和处理自己的数据, 不需要将数据传输到中心服务器。联邦学习框架通过服务器与客户端协作计算全局 ML 模型, 以此来克服传统方法中存在的问题, 展示在图 1 中, 这个框架的介绍是在[3]中进行的, 其中作者提出联邦平均算法(FedAvg)作为主要算法。首先由中央服务器初始化一个全局模型, 并将其: ① 发送给所有参与的客户端。每个客户端接收到全局模型后, 会根据自己持有的数据进行, ② 本地训练, 更新模型参数, 并将更新后的信息, ③ 上传到服务器。最后, ④ 服务器通过平均更新模型得到一个新的全局模型。这个过程不断迭代, 直到全局模型的性能达到预期目标或者满足特定的终止条件。通过这种方式, 联邦学习能够在不直接共享原始数据的前提下, 实现跨多个客户端的协同

学习, 既保护了数据隐私, 又提高了模型的泛化能力。总之, 联邦学习作为一种创新的机器学习框架, 为处理分布式数据下的协作学习问题提供了有效的解决方案, 具有广泛的应用前景和重要的研究价值。

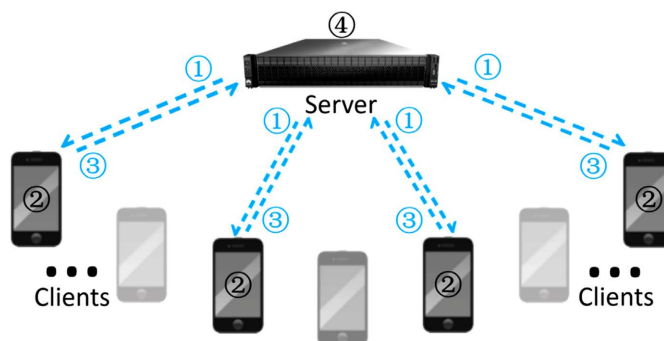


Figure 1. Federal learning framework

图 1. 联邦学习框架

尽管 FL 为数据隐私提供了更好的保护, 但也面临一些挑战, 如通信效率、训练速度等问题。因此, 本文提出一种新 FL 协议, FDLADMM 算法是一种基于原始 - 对偶优化的算法, 能够自动适应系统规模, 并相比现有技术显著提高收敛速度。FDLADMM 每轮涉及少数客户端, 并为每位用户提供丰富的计算选择, 减少了设备之间的通信次数, 从而优化了模型训练速度。同时使用双变量引导本地训练过程, 避免了客户端漂移问题, 即防止模型过度拟合于特定客户端的数据。

2. 相关工作

在联邦学习的研究领域, 尽管 FedAvg 因其简洁性而广泛应用于联邦学习, 但它未能解决数据偏移问题, 即数据量较小的设备对全局模型的贡献较小, 从而影响模型的泛化性能, 此外, FedAvg 可能会使跨客户端的数据分布发生分歧。作者在[4]中提出了 FedProx, 它增加了局部训练问题的近似项, 以抵消数据分布造成的可能的分歧。此外, 这种方法通过允许在每个选定的客户端上进行不同数量的工作来适应系统差异性。尽管如此, FedProx 的性能对于近似系数的选择是敏感的, 其调整取决于系统规模大小的先前知识[4]。SCAFFOLD [5]引入了客户端和服务端控制变量, 它们有效地作为本地和全局信息的跟踪变量。结果显示, 通过利用跨客户端数据的相似性, 它的性能优于 FedAvg/Prox, 但是由于额外的控制变量, 它每轮的通信成本增加了一倍。另一个工作分支源于原始 - 对偶优化方法, 如 FedHybrid [6]和 FedPD [7]方法, 这些方法在分布式优化中获得了很大的成功。FedHybrid 是一个同步方法, 要求所有客户端在每轮中进行更新。该方法通过将客户端分为执行梯度更新和执行牛顿更新的两组来应对系统差异性。然而, 这种方法并未适用于联邦深度学习, 因为牛顿法不适合用于大型神经网络, 并且其收敛性建立在强凸性假设上。FedPD 在客户端采用基于梯度下降的本地训练, 并进行可变量的工作。类似于本文提出的方法, FedPD 使用双变量来捕捉局部模型和全局模型之间的差异。然而, FedPD 采用了全客户端参与的策略, 即所有客户端在每轮中更新其本地模型和双变量, 并且以固定概率与服务器通信来更新全局模型。

与之相比, 本文提出的 FDLADMM 算法基于原始 - 对偶优化算法, 能够有效适应系统规模, 同时优化通信效率和模型训练速度。在 FDLADMM 中, 服务器每轮仅与部分客户端通信, 客户端只在被选中时执行本地训练, 此外, FDLADMM 使用双变量来指导模型进行本地训练, 这有效地解释了在没有调优的情况下自动适应数据分布, 这解决了在 FL 设置中遇到的一个关键挑战, 即避免客户端漂移, 以防止模型过度拟合特定选定客户的数据。本文的方法与随机 ADMM [8]紧密相关, 在 FL 环境中相当于随机激活客户端与服务器, 但是它考虑了在不同硬件条件和数据量下, 局部问题可能不会精确解决的现实情况, 允

许在客户端进行可变数量的训练。还有一个相关研究领域是异步 ADMM [9] [10], 旨在解决前述的掉队问题, 但是它所依赖的如有界延迟等假设, 在 FL 环境中可能不实际。

3. 算法

$$\underset{\theta \in \mathbb{R}^D}{\text{minimize}} = \sum_{i=1}^m f_i(\theta) \quad (1)$$

FL 的目标是将其构建为一个损失最小化问题: 最小化 $\sum_{i=1}^m f_i(\theta)$, 其中 m 是客户端的数量, $f_i(\theta)$ 表示加权的本地训练损失。常见的权重选择包括 $\alpha_i = n_i/N$, 其中 $N = \sum_{i=1}^m n_i$ (即根据它们数据量比例给客户端赋予权重), 或者 $\alpha_i = 1$ (即所有客户端平等权重, 这有助于避免对拥有更多数据的客户端过度拟合, 同时这也是本文实验中采用的选择)。FL 的目标是针对分布在不同客户端的数据, 对模型参数 θ 解决公式(1)所示的问题。与数据中心分布式机器学习相比, 通信是 FL 应用的瓶颈。联邦系统中大量的客户端数量使得大规模消息传输的同步通信变得不可行, 这是由于带宽限制和延迟节点的存在。此外, 大数据量的参与以及原始数据中个人信息的可能泄漏, 阻止了服务器收集本地数据进行集中处理。因此, 本文提出的 FDLADMM 算法, 在优化通信效率和训练速度的前提下, 有效的保护了数据隐私。

3.1. FDLADMM 算法

如公式(2)所示, 其所描述的最优化问题在数学上与公式(1)是等效的, 因为它们的最优解是相同的。这种等效性使得公式(2)更适合在 FL 环境中进行解释和应用。

$$\underset{w, \theta \in \mathbb{R}^D}{\text{minimize}} = \sum_{i=1}^m f_i(\theta_i), \text{ s.t } w_i = \theta, \forall i \in [m] \quad (2)$$

在公式(2)中, θ_i 可以被解释为客户端 i 持有的本地模型, 这代表了每个参与方(如移动设备、智能体、个人用户等)在本地训练出的模型参数。而 θ 则可以被解释为服务器持有的全局模型, 这是各个参与方本地模型的整合和更新结果。在 t 轮的初期, 系统将选取一组客户端(标记为 S^t), 并且这些客户端将会加载来自服务器的模型 θ^t 。需要指出的是, FDLADMM 的一个积极因素在于其客户端的选取机制: 该机制可以是基于设备运行状况(如能量和带宽)的动态规则, 其必须确保所有客户端都有一定的概率参与进来。然后进行本地训练, 这一过程涉及优化函数 $L = (w_i, v_i^t, \theta^t)$, 以求达到最小化, 其中 v_i^t 定义为每个客户端 i 持有的本地对偶变量, 且 v_i^t 属于 $\mathbb{R}^d$, w_i 定义为最小化 $L = (w_i, v_i^t, \theta^t)$ 的解, $\rho > 0$ 是二次项的系数。

$$L_i(w_i, v_i^t, \theta^t) = f_i(w_i) + (v_i^t)^T (w_i - \theta^t) + \rho/2 \|w_i - \theta^t\|^2 \quad (3)$$

对偶变量和二次项的结合有助于在更新模型参数(使用本地数据)和与服务器模型保持一致之间取得平衡(服务器模型用于整合所有参与者的信息)。本文注意到, FedProx 以 $v_i^t = 0$ 的方式类似地解决了(3)式, 虽然这在一定程度上为客户端偏移提供了保障。然而, FedProx 的竞争性能依赖于对 ρ 的精确调整[4] [5]。对偶变量的添加使 FDLADMM 能够自动适应数据分布和系统规模, 并显著减轻了超参数调整问题。除了本地模型更新外, 还会更新每个选定客户端的对偶变量, 本文将原始变量和对偶变量结合起来, 形成所谓的增强模型 $w_i + \rho^{-1}v_i^t$, 并使用 $\Delta_i^t \in \mathbb{R}^d$ 来表示客户端 i 向服务器发送的更新信息, 即连续增强模型之间的差异:

$$\Delta_i^t = (w_i^{t+1} + \rho^{-1}v_i^{t+1}) - (w_i^t + \rho^{-1}v_i^t) \quad (4)$$

服务器在从选定的客户端收集更新消息后更新全局模型:

$$\theta^{t+1} = \theta^t + |S^t|^{-1} \eta \sum_{i \in S^t} \Delta_i^t \quad (5)$$

其中 $\eta > 0$ 是服务器聚合步长。不同的 η 值适用于系统规模和统计变化各异的场景。本文通过经验观察到, 将 η 设置为 1 可以实现快速训练速度, 而设置 η 为 $|S^t|/m$ 有助于在检测到显著的数据差异性时消除振荡行

为。与 FedAvg/Prox 不同, 后者仅使用客户端 $i \in S'$ 的当前模型信息来更新全局模型, FDLADMM 有效地融入了历史信息, 这是通过公式(5)中的更新规则实现的。

3.2. 算法总结

本文阐述了 FDLADMM 如何对 FedAvg 和 FedProx 进行了扩展。如公式(3)所定义, 当设置对偶变量 $v_i = 0$ 时, 本文就回到了 FedProx 的局部训练问题。如果进一步将二次项系数 ρ 设置为 0, 那么就转变为了 FedAvg 的局部训练问题。这两个参数的引入主要是为了应对联邦学习中的一个基本挑战, 即如何平衡广泛的本地训练(以加快整个 FL 过程的收敛速度, 从而减少通信轮次总数), 以及防止客户端模型过度拟合其本地数据。因此本文认为 FDLADMM 是一种有效的策略, 能够利用本地设备的能力来减少通信成本。

4. 实验评估

4.1. 数据集介绍

本文分别在两种图像分类的数据集上进行实验, 数据集的具体介绍如下。

4.1.1. Fashion MNIST 数据集

Fashion-MNIST 是一个图像分类数据集, 用于机器学习算法的基准测试。Fashion-MNIST (FMNIST) 由 Zalando Research 提供, 包含 70,000 个服饰图像, 分为 10 个类别, 其中训练集包含 60,000 张图像, 测试集包含 10,000 张图像, 数据集的图像大小为 28×28 像素[11]。

4.1.2. CIFAR-10 数据集

CIFAR-10 是一个包含日常物品的彩色图像数据集, 由 Hinton 的学生 Alex Krizhevsky 和 Ilya Sutskever 整理, 旨在识别广泛的物体类别[12]。该数据集涵盖了十个类别, 包括飞机、汽车、鸟类、猫、鹿、狗、蛙类、马、船和卡车, 每个类别包含 6000 张 32×32 像素的 RGB 彩色图像。整个数据集由 50,000 张训练图像和 10,000 张测试图像构成。

4.2. 实验环境参数

本文实验具体的硬件配置如表 1 所似乎。在软件方面, 实验采用了 Python 编程语言, 版本为 3.8.0, 确保代码的兼容性。

Table 1. Experimental setup
表 1. 实验设置

参数	参数取值
操作系统	Ubuntu 20.04
开发语言	Python 3.8.0
CUDA	11.3
CPU	Xeon(R) Platinum 8358P
GPU	RTX 3090(24G)

4.3. 模型参数

本文使用两种流行的具有两个卷积层的卷积神经网络(Convolutional Neural Network, CNN)模型[3][13], 它们都具有一个卷积模块(两个 5×5 卷积层, 每个后跟 2×2 最大池层)和一个全连接模块。两个

模型的输入是一个分别为 784 和 3072 维的平面图像, 而两个模型的输出都是一个从 0 到 9 的类标签。
表 2 总结实验中使用的数据集、模型大小和目标精度。并根据表 2 详细介绍的模型参数, 将本文提出的方法与 FedAvg、FedProx 和 SCAFFOLD 进行了比较。

Table 2. Experimental setup
表 2. 实验设置

模型	参数数量	数据集	目标准确率
CNN 1	1,663,370	FMNIST	80%
CNN 2	1,105,098	CIFAR-10	45%

4.4. 结果分析

首先, 本文使用了 FMNIST 和 CIFAR-10 数据集, 通过设置相同的客户端数量, 比较了相同的通信次数时模型的测试精度, 测试精度的具体计算如式(6)所示。

$$ACC = \frac{TP}{TP + FN} \quad (6)$$

其中, TP 表示将正样本预测为正样本总数, FN 表示将正样本预测为负样本总数。实验结果如图 2 所示, 可以注意到, 在所有测试的情况下, FDLADMM 始终优于所有基线方法。其次, 本文扩大系统规模, 在 FMNIST 数据集上, 本文分别比较了 500 个客户端和 1000 个客户端下 FDLADMM 的性能。本文首先为每种算法在 200 个客户端设置下调整了超参数以获得最佳性能, 然后保持超参数不变的情况下扩大系统规模, 结果如图 3 所示。需要注意的是, 增加客户端会引入额外的双变量到 FDLADMM 中, 相同数量的数据会在更多的指导下进行处理, 这有助于在较大规模上使 FDLADMM 相对于基线方法的改进更大。因此, FDLADMM 被证明更快地实现了对本地信息的有效整合。本文得出结论, FDLADMM 具有强大的可扩展性, 增加规模时的性能提升归因于与服务器的信息交换数量增加, 以及额外的双变量提供的指导来优化学习过程。

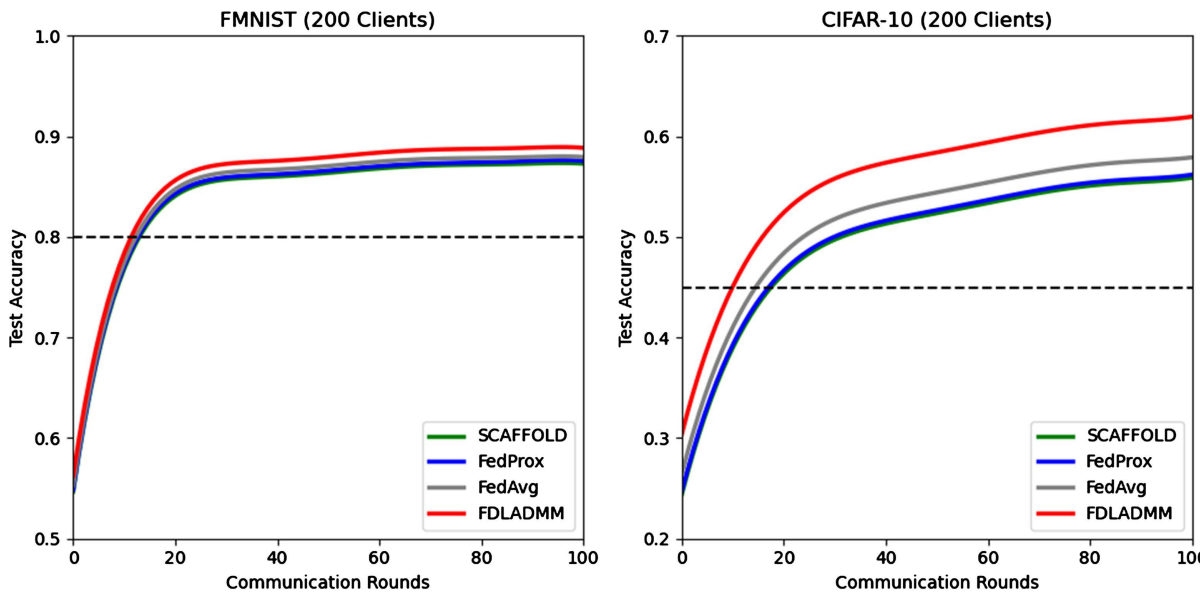


Figure 2. Impact of different methods on system performance
图 2. 不同方法对系统性能的影响

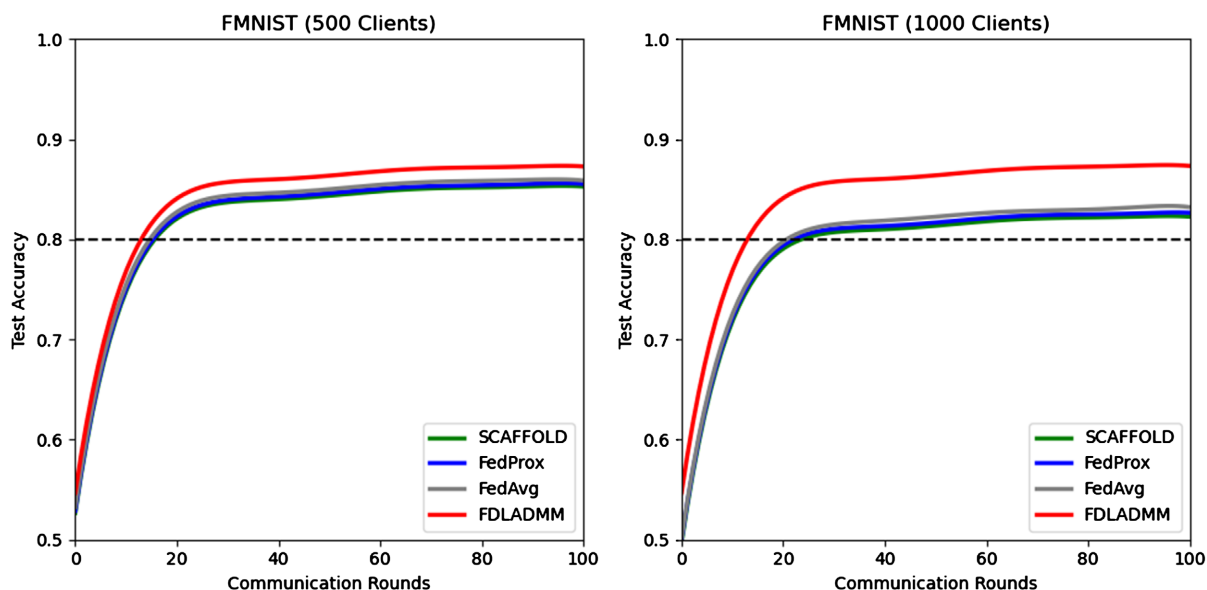


Figure 3. Impact of different system size on system performance

图 3. 不同系统规模对系统性能的影响

5. 总结

本文深入探讨了面向复杂任务协作的分布式学习技术,并特别聚焦于联邦学习框架在此背景下的应用和挑战。由于设备之间的通信带宽和计算资源的限制、以及系统规模差异性等难题,本文提出了 FDLADMM 算法,它能够有效适应系统规模,同时优化通信效率和模型训练速度。这一新颖的联邦学习框架,通过允许设备之间在不共享原始数据的情况下共同训练模型,显著减少了通信需求,并且随着系统规模增大,无需进行超参数调整即可有效适应。这一创新的方法为应对复杂任务协作中的挑战提供了一种可行且高效的解决方案,并有望在未来的研究和实践中得到广泛应用。

基金项目

国家自然科学基金资助项目(61602305, 61802257);上海市自然科学基金资助项目(18ZR1426000, 19ZR1477600)。

参考文献

- [1] Bottou, L. (2010) Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT 2010*, Physica-Verlag HD, 177-186. https://doi.org/10.1007/978-3-7908-2604-3_16
- [2] Dean, J., Corrado, G., Monga, R., et al. (2012) Large Scale Distributed Deep Networks. *Advances in Neural Information Processing Systems*, 25. <https://www.researchgate.net/publication/266225209>
- [3] McMahan, B., Moore, E., Ramage, D., Hampson, S. and Arcas, B.A. (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. *Artificial Intelligence and Statistics, PMLR*, 1273-1282. <https://doi.org/10.48550/arXiv.1602.05629>
- [4] Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A. and Smith, V. (2020) Federated Optimization in Heterogeneous Networks. *Machine Learning and Systems*, 429-450. arXiv: 1812.06127.
- [5] Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S. and Suresh, A.T. (2020) SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. *International Conference on Machine Learning, PMLR*, 5132-5143. <https://doi.org/10.48550/arXiv.1910.06378>
- [6] Niu, X. and Wei, E. (2021) FedHybrid: A Hybrid Primal-Dual Algorithm Framework for Federated Optimization. arXiv: 2106.01279.

- [7] Zhang, X., Hong, M., Dhople, S., Yin, W. and Liu, Y. (2021) FedPD: A Federated Learning Framework with Adaptivity to Non-IID Data. *IEEE Transactions on Signal Processing*, **69**, 6055-6070.
<https://doi.org/10.1109/tsp.2021.3115952>
- [8] Zhang, R. and Kwok, J. (2014) Asynchronous Distributed ADMM for CONSENSUS Optimization. ICML, 1701-1709.
- [9] Chang, T., Hong, M., Liao, W. and Wang, X. (2016) Asynchronous Distributed ADMM for Large-Scale Optimization—Part I: Algorithm and Convergence Analysis. *IEEE Transactions on Signal Processing*, **64**, 3118-3130.
<https://doi.org/10.1109/tsp.2016.2537271>
- [10] Zheng, Y., Song, Y., Hill, D.J. and Zhang, Y. (2018) Multiagent System Based Microgrid Energy Management via Asynchronous Consensus ADMM. *IEEE Transactions on Energy Conversion*, **33**, 886-888.
<https://doi.org/10.1109/tec.2018.2799482>
- [11] Xiao, H., Rasul, K. and Vollgraf, R. (2017) Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv: 1708.07747.
- [12] Krizhevsky, A. (2009) Learning Multiple Layers of Features from Tiny Images. University of Toronto.
<https://learning2hash.github.io/publications/cifar2009learning/>
- [13] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. <https://doi.org/10.1109/5.726791>