中国科学: 数学 2026 年 第 56 卷 第 ? 期: 1 ∼ 28
SCIENTIA SINICA Mathematica
综 述

《中国科学》杂志社
SCIENCE CHINA PRESS
CrossMark
click for updates

分布式统计学习的若干进展

刘卫东, 毛晓军*, 余政嵘

上海交通大学数学科学学院, 上海 200240
E-mail: weidongl@sjtu.edu.cn, maoxj@sjtu.edu.cn, yuzhengrong@sjtu.edu.cn

收稿日期: 2025-11-19; 接受日期: 2026-03-21; 网络出版日期: 2026-05-12; * 通信作者
国家自然科学基金 (批准号: 72495121, 12422111 和 12371273) 资助项目

摘要 随着海量数据的产生, 传统统计学习方法在处理大规模数据时面临挑战, 推动了分布式学习成为重要的研究方向. 然而, 随着技术的进步, 分布式学习对可信保障的需求日益增加, 尤其是在稳健性、隐私保护和公平性方面. 同时, 分布式学习的应用场景也在不断扩展, 涵盖了半监督学习和在线学习等新兴领域. 本文综述分布式学习的核心框架和优化算法, 并重点探讨其在不同可信需求和应用场景中的最新研究进展.

关键词 分布式算法 稳健性 隐私保护 公平性 半监督学习 在线学习

MSC (2020) 主题分类 68W15, 68T05, 62F35, 68P27, 68W27

1 引言

随着大数据时代的到来, 传统的统计学习方法在处理海量数据时面临显著挑战. 根据国际数据公司 (International Data Corporation) 的报告, 2011 年全球创建和复制的数据总量已达到 1.8 泽字节 (约为 10^{21} 字节), 并在五年内增长了近九倍 (参见文献 [46]). 在传统的集中式统计学习方法中, 所有数据通常存储并计算于单一节点. 然而, 在大数据背景下, 单一计算节点的处理能力已无法满足日益增长的数据需求. 首先, 海量原始数据传输至计算节点可能引发网络拥堵, 并导致较高的通信延迟. 例如, 在医疗领域, 如果需要将全国范围内各地医院的电子病历数据传输至单一计算节点进行处理, 庞大的数据量和有限的网络带宽可能导致传输时间过长, 甚至造成网络阻塞. 其次, 集中式训练将所有计算集中于单一节点, 这会使得该节点承受巨大的计算负担, 从而产生显著的计算延迟. 以金融行业为例, 若需要在单一节点上处理大量的交易数据以训练风险评估模型, 计算资源可能无法满足实时处理需求, 从而导致模型训练时间过长, 进而影响业务响应速度. 最后, 训练数据常包含敏感的个人信息, 出于隐私保护的考虑, 用户往往不愿意将原始数据共享给他人. 例如, 患者的医疗记录包含个人健康信息, 患者可能因担心隐私泄露而不愿将数据提供给外部机构进行训练.

英文引用格式: Liu W D, Mao X J, Yu Z R. Recent advances in distributed statistical learning (in Chinese). Sci Sin Math, 2026, 56: 1–28, doi: 10.1360/SSM-2025-0292

© 2026《中国科学》杂志社

为应对上述挑战, 分布式学习作为一种有效的替代方法应运而生. 在这一方法中, 数据拥有者通过分布式协作训练机器学习模型, 而无需将原始数据上传至中心节点进行集中处理. 在分布式学习中, 节点利用其本地计算和通信资源进行本地训练, 并与中心节点或其他节点交换信息. 这一框架不仅有效缓解了集中式学习中的计算和通信负担, 还在保护数据隐私方面具有显著优势. 鉴于其巨大的潜力, 分布式学习在近年来得到了广泛研究.

从统计学角度看, 分布式学习主要涉及三个核心问题. 第一, 在通信受限和数据分散存储的条件下, 如何构造既具有计算可扩展性又保持较高统计效率的分布式估计方法, 并进一步在分布式环境下对未知参数进行不确定性量化, 例如, 构造置信区间或进行假设检验; 第二, 在稳健性、隐私保护和公平性等实际约束下, 如何兼顾统计精度、计算效率与系统可靠性; 第三, 在半监督学习、在线学习和非参数统计学习等不同任务与应用场景下, 如何将分布式统计学习方法推广到更丰富的任务设定中. 围绕上述问题, 近年来统计学与机器学习领域提出了大量分布式学习方法, 并逐步形成了若干具有代表性的研究方向. 本文主要沿着这三个方面展开综述.

从节点连接方式的角度来看, 分布式架构主要分为中心化架构和去中心化架构, 如图 1 所示. 这两种架构的主要区别在于节点间信息交互的方式和应用场景. 在中心化架构中, 所有节点都与一个中心节点连接, 数据和计算任务的协调与汇总通常由该中心节点完成. 中心化架构的优势在于易于管理和控制, 且数据的集中处理有助于提高计算效率, 但也容易导致中心节点负担过重. 与此不同, 在去中心化架构中, 各节点之间没有明确的中心节点, 数据和计算任务在多个节点间分布与协作. 每个节点独立执行本地计算, 并通过局部的信息交换进行协作. 这种架构特别适合无法集中处理的场景, 如边缘计算和物联网. 去中心化架构具有更好的扩展性和灵活性, 在不需要中心节点协调的情形下使得所有节点间实现高效的协作.

如前所述, 分布式学习是由多个节点组成的开放系统. 因此相比单机模型, 分布式学习中的节点更易受到各种破坏性事件的影响. 因此, 确保分布式学习的可信性 (trustworthy) 至关重要. Wei 和 Liu [132] 指出, 稳健性、隐私保护性和公平性是提升分布式学习可信性的三大关键因素, 共同构成了可信分布式学习的核心框架. (1) 稳健性: 稳健性指系统在学习或推理过程中处理异常情形的能力. 具体而言, 当系统遇到异常信息、数据异常值或重尾噪声等情形时, 能够持续正常运行并作出合理的推理与决策. (2) 隐私保护性: 隐私保护性指模型在训练和推理过程中保障数据隐私不被泄露. 例如, 在分

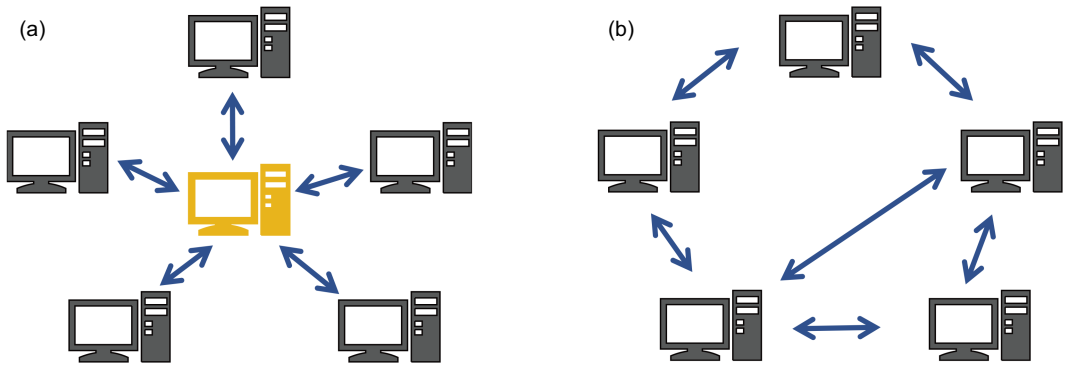


图 1 (网络版彩图) 中心化与去中心化架构对比. 在中心化架构中, 所有节点依赖于一个中心节点进行参数聚合与更新; 而在去中心化架构中, 各节点之间通过点对点的方式直接通信与协作, 从而避免单点故障并提升系统稳健性. (a) 中心化架构; (b) 去中心化架构

布式学习中, 无法通过梯度信息推断出节点的原始数据. (3) 公平性: 公平性指模型能够有效处理数据偏差和算法偏差, 确保对个人或群体的公正性. 偏见指模型在训练过程中对特定群体或个体的偏好或不公正倾向. 这三个方面相互交织, 共同保障了分布式学习的可信性.

除了上述核心问题之外, 近年来, 许多研究还关注分布式统计学习在不同场景下的应用与拓展, 如半监督学习、在线学习和非参数统计学习等, 从而进一步丰富了分布式学习的研究内容. 其中, 半监督学习面向标签稀缺或数据不完全的情形, 结合少量标注数据与大量未标注数据进行训练; 在线学习强调模型在动态环境中的持续更新能力; 非参数统计学习在不预先指定有限维参数模型的条件下, 提供了更灵活的建模方式.

本文对现有的分布式统计学习研究进行了全面综述, 重点讨论了该领域的核心框架和优化算法. 本文余下内容的结构安排如下. 第 2 节主要介绍分布式统计学习的两种基本框架, 综述这两种框架下的常用优化算法、统计性质和分布式统计推断问题. 第 3 节重点讨论可信分布式统计学习, 包含稳健性 (包括 Byzantine 稳健性和重尾噪声与离群值稳健性)、隐私保护性和公平性这三个重要组成部分. 第 4 节探讨分布式统计学习在其他场景中的应用, 重点包括半监督学习、在线学习和非参数统计学习. 第 5 节展望未来研究的方向.

为便于后文统一表述, 本文先对若干基本概念作简要说明, 并在表 1 中约定常用符号.

可信分布式统计学习 可信分布式统计学习是指, 在分布式学习过程中, 同时兼顾模型性能与系统可靠性的一类学习框架. 在本文中, “可信” 主要包含三个核心维度: 稳健性、隐私保护性和公平性. 其中, 稳健性强调模型在异常值、重尾噪声或恶意攻击下的稳定性, 隐私保护性关注训练和通信过程中敏感信息不被泄露, 公平性则强调模型对不同群体或个体的预测与决策应尽可能公正.

去中心化 M- 估计 M- 估计是统计学中通过最小化某个经验损失函数或求解相应估计方程来构造参数估计量的一类广义估计方法. 去中心化 M- 估计是指, 在无中心节点的网络中, 各节点仅依赖局部数据和邻居通信, 协同求解一个全局 M- 估计问题.

统计效率 统计效率用于衡量一个估计方法利用样本信息的充分程度. 一般而言, 若两个估计量在相同样本规模下都用于估计同一目标参数, 则误差更小、方差更低或极限分布更集中的方法通常被认为具有更高的统计效率. 在本文中, 统计效率既可以从有限样本角度理解为估计误差的大小, 也可以从渐近角度理解为渐近方差的大小.

表 1 本文常用符号说明

符号	含义
m	分布式系统中的节点总数
N	全部节点上的总样本量
n	每个节点上的样本量
p	待估参数的维数
$\boldsymbol{\theta}$	待估参数
$\boldsymbol{\theta}_i$	第 i 个节点上的局部参数
$\boldsymbol{\theta}^*$	真实参数
$\widehat{\boldsymbol{\theta}}$	全局最优估计量, $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$
$\widehat{\boldsymbol{\theta}}_i^{(t)}$	第 i 个节点上在 t 时刻的局部估计量
$f_i(\boldsymbol{\theta})$	第 i 个节点上的局部损失函数
$f(\boldsymbol{\theta})$	全局损失函数, 通常为 $f(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^m f_i(\boldsymbol{\theta})$
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	去中心化通信网络对应的图结构, 其中 $\mathcal{V}$ 为通信图中的节点集合, $\mathcal{E}$ 为通信图中的边集合

最优统计速率 最优统计速率是指, 在给定模型、参数空间和样本规模下, 估计误差随样本量增长所能达到的最好收敛阶. 它反映了一个统计问题本身的难度, 也为比较不同分布式方法的优劣提供了基准. 例如, 在许多低维参数估计问题中, 最优误差阶通常为 $\mathcal{O}_{\mathbb{P}}(N^{-1/2})$; 在高维稀疏问题中, 则常见为 $\mathcal{O}_{\mathbb{P}}(\sqrt{s\log p/N})$. 当一个分布式方法达到与集中式方法相同的误差阶时, 通常称其达到了最优统计速率, 或至少达到了与单机方法同阶的统计精度.

通信复杂度 通信复杂度是衡量分布式算法在节点之间或节点与中心节点之间交换信息所需代价的重要指标. 它可以从不同角度刻画, 如通信轮次、每轮传输向量的维数和总传输比特数等. 在分布式统计学习中, 通信复杂度往往直接决定算法的可扩展性和实际部署成本. 尤其当参数维数较高, 网络规模较大或通信资源有限时, 如何降低通信复杂度就成为算法设计的核心问题之一.

通信复杂度 - 统计精度权衡 通信复杂度 - 统计精度权衡是分布式统计学习中的一个核心问题, 指的是算法在减少通信代价的同时, 往往可能牺牲一定的统计精度; 反之, 若希望逼近集中式学习的最优误差或最优渐近方差, 通常又需要付出更多通信开销. 换言之, 通信越少, 估计可能越粗糙; 通信越充分, 估计通常越精确, 但系统代价也越高.

2 分布式统计学习的两种框架

分布式统计学习的核心问题在于, 在通信受限和数据分散存储的条件下, 如何设计既具有计算效率又具有良好统计性质的学习方法. 根据系统结构的不同, 现有方法通常可分为中心化框架和去中心化框架两类.

为更直观地说明中心化与去中心化框架的差异及其适用条件, 表 2 从多个维度对两类框架进行了归纳比较. 总体而言, 中心化框架更适用于网络结构稳定、中心节点可靠且任务目标强调统一聚合的场景; 而去中心化框架则更适用于无固定中心节点、节点地位对等、对系统稳健性和可扩展性要求较高的场景.

2.1 中心化分布式统计学习

在中心化分布式统计学习框架中, 所有数据处理通过中心节点协调和控制. 该系统由多个分布在

表 2 中心化与去中心化分布式统计学习框架的比较

主要特征	中心化框架	去中心化框架
网络结构	存在中心节点, 负责信息聚合, 协调与广播	无中心节点, 各节点通过邻域连接协同完成计算与更新
通信模式	节点与中心节点双向通信	节点之间点对点通信
通信瓶颈	中心节点容易成为通信瓶颈, 尤其在大规模网络中更为明显	无单点通信瓶颈, 但通常需要更多轮邻域通信以达成共识
单点故障风险	存在单点故障风险, 中心节点失效可能影响整体运行	无中心节点, 对单点故障通常更不敏感
统计效率	更容易逼近集中式学习的统计效率, 理论分析较成熟	统计效率受网络拓扑、权重矩阵和步长等影响
隐私与安全风险	中心节点可能成为隐私泄露或攻击的重点目标	避免中心汇聚带来的风险, 但仍需应对通信中的安全与隐私挑战
典型部署环境	具备中心协调节点的环境, 如云平台、数据中心和多服务器集群等	无固定中心节点的环境, 如传感器网络、边缘计算和无人机集群等

不同位置的本地节点组成, 这些节点通过网络连接, 将本地数据或计算结果传输至中心节点进行处理. 具体而言, 各节点负责处理本地数据并计算初步统计量或训练模型, 然后将结果上传给中心节点, 中心节点基于收集到的信息进行全局优化. 这一框架设计相对简单, 因为所有节点依赖于中心节点进行信息传递和聚合, 易于实现且具有良好的组织结构. 中心节点能够有效地控制全局一致性, 确保各节点计算结果的有效整合. 因此, 这一框架中的核心问题在于, 在通信受限条件下, 分布式估计能否达到与集中式方法相当的统计效率, 并进一步保留集中式方法中的统计推断能力, 以及这种统计效率对节点数量和通信次数的依赖关系如何刻画.

在中心化框架中, 一般考虑有 m 个节点, 每个节点与中心节点相连, 组成中心化网络, 每个节点都能与中心节点进行信息传输. 设共有 N 个样本, 这些样本平均分配在每个节点中, 每个节点拥有 $n = N/m$ 个局部样本. 在这样的场景下, 通常需要求解以下优化问题:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} f(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\boldsymbol{\theta}), \tag{2.1}$$

其中 $f_i(\boldsymbol{\theta})$ 为节点 i 上的损失, $f(\boldsymbol{\theta})$ 为全局损失, $\boldsymbol{\theta}$ 为模型参数. 为了解决问题 (2.1), 已有多种方法被提出. 根据通信次数的不同, 这些方法可分为一次方法 [81,102,162] 和迭代方法 [40,56,60,105,125]. 接下来, 分别介绍这两类方法.

在中心化框架中, 最先被研究的问题是, 仅通过一次通信对各节点的局部估计进行聚合, 是否仍能获得与全样本估计相当的统计精度. 针对这一问题, 一次方法的基本思想是 “分而治之” (divide and conquer, DC). 具体而言, 节点 i 使用本地数据计算参数的估计值 $\widehat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}} f_i(\boldsymbol{\theta})$, 然后将该估计值传输至中心节点, 中心节点通过计算平均值 $\overline{\boldsymbol{\theta}} = \sum_{i=1}^m \widehat{\boldsymbol{\theta}}_i/m$ 得到最终的全局估计. 一次方法的优点在于较高的通信效率, 因为节点只需要传输一次局部估计至中心节点, 便可得到全局估计. 围绕平均估计量 $\overline{\boldsymbol{\theta}}$ 能否保持全样本估计的统计效率, 已有大量研究对其理论性质进行了分析. 例如, Zhang 等 [162] 证明了, 在某些假设条件下, 估计量的误差满足以下关系:

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2] \leqslant \frac{C_1}{N} + \frac{C_2}{n^2} + \mathcal{O}\left(\frac{1}{Nn} + \frac{1}{n^3}\right), \tag{2.2}$$

其中 C_1 和 C_2 为常数. 如果局部样本量 n 足够大, 以满足 $n^{-2} \leqslant N^{-1}$, 则 (2.2) 中的误差阶简化为 $\mathcal{O}(N^{-1})$, 此时与全局估计量的误差阶相同. 这表明, 为了达到与集中式模型相同的估计效果, 局部样本量不能过小. 此外, Rosenblatt 和 Nadler [102] 进一步研究了该理论, 证明了估计量与全局估计量在一阶误差上等价, 但二阶误差项不可忽视, 尤其在非线性模型下.

上述结果表明, 一次方法的主要统计瓶颈在于高阶偏差项随局部样本量减小而变得不可忽视. 此外, 当模型是非线性时, 平均聚合的效果会进一步变差. 从统计推断角度看, 仅基于一次平均构造置信区间或进行假设检验也更具挑战 (参见文献 [60]). 这说明对于分布式系统构建允许合理迭代次数的算法十分重要. 为了进一步缓解一次方法中的高阶偏差问题, Huang 和 Huo [56] 提出了均值估计量的一次微调方法, 首先得到均值估计量 $\widehat{\boldsymbol{\theta}}^{(0)} = \overline{\boldsymbol{\theta}}$, 并将其广播至所有节点. 然后, 在所有节点上计算局部梯度 $\nabla f_i(\widehat{\boldsymbol{\theta}}^{(0)})$ 和局部 Hessian 矩阵 $\nabla^2 f_i(\widehat{\boldsymbol{\theta}}^{(0)})$, 并将计算结果传输至中心节点, 得到全局梯度 $\nabla f(\widehat{\boldsymbol{\theta}}^{(0)}) = \sum_{i=1}^m \nabla f_i(\widehat{\boldsymbol{\theta}}^{(0)})/m$ 和 Hessian 矩阵

$$\nabla^2 f(\widehat{\boldsymbol{\theta}}^{(0)}) = \sum_{i=1}^{m} \nabla^2 f_i(\widehat{\boldsymbol{\theta}}^{(0)})/m.$$

因此, 一次微调的估计量定义为

$$\widehat{\boldsymbol{\theta}}^{(1)} = \widehat{\boldsymbol{\theta}}^{(0)} - \nabla^2 f(\widehat{\boldsymbol{\theta}}^{(0)})^{-1}\nabla f(\widehat{\boldsymbol{\theta}}^{(0)}). \tag{2.3}$$

相比于一次估计量, 该估计量的通信开销为 $\mathcal{O}(mp^2)$, 但其统计性质更优. Huang 和 Huo [56] 证明了

$$\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^{(1)}-\boldsymbol{\theta}^*\|_2^2]\leqslant\frac{C_1}{N}+\mathcal{O}\left(\frac{1}{n^4}+\frac{1}{N^2}\right),$$

其中 C_1 为常数, 并给出了其渐近正态性. 这表明, 与 (2.2) 相比, 该方法的高阶误差项进一步减小. 因而, 与简单平均方法相比, 该方法在更弱的局部样本量条件下即可达到与集中式方法同阶的误差, 从而放宽了对节点数量的限制. 上述方法的自然扩展是多次逼近的估计量. 令 $\widehat{\boldsymbol{\theta}}^{(t)}$ 为第 t 步的估计量, 可以将迭代 (2.3) 拓展至多次迭代. 然而, 每次迭代都需要计算和传输大量 Hessian 矩阵, 导致通信代价较高. 为此, Shamir 等 [105] 从优化实现角度提出了一种近似 Newton 法, 能够在不直接传递完整 Hessian 矩阵的情形下进行 Newton 型迭代. 类似地, Jordan 等 [60] 提出了 CSL (communication-efficient surrogate likelihood) 方法, 利用中心节点的 Hessian 矩阵 $\nabla^2 f_{\text{center}}(\widehat{\boldsymbol{\theta}}^{(t)})$ 替代 $\nabla^2 f(\widehat{\boldsymbol{\theta}}^{(t)})$, 由此迭代为

$$\widehat{\boldsymbol{\theta}}^{(t+1)}=\widehat{\boldsymbol{\theta}}^{(t)}-\nabla^2 f_{\text{center}}(\widehat{\boldsymbol{\theta}}^{(t)})^{-1}\nabla f(\widehat{\boldsymbol{\theta}}^{(t)}),$$

从而避免传递 Hessian 矩阵, 减少通信代价. 在一定假设下, Jordan 等 [60] 证明了

$$\|\widehat{\boldsymbol{\theta}}^{(t+1)}-\widehat{\boldsymbol{\theta}}\|_2\leqslant\frac{C_1}{\sqrt{n}}\|\widehat{\boldsymbol{\theta}}^{(t)}-\widehat{\boldsymbol{\theta}}\|_2,\quad\forall t\geqslant 0,$$

其中 $\widehat{\boldsymbol{\theta}}=\arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$, C_1 为常数. 因此, 经过 $\lceil\log m/\log n\rceil$ 轮次的迭代后, 估计量即可逼近 $\widehat{\boldsymbol{\theta}}$, 达到与集中式方法相当的统计精度.

上述内容主要讨论了中心化分布式方法在点估计层面如何通过有限通信逼近 $\widehat{\boldsymbol{\theta}}$. 然而, 仅有点估计误差的控制还不够, 更关键的问题在于分布式估计量能否进一步继承集中式方法的渐近分布与推断能力. 因此, 近年来相关研究开始关注中心化分布式框架下的统计推断问题. 在现有方法中, 最常见的一类思路是, 基于分布式估计量的渐近正态性开展 Wald 型推断, 即先建立估计量的极限正态分布, 再通过估计其渐近方差来构造置信区间或进行假设检验. 例如, Battey 等 [7] 在分治框架下系统研究了分布式统计推断问题, 在统一的似然框架下提出了基于子样本统计量聚合的估计与检验方法, 并分析了节点数量增长对统计效率的影响, 说明在适当条件下分布式方法可以保持与集中式方法相同的推断效率. 当然一些工作也从不同角度建立了分布式估计量的渐近正态性, 从而为统计推断提供理论依据, 如加权分治法 [47]、一次微调方法 [56] 和 CSL 方法 [33, 60] 等. 除 Wald 型推断外, 另一类重要思路是, 基于似然比或经验似然统计量开展检验. 这类方法的优势在于, 可以直接构造具有标准极限分布的检验统计量, 从而在一定程度上减轻对渐近方差估计的依赖, 因而尤其适用于分布式环境. 例如, Zhou 等 [166] 提出了 DEL (distributed empirical likelihood) 方法, 用于多个不平衡数据源的分布式数据整合分析, 并证明了相应的经验似然比统计量在大样本下仍满足 Wilks 定理. 此外, 还有一些针对其他统计问题的分布式推断研究, 如针对对称统计量的分布式推断 [21] 以及主成分分析的分布式推断方法 [41] 等.

总体而言, 中心化分布式统计学习的研究不仅关注分布式点估计能否达到与集中式方法相同的误差阶, 也关注分布式算法输出能否继承集中式估计量的渐近分布与推断能力. 一次方法主要解决了低通信约束下点估计的可行性问题, 但其高阶偏差往往不可忽视; 允许适当迭代的修正方法则在减小偏差的同时, 为建立渐近正态性和开展统计推断提供了基础.

2.2 去中心化分布式统计学习

在中心化分布式统计学习中, 中心节点负责统一收集和处理其他本地节点的信息. 然而, 中心节点的存在会给系统带来隐患和缺陷. (1) 中心节点故障的风险. 由于所有节点依赖中心节点进行信息

聚合和计算, 一旦中心节点发生故障, 整个系统将停止工作. (2) 通信网络带宽的限制. 中心节点需要与所有其他节点交换信息, 随着网络规模的扩大, 中心节点的网络带宽可能成为瓶颈. (3) 隐私和安全问题. 在处理敏感信息 (如医疗和金融数据等) 时, 中心节点很有可能成为潜在攻击的目标, 导致数据泄露. 为了解决这些问题, 近些年来已有越来越多的研究关注去中心化的分布式统计学习方法. 这一框架中的核心问题在于, 在缺乏中心节点的条件下, 如何仅通过局部通信实现全局参数共识, 并进一步保证分布式算法输出能够充分逼近集中式估计量, 从而继承其统计效率与渐近性质. 需要说明的是, 去中心化框架中的全局经验风险最小化解 $\widehat{\boldsymbol{\theta}}$ 与集中式估计量对应同一优化目标, 因而在适当条件下具有相同的统计性质; 此时的关键并不在于重新定义统计目标, 而在于分析有限通信约束、网络结构和数据异质性会如何影响分布式算法对 $\widehat{\boldsymbol{\theta}}$ 的逼近精度及其统计推断能力.

在去中心化分布式统计学习中, 考虑有 m 个彼此相连的节点, 组成一个去中心化网络. 该网络通常可以由图 $\mathcal{G}=(\mathcal{V},\mathcal{E})$ 描述, 其中 $\mathcal{V}=\{1,2,\ldots,m\}$ 为节点集合, $\mathcal{E}\in\mathcal{V}\times\mathcal{V}$ 为边集合. 如果节点对 $(i,j)\in\mathcal{E}$, 则节点 i 与 j 之间是相互连接的, 可进行通信; 若 $(i,j)\notin\mathcal{E}$, 则节点 i 与 j 之间不能直接传输信息. 在这种场景下, 我们希望求解优化问题 (2.1). 当图 $\mathcal{G}$ 连通时, 可以将问题 (2.1) 写为如下等价的共识 (consensus) 优化问题:

$$\min_{\{\boldsymbol{\theta}_i\}_{i=1}^m}\frac{1}{m}\sum_{i=1}^m f_i(\boldsymbol{\theta}_i)\quad\text{s.t.}\quad\boldsymbol{\theta}_i=\boldsymbol{\theta}_j,\quad\forall\,(i,j)\in\mathcal{E},\tag{2.4}$$

其中 $\boldsymbol{\theta}_i$ 为节点 i 的参数. 围绕上述共识优化框架, 已有大量工作研究不同统计学习任务在去中心化网络中的实现, 包括去中心化平均 [30, 135, 136]、机器学习 [45, 87, 96]、稀疏优化 [79] 和矩阵填补 [80] 等.

与问题 (2.1) 相比, 问题 (2.4) 反映了去中心化学习首先需要解决的基本问题, 即在没有中心节点进行统一聚合与广播的情形下, 如何通过邻域通信实现全局一致性. 此时, 每个节点都存储自己的参数 $\boldsymbol{\theta}_i$, 并且在训练过程中必须通过节点之间的通信确保所有节点上的参数达成一致. 为了解决问题 (2.4), 近年来已有多项研究提出了高效的算法. 最基础的算法是去中心化平均算法 [13, 61, 135, 136]: 假设节点 i 收集的数据为 $\boldsymbol{x}_i$, 目标是得到所有节点平均值 $\overline{\boldsymbol{x}}=\sum_{i=1}^m\boldsymbol{x}_i/m$ 的估计. 令第 i 个节点在第 t 次通信时的估计为 $\widehat{\boldsymbol{x}}_i(t)$, 则加权线性迭代方法 [135] 的迭代步骤为

$$\widehat{\boldsymbol{x}}_i(t+1)=\sum_{j\in\mathcal{N}_i\cup\{i\}}W_{ij}\widehat{\boldsymbol{x}}_j(t),\tag{2.5}$$

其中 $j\in\mathcal{N}_i$ 表示节点 i 的邻居节点集合, W_{ij} 为节点 i 和 j 之间的权重, 满足对于所有 $(i,j)\notin\mathcal{E}$, $W_{ij}=0$, 否则 $W_{ij}>0$. 记

$$\widehat{\boldsymbol{x}}(t)=[\widehat{\boldsymbol{x}}_1(t),\widehat{\boldsymbol{x}}_2(t),\ldots,\widehat{\boldsymbol{x}}_m(t)]^\top,\quad\boldsymbol{W}=[W_{ij}]_{i,j=1}^m\in\mathbb{R}^{m\times m},$$

则迭代 (2.5) 可以简洁地表示为

$$\widehat{\boldsymbol{x}}(t+1)=\boldsymbol{W}\widehat{\boldsymbol{x}}(t).$$

Xiao 和 Boyd [135] 证明了 $\lim_{t\to\infty}\widehat{\boldsymbol{x}}(t)=\mathbf{1}_m\overline{\boldsymbol{x}}$ 的充分必要条件为: 矩阵 $\boldsymbol{W}$ 具有双随机 (doubly stochastic) 性质 (即 $\mathbf{1}_m^\top\boldsymbol{W}=\mathbf{1}_m^\top$, $\boldsymbol{W}\mathbf{1}_m=\mathbf{1}_m$), 并且最大奇异值 $\sigma_{\max}(\boldsymbol{W}-\mathbf{1}_m\mathbf{1}_m^\top/m)<1$. 最大奇异值越小, 收敛速度越快. 为了加快收敛, 一类启发式的构造方法是利用图的 Laplace 矩阵为每个边分配权重, 如 Metropolis 矩阵 [135] 和惰性 Metropolis 矩阵 [92] 等. 除此之外, Xiao 等 [136] 提出了一种通过最小化偏差来选择权重的方法. 上述文献主要解决了去中心化网络中如何通过局部通信实现全局信息融合的问题, 为后续分布式优化与统计估计提供了共识基础.

在实现参数共识的基础上, 下一步需要解决的问题是如何在去中心化网络中完成全局优化. 为此, 研究者将去中心化平均思想与传统优化方法结合, 提出了用于解决问题 (2.4) 的一系列去中心化优化算法, 包括去中心化 (随机) 梯度下降算法[75,93,155]、去中心化二阶算法[6,36,91]、去中心化对偶平均算法[27,34] 和去中心化交替方向乘子算法[4,109,159] 等. 下面主要介绍最常使用的去中心化 (随机) 梯度下降算法及其扩展 (即去中心化一阶算法). 令节点 i 在第 t 次通信时的参数估计为 $\boldsymbol{\theta}_i^{(t)}$, 则去中心化梯度下降 (decentralized gradient descent, DGD) 算法的迭代为

$$\boldsymbol{\theta}_i^{(t+1)} = \sum_{j\in\mathcal{N}_i\cup\{i\}} W_{ij}\boldsymbol{\theta}_j^{(t)} - \eta(t)\nabla f_i(\boldsymbol{\theta}_i^{(t)}), \tag{2.6}$$

其中 $\eta(t)$ 为步长. Nedic 和 Ozdaglar[93] 在一定假设下建立了目标函数值与最优值之间的误差上界. 此外, Yuan 等[155] 分析了 DGD 方法的收敛性: 在目标函数满足凸性、有下界和 Lipschitz 连续的条件下, 只要步长 $\eta(t)=\eta$ 足够小, 目标函数值以 $\mathcal{O}(1/t)$ 的速度收敛到 $\mathcal{O}(\eta)$ 的水平; 在目标函数强凸时, 所有局部估计 $\boldsymbol{\theta}_i^{(t)}$ 以线性速度收敛到最优值的 $\mathcal{O}(\eta)$ 邻域内. 这意味着只有在使用递减步长时, DGD 方法才能够收敛到精确的最优值.

近年来, 已有丰富的研究拓展和改进了 DGD 方法. Shi 等[108] 提出了精确一阶共识算法 (exact first-order algorithm, EXTRA), 与 DGD 相比, 该方法采用了额外的迭代, 使得在常数步长下, 针对强凸优化问题 (2.1), 每个节点的估计值以 $\mathcal{O}(C^{-t})$ 速度收敛到最优值. Qu 和 Li[98] 利用目标函数的光滑性和过往的梯度信息, 提出了一种带梯度追踪 (gradient tracking) 的去中心化算法, 使得算法在目标函数光滑时与中心化梯度下降相匹配. 以往的研究通常假设各节点之间的数据分布较为一致 (方差有界), 然而当数据分布差异较大时, 这些方法的收敛速度往往会减慢. 为了解决上述问题, Tang 等[115] 利用前两次迭代的信息和降方差更新方法, 提出了 D^2 算法, 提高了去中心化梯度下降算法在节点间数据分布差异较大时的收敛速度. 此外, Lian 等[76] 和 Wu 等[134] 研究了异步情形下 (即节点不同步计算和通信) 的去中心化梯度下降算法. Richards 和 Rebeschini[100] 及 Richards 等[101] 分析了基于非参数回归的去中心化梯度下降算法的收敛性. 总体而言, 这类去中心化一阶方法主要回答的是, 在有限通信轮次、给定网络拓扑和可能存在数据异质性的条件下, 分布式迭代能否有效逼近集中式解, 以及步长、网络结构与异质性如何影响这一逼近误差.

然而, 仅有算法收敛还不够, 更关键的问题是去中心化通信约束是否会影响全局估计的统计效率与渐近分布. 因此, 近年来已有研究开始系统分析去中心化分布式估计的统计性质与统计推断问题. 代表性的工作包括 Wu 等[133] 在网络弱平衡条件下, 给出了网络梯度下降算法的算法收敛结果, 并且针对一般的去中心化 M- 估计 $\widehat{\boldsymbol{\theta}}$, 建立了 $\mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$ 的最优统计效率. 此外, Wu 等[133] 指出影响去中心化梯度估计统计效率的三个主要因素为学习率、网络结构和节点间数据分布的差异. 另外, Zhang 等[161] 针对不同节点数据存在异质性的情形, 提出了一种基于分布式生成树的融合套索 (fused LASSO) 回归方法, 并对所得估计建立了统计一致性和渐近正态性的理论保证. 由于去中心化估计的渐近分布难以直接推导, 其渐近方差也难以在分布式网络中计算, 因而如何在此基础上进一步开展统计推断仍具有较大挑战性. 在统计推断方面, 现有研究主要包括基于 Wald 统计量的方法以及基于似然比或经验似然的方法. 在 Wald 型推断方面, Gu 和 Chen[48] 在去中心化框架下研究了统计推断问题, 在建立去中心化 Polyak-Ruppert 平均估计量渐近正态性的基础上, 构造了相应的置信域并给出了理论保证. 另一类重要思路是基于似然比或经验似然构造检验统计量. 例如, Du 和 Wang[31] 研究了去中心化网络中的经验似然推断方法, 提出了一种基于经验似然的去中心化算法. 通过引入 Lagrange 乘数与惩罚函数来求解全局似然比统计量, 并证明了即使在节点数量趋于无穷时, 所得似然比统计量仍渐近服

表 3 中心化与去中心化分布式统计学习典型方法比较

类别	方法	估计/优化误差	达到集中式统计效率的条件	通信复杂度
中心化	DC [102, 162]	$\mathcal{O}(N^{-1/2}+n^{-1})$	$m \leqslant \sqrt{N}$	$\mathcal{O}(mp)$
	一次微调 [56]	$\mathcal{O}(N^{-1/2}+n^{-2})$	$m \leqslant N^{3/4}$	$\mathcal{O}(mp^2)$
	CSL [60]	$\mathcal{O}(N^{-1/2}+n^{-(T+1)/2})$	$T \geqslant \lceil \log m/\log n \rceil$	$\mathcal{O}(Tmp)$
去中心化	DGD [93, 155]	$\mathcal{O}(\eta + C^{-T})$	$T \geqslant \log_C N/2$, $\eta \leqslant N^{-1/2}$	$\mathcal{O}(T\lvert\mathcal{E}\rvert p)$
	EXTRA [108]	$\mathcal{O}(C^{-T})$	$T \geqslant \log_C N/2$	$\mathcal{O}(T\lvert\mathcal{E}\rvert p)$
	DIGing [98]	$\mathcal{O}(C^{-T})$	$T \geqslant \log_C N/2$	$\mathcal{O}(T\lvert\mathcal{E}\rvert p)$

注: 表中误差阶表示分布式方法相对于集中式估计量的估计误差或优化误差. 当分布式附加误差不超过集中式统计误差 $\mathcal{O}(N^{-1/2})$ 时, 相应方法可达到与集中式方法同阶的统计效率. 去中心化方法部分主要基于强凸目标函数情形下的理论结果进行概括, 其中收敛速率常数 C 通常依赖于网络拓扑结构、权重矩阵和步长选择. 对于去中心化算法, 通过控制优化误差不超过统计误差 $\mathcal{O}(N^{-1/2})$ 可得到达到统计效率所需要的迭代次数.

从 χ^2 分布.

为更清晰地比较中心化与去中心化框架下典型方法的差异, 表 3 对若干代表性方法进行了归纳总结.

总体来看, 去中心化分布式统计学习的研究大体围绕两个层面展开. 第一类工作关注如何在仅允许邻域通信的条件下实现全局参数共识并逼近集中式最优解, 其重点在于共识机制、算法收敛速度以及网络拓扑对逼近误差的影响; 第二类工作则进一步研究分布式算法输出能否继承集中式估计量的统计效率与渐近分布, 并分析学习率、网络结构和数据异质性对统计推断的影响.

3 可信的分布式统计学习

分布式统计学习的发展使得无需集中数据即可进行大规模学习. 然而, 这种开放系统特性也带来了诸多新的挑战和风险. 由于分布式学习中的节点可能来自不同的地理区域, 数据在收集和处理过程中容易受到噪声和偏差等多种威胁. 同时, 模型训练和推理阶段也可能遭遇恶意攻击和隐私泄露. 因此, 可信分布式学习显得尤为重要, 它通常涉及稳健性、隐私保护和公平性三个核心方面 (参见文献 [132]), 接下来详细介绍相关研究进展及解决方案.

3.1 稳健的分布式统计学习

稳健的分布式统计学习面临诸多挑战, 其核心目标是在各种干扰和异常情形下, 仍能进行高效准确的参数估计和模型推断. 为此, 稳健的分布式统计学习主要划分如下: (1) Byzantine 稳健性; (2) 重尾噪声与离群值稳健性.

保证系统稳健性是分布式系统面临的固有挑战之一, Blanchard 等 [10] 指出, 将计算任务分布到多个节点上会引入更高的故障风险, 这些风险包括崩溃、计算错误、进程停滞和数据分布不均等. 在极端情形下, 系统还可能遭遇恶意攻击, 这些攻击试图破坏整个系统. 最稳健的系统应能够容忍 Byzantine 故障, 即允许部分节点表现出任意行为 (参见文献 [65]). 在这种背景下, 异常节点 (通常称为 Byzantine 节点) 可能通过发送恶意信息 (如伪造参数或篡改梯度) 破坏系统稳健性. 如图 2 所示, Byzantine 节点若注入错误的梯度信息, 可能使整个分布式机器学习过程偏离正确的优化路径, 从而显著降低模型性能. 为应对此类挑战, 越来越多的研究开始关注基于 Byzantine 稳健性的分布式统计学习方法, 以期构建在恶劣环境下依然具备高安全性的系统.

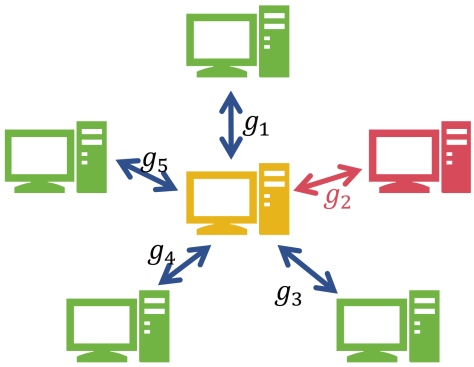


图 2 (网络版彩图) 中心化架构下的 Byzantine 攻击示意图. 大多数节点 (绿色) 上传的更新 g_1, g_3, g_4, g_5 为正常的局部梯度, 而 Byzantine 节点 (红色) 上传的 g_2 可能是任意伪造的信息, 从而干扰中心节点 (黄色) 在聚合时的方向, 导致全局模型偏移

除了系统遭受攻击外, 数据本身可能伴随重尾噪声, 并且可能还会存在离群值. 这种非理想的数据分布在许多系统中普遍存在, 并对模型的稳定性和推断精度构成严峻挑战. 在这种情形下, 如果采用基本的损失函数, 模型很难具备在真实数据下的泛化能力, 而且会导致参数估计的不稳健性. 因此, 研究者们关注在重尾噪声和离群值影响下, 如何优化统计模型并设计自适应方法, 有效减轻这些不利因素对系统性能的负面影响, 以实现稳健性. Koenker 和 Bassett [62] 首次提出了分位数回归 (quantile regression, QR), 其损失函数定义为

$$\rho_\tau(x) = x(\tau - \mathbb{I}\{x \leqslant 0\}),$$

其中 $\mathbb{I}$ 为示性函数, τ 为分位数水平. 分位数回归通过优化该损失函数, 能够估计数据的不同分位数. 从而避免了仅关注均值的局限性. 因此, 分位数回归能够更全面地描述数据的分布特征, 在面对重尾噪声和离群值时, 它比传统方法更具稳健性和灵活性, 因而受到了广泛关注 (参见文献 [9,74,126,165]).

3.1.1 Byzantine 稳健的分布式统计学习

在分布式统计学习中, Byzantine 节点可能通过发送错误参数或篡改梯度信息严重干扰模型更新, 进而破坏全局学习的准确性与稳定性. 因此, Byzantine 稳健分布式学习的核心问题在于, 当部分节点返回恶意或异常信息时, 分布式估计能否仍然保持一致性、最优统计效率和有效的统计推断能力, 以及这些性质如何依赖于 Byzantine 节点比例. 针对这一挑战, 研究者们提出了多种 Byzantine 稳健算法, 现有方法大致可分为中心化 [10,25,89,94,113,117,119,150] 和去中心化 [37,42,43,50,51,148] 两大类.

在中心化场景下, 系统中存在一个中心节点, 负责收集各工作节点上传的梯度或模型参数并进行全局聚合. 由于 Byzantine 节点可能故意上传异常梯度或篡改参数, 简单的均值聚合容易受到恶意信息的显著影响, 从而导致模型更新方向偏移甚至训练失效. 因此, 中心化 Byzantine 稳健方法的核心在于设计稳健聚合策略, 以削弱异常节点对全局更新的干扰. 若记 t 时刻的模型参数为 $\boldsymbol{\theta}^{(t)}$, 各节点上传的梯度为 $\{\boldsymbol{g}_i = \nabla f_i(\boldsymbol{\theta}^{(t)})\}_{i=1}^m$, 定义稳健聚合算子为 $\mathcal{A}$, 则其统一更新形式可写为

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \mathcal{A}(\boldsymbol{g}_i : i \in [m]).$$

现有中心化方法主要可分为两类. 第一类方法直接设计稳健聚合规则, 其目标是在聚合阶段抑制或过滤异常梯度的影响. 典型代表包括 Krum、坐标级中位数和截断均值等方法. 具体而言, Krum 方

法通过选择与多数节点梯度更一致的候选更新来降低异常节点的干扰 (参见文献 [10]); 坐标中位数和截断均值则利用逐坐标的稳健统计量削弱极端值的影响, 可以证明坐标中位数与截断均值两类聚合方法在强凸损失下可达到最优统计速率 (参见文献 [150]). 这类方法结构简单, 实现直接, 并在强凸情形下具有较好的理论保证. 除此之外, 中心化场景下还有不少围绕稳健聚合规则或优化策略展开的改进工作, 如文献 [3, 49] 等.

第二类方法采用分组聚合或稳健均值估计思想, 其核心目标是在存在异常节点时同时控制估计偏差与方差. 代表性工作包括 MOM (median-of-means) 及其改进方法. 此类方法首先将节点划分为若干组, 在组内进行均值汇总, 再在组间进行稳健聚合, 从而同时提升 Byzantine 稳健性 (参见文献 [25,89]). Chen 等 [25] 证明了, 该策略最多容忍 $f < m/2$ 个 Byzantine 节点, 且其估计误差为

$$\mathcal{O}\left(\sqrt{\max\{1, f\}\, p/N}\right),$$

相较于最优误差 $\mathcal{O}(\sqrt{p/N})$, 最多仅增加 $\sqrt{f}$ 阶. Minsker [89] 进一步证明了该估计量的渐近分布为

$$\sqrt{m}(\widehat{\mu} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \frac{\pi}{2}\sigma^2\right).$$

为了改进其方差系数 $\pi/2$, Tu 等 [119] 提出了 VRMOM (variance reduced median-of-means) 方法, 该方法利用多个分位数水平提升统计效率, 且在计算复杂度上与 MOM 方法相当. 通过这种方式, 可以在不增加计算复杂度的前提下使统计效率接近最优. 基于这一思想, Tu 等 [119] 还将相应的稳健估计策略与 Jordan 等 [60] 提出的 CSL 方法结合, 构造了 RCSL (robust communication-efficient surrogate likelihood) 方法, 在通信轮次不变的条件下实现了更快的收敛速率, 并提供了用于统计推断的渐近正态性结果.

在去中心化场景下, 由于缺乏中心服务器, 每个节点仅能与其邻居节点进行通信, 因此设计 Byzantine 稳健算法面临更大挑战. 具体而言, 在参数更新过程中, 每个节点需要综合考虑自身以及邻居节点的梯度和参数等信息, 而部分邻居可能是 Byzantine 节点, 这些节点可能发送篡改的信息, 进而破坏聚合过程. 为此, 研究者们提出了一系列去中心化 Byzantine 稳健方法. 例如, Fang 等 [42] 基于去中心化梯度下降算法 (2.6) 提出了 Byzantine 稳健的去中心化梯度下降框架. 假设节点 i 在时刻 t 收到来自其邻居传递的参数信息 $\{\boldsymbol{\theta}_j^{(t)}\}_{j\in\mathcal{N}_i}$, 选取稳健的聚合策略 $\mathcal{A}$, 则可以将迭代 (2.6) 转化为 Byzantine 稳健的形式:

$$\boldsymbol{\theta}_i^{(t+1)} = \mathcal{A}(\boldsymbol{\theta}_j^{(t)} : j \in \mathcal{N}_i \cup \{i\}) - \eta(t)\nabla f_i(\boldsymbol{\theta}_i^{(t)}).$$

Fang 等 [42] 表示, 上述聚合策略 $\mathcal{A}$ 可以选用如 Krum、中位数和截断均值等方法. 因此, 该框架具有良好的可扩展性, Fang 等 [42] 还提供了强凸问题和一类非凸问题的收敛保证. 尽管上述方法能够防御简单的 Byzantine 攻击, 但由于这些基于距离的策略往往存在脆弱性 (梯度之间的距离接近并不意味着其性能相似), 因此它们很容易在面对更复杂的 Byzantine 攻击时失效. 为了克服上述困难, Guo 等 [50] 提出了 UBAR (uniform Byzantine-resilient aggregation rule) 方法, 该方法通过两阶段过程来抵御 Byzantine 攻击. 为应对简单的 Byzantine 攻击, 每个节点首先根据其参数与邻居节点参数之间的距离, 选择若干潜在的正常节点. 随后, 每个节点利用自己的训练样本来测试所选参数的表现, 进一步排除那些会导致更大损失的邻居. 通过这种方式, UBAR 方法能够有效地抵御复杂的 Byzantine 攻击.

总而言之, 中心化方法主要依托中心节点设计稳健聚合策略或稳健均值估计机制, 其重点在于控制异常节点对全局更新方向和估计误差的影响; 去中心化方法则需要在局部通信约束下在各个节点处进行异常信息过滤.

3.1.2 重尾噪声与离群值稳健的分布式统计学习

由于重尾噪声和离群值源自数据本身, 这些问题在分布式统计学习中同样会被继承, 并可能进一步因样本拆分与有限通信而放大其对估计稳定性的影响. 因此, 重尾噪声与离群值稳健的分布式学习所关注的核心问题在于: 当误差分布偏离轻尾假设或样本中存在异常观测时, 分布式估计能否保持稳健性. 针对这一挑战, 研究者们在分布式场景下引入了分位数回归, 并提出了两类主要的方法, 包含中心化方法 [20, 22–24, 114, 121, 122] 和去中心化方法 [82, 90, 107, 124].

在中心化场景下, Volgushev 等 [121] 提出了利用分治法解决分位数回归问题. 首先, 在每个节点上得到分位数回归估计

$$\boldsymbol{\theta}_{\mathrm{QR},i} = \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{n} \rho_\tau(y_{ij} - \boldsymbol{\theta}^\top \boldsymbol{x}_{ij}),$$

聚合后得到估计量 $\overline{\boldsymbol{\theta}} = \sum_{i=1}^{m} \boldsymbol{\theta}_{\mathrm{QR},i}/m$. Volgushev 等 [121] 进一步证明了该估计量与单机情形下的估计量 $\widehat{\boldsymbol{\theta}}$ 之间满足

$$\|\overline{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\|_2 = \mathcal{O}_{\mathbb{P}}\left(\frac{m\log N}{N} + \frac{m^{1/4}(\log N)^{7/4}}{N^{3/4}}\right) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{N}}\right).$$

因此, 为了使估计量达到最优统计误差, 要求每台本地机器至少存储 $\mathcal{O}(N^{1/2}\log N)$ 的数据, 这在本地机器内存有限的情形下是不可行的, 而且该方法仍未能克服分治法对节点个数的限制, 因此无法应用于拥有大规模节点的网络. 为了克服这些困难, Chen 等 [24] 提出了 LEQR (linear estimator for quantile regression) 方法. 该方法不直接使用 $\boldsymbol{\theta}_{\mathrm{QR},i}$, 而是通过核光滑方法, 用函数 $K_h(x) = x(H(x/h) + \tau - 1)$ 近似 $\rho_\tau(x)$, 其中 $H(x/h)$ 为 $\mathbb{I}\{x > 0\}$ 的光滑近似. 随后, 利用一阶最优条件构建迭代式 $\widehat{\boldsymbol{\theta}}^{(t)} = (\sum_{i=1}^{m} \boldsymbol{V}_i)^{-1}(\sum_{i=1}^{m} \boldsymbol{U}_i)$, 其中 $\boldsymbol{V}_i$ 和 $\boldsymbol{U}_i$ 分别为节点 i 利用上一时刻计算的相应矩阵和向量. Chen 等 [24] 不仅指出该估计量只需要几轮迭代即可达到与单机相同的统计效率, 还建立了渐近正态性, 在 t 满足轻微假设下, 对于任意的 $\boldsymbol{v} \in \mathbb{R}^p$, 有

$$\sqrt{N}\frac{\boldsymbol{v}^\top(\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*)}{\sqrt{\boldsymbol{v}^\top \boldsymbol{D}^{-1}\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]\boldsymbol{D}^{-1}\boldsymbol{v}}} \xrightarrow{d} \mathcal{N}(0, \tau(1-\tau)), \quad n, N \to \infty,$$

并成功克服了分治法中的节点存储限制. 然而, LEQR 方法主要适用于低维分位数回归问题, 无法在高维情形下识别回归系数的稀疏性. 为了解决这一问题, Chen 等 [23] 提出了基于 ℓ_1- 正则化的分位数回归方法, 以保证在高维情境下回归系数的稀疏性. 为处理带有 ℓ_1 惩罚的分位数损失, Chen 等 [23] 通过将响应变量 y_i 转化为 $\widetilde{y}_i = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}}_0 - \widehat{f}^{-1}(0)(\mathbb{I}\{y_i \leqslant \boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}}_0\} - \tau)$, 其中 $\widehat{\boldsymbol{\theta}}_0$ 和 $\widehat{f}^{-1}(0)$ 均为代估参数, 进而将原稀疏分位数回归问题转化为 $(\boldsymbol{x}_i, \widetilde{y}_i)$ 的稀疏线性回归问题, 并通过近似 Newton 法迭代得到估计量 $\widehat{\boldsymbol{\theta}}^{(t)}$. Chen 等 [23] 证明了, 在一定条件下, 估计量满足

$$\|\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{s\log N}{N}} + s^{\frac{2t+1}{2}}\left(\frac{\log N}{n}\right)^{\frac{t+1}{2}}\right),$$

其中 s 为 $\boldsymbol{\theta}^*$ 的稀疏水平. 因此在 t 很小时, 该估计量能达到与单机情形相同的统计效率 $\mathcal{O}_{\mathbb{P}}(\sqrt{s\log N}/N)$. 此外, 也有基于其他模型的分位数回归研究. Chen 等 [22] 研究了稳健的降秩回归 (reduced rank regression, RRR) 问题, 所提出的分布式稳健低秩估计量方法达到了接近最优的速率, 同时为该估计量的秩一致性提供了理论保证.

在去中心化场景下, Wang 和 Li [124] 研究了分位数回归问题, 并将其扩展到高维情形. 他们通过去中心化次梯度下降方法求解该问题, 并证明了算法的收敛性. 然而, 该方法未能提供快速收敛的理论保障. 为改进上述结果, Liu 等 [82] 提出了 deSMR (decentralized surrogate median regression) 方法来解决该问题. 首先, 通过类似于 Chen 等 [23] 提出的方法, 将原始问题转化为稀疏线性回归问题, 然后利用去中心化的算法求解. 考虑到去中心化场景中参数共识的需求, Liu 等 [82] 指出, 这两个步骤需要通过双层迭代进行. 外层迭代在每步中将原始损失函数转化为稀疏线性回归问题, 内层迭代则使用去中心化交替方向乘子算法求解该问题. 假设外层迭代进行 v 次、内层迭代进行 t 次时得到的估计量为 $\boldsymbol{\theta}_i^{v,t}$. 他们首先证明了内层迭代的收敛性, 对于任意 v, 都有

$$\|\boldsymbol{\theta}_i^{v,t} - \widehat{\boldsymbol{\theta}}^v\|_2 = \mathcal{O}_{\mathbb{P}}(\gamma^t),$$

其中 $\widehat{\boldsymbol{\theta}}^v$ 为外层迭代 v 次后全局损失函数的最优值, $\gamma \in (0,1)$. 进一步地, 他们证明, 在一定条件下, 最终的估计量满足以下收敛性:

$$\|\boldsymbol{\theta}_i^{v,t} - \boldsymbol{\theta}^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{s\log N}{N}} + s^{\frac{2v+1}{2}}\left(\frac{\log N}{n}\right)^{\frac{2v+1}{2}} + \gamma^t\right).$$

因此, 在 v 和 t 均不需要很大时, 即为最优统计效率.

总而言之, 针对重尾噪声与离群值的情形, 现有分布式统计学习方法通过分位数回归、分治策略及多轮迭代等技术, 在保证稳健性的同时实现了接近集中式方法的统计效率与渐近推断性质.

3.2 保护隐私的分布式统计学习

在分布式统计学习中, 虽然局部节点没有直接共享自己的本地数据, 但是攻击者仍然可以通过分析通信数据 (如本地参数和梯度等), 挖掘本地节点中的私人信息 (参见文献 [86]). 例如, Shokri 等 [110] 指出, 攻击者仅通过访问一个经过训练的机器学习模型参数, 就能推断样本所对应的个体是否为该医院的患者, 从而侵犯个人隐私. 因此, 保护隐私的分布式统计学习所关注的核心问题在于: 在限制信息泄露的同时, 分布式估计能否仍保持与非隐私方法相当的统计效率、渐近性质与统计推断能力, 以及隐私预算、通信机制和样本规模会如何影响这一权衡关系. 为了防止攻击者通过本地节点传递的模型参数或者模型输出对原始数据进行反向推断, 近年来, 已经有越来越多的研究开始关注基于隐私保护的分布式统计学习.

为了系统研究隐私约束下分布式学习的统计性质, 首先需要一个严格的数学框架来刻画隐私保护标准. 在众多框架中, 差分隐私 (differential privacy, DP [35]) 是最常用的一类标准. 具体而言, 假设 $\mathcal{A}$ 是一个随机算法, 则算法 $\mathcal{A}$ 被称为 (ϵ,δ)- 差分隐私, 当且仅当对于所有仅相差一个元素的数据集 $\mathcal{D}_1$ 和 $\mathcal{D}_2$, 以及 $\mathcal{A}$ 值域中的所有子集 $\mathcal{S}$, 有

$$\mathbb{P}(\mathcal{A}(\mathcal{D}_1) \in \mathcal{S}) \leqslant \mathrm{e}^{\epsilon}\mathbb{P}(\mathcal{A}(\mathcal{D}_2) \in \mathcal{S}) + \delta. \tag{3.1}$$

满足差分隐私的算法在某个训练样本发生变化时, 输出端产生的变化在统计意义上是微小的, 这能有效阻止攻击者通过算法的输出推测样本的变化情况, 从而实现隐私保护. 在经典差分隐私框架的基础上, 一些研究拓展了差分隐私的概念, 以适应不同应用场景. 在分布式学习背景下, Liu 等 [83] 提出了分布式分组差分隐私 (distributed group differential privacy, DGDP) 的概念, 将 DP 定义中 “仅相差一个元素的数据集” 拓展为了 “仅在一个局部节点上不同的数据集”, 保护了本地节点上所有数据的隐私,

更适用于分布式学习的场景. 接下来, 在差分隐私的框架下, 分别介绍中心化和去中心化的分布式学习方法.

在中心化框架中, 每个本地节点希望从两个方面保护本地数据的隐私 (参见文献 [95]): (1) 本地节点不愿意与其他节点或第三方共享本地原始数据; (2) 在中心节点聚合参数时, 无法通过传输到中心节点的本地参数反推本地数据的信息. 正如第 2.1 小节中所介绍的, 中心化分布式学习的流程可以总结为局部训练、参数传输、参数聚合和参数广播. 中心化 DP 算法的基本思想是在上述过程中引入随机噪声, 使得攻击者无法从传输的参数中推断出原始数据, 从而实现隐私保护. 根据噪声添加的位置, 中心化分布式 DP 算法可分为输出扰动 [57,95] 和梯度扰动 [57,116,131,147] 两种.

输出扰动的想法最为直观, 已在传统的 DP 研究中得到充分探讨 (参见文献 [19,35]). Pathak 等 [95] 首次在分布式的场景下基于输出扰动的想法提出了一种差分隐私的分布式逻辑回归算法. 进一步地, Jayaraman 等 [57] 将分布式差分隐私算法扩展到一般的经验风险最小化问题. 具体而言, 假设有 m 个节点, 第 i 个节点上需要优化含有正则项的经验损失函数

$$\sum_{j=1}^{n_i} f(\boldsymbol{\theta}; \boldsymbol{x}_{ij}, y_{ij}) + \lambda\Omega(\boldsymbol{\theta}),$$

其中 $\Omega(\boldsymbol{\theta})$ 表示正则项. 假设在第 t 次迭代中得到的本地模型参数为 $\boldsymbol{\theta}_i^{(t)} \in \mathbb{R}^p$. 为了保护隐私, 在中心节点上使用如下的扰动聚合参数:

$$\widehat{\boldsymbol{\theta}}_{\text{priv}}^{(t)} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\theta}_i^{(t)} + \xi,$$

其中 ξ 表示随机噪声. Jayaraman 等 [57] 指出, 当损失函数 f 是 L-Lipschitz 光滑并且数据有界时, 只需添加的噪声满足 $\xi \sim \text{Lap}(2L/mn_{(1)}\lambda\epsilon)$, 则聚合估计量 $\widehat{\boldsymbol{\theta}}_{\text{priv}}^{(t)}$ 满足 ϵ- 差分隐私, 其中 Lap 为 Laplace 分布, $n_{(1)} = \min\{n_1, \ldots, n_m\}$ 表示所有节点上最小的样本量. 此外, Jayaraman 等 [57] 还给出了在差分隐私限制下, 所得估计量的经验误差以及期望经验误差的上界.

输出扰动的计算效率较高, 仅需要一次安全的多方计算来聚合模型并添加噪声. 然而, 输出扰动适用于数据分布均匀的场景, 但当参与方数量较多且每个参与方的数据量较小时, 模型的准确性可能下降. 在这种场景下, 可以采用梯度扰动的方式来实现差分隐私. 具体而言, 在每次迭代中, 每个节点通过安全的多方计算聚合本地梯度, 并在聚合后的梯度上添加上一个 Gauss 噪声, 以确保差分隐私. Jayaraman 等 [57] 指出, 在 t 次迭代之后, 若向聚合梯度中加入服从多元 Gauss 分布 $\mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}_p)$ 的噪声, 其中 $\sigma^2 = \{8L^2t\log(1/\delta)\}/\{m^2n_{(1)}^2\epsilon^2\}$, 则得到的参数估计值是 (ϵ, δ)- 差分隐私的. 虽然梯度扰动能够提供更为完善的 (ϵ, δ)- 差分隐私, 并且适用于数据分布不均匀、节点数量较高的情形, 但相应地, 它比输出扰动带来更大的计算开销.

在去中心化的分布式学习框架中, 与中心化场景类似, 本地节点同样需要在参数交换和模型更新过程中保护数据隐私. 由于去中心化算法通常通过迭代方式进行, 因此去中心化差分隐私算法的核心思想是在每次迭代的通信和更新过程中引入随机噪声, 从而防止攻击者通过传递的信息 (如梯度或模型参数) 反推本地数据的敏感信息. 近年来, 许多研究提出了不同的去中心化差分隐私机制 [8,26,128,129,142,151,157,160].

例如, Yu 等 [151] 基于去中心化随机梯度下降 (见 (2.6) 的随机梯度版本) 提出了一种差分隐私算法 DP^2-SGD (decentralized parallel stochastic gradient descent with DP). 在该方法中, 每个节点 i 在第 t 次迭代时持有本地模型参数 $\boldsymbol{\theta}_i^{(t)}$ 和一个随机扰动变量 $\boldsymbol{\xi}_i^{(t)}$. 在每轮迭代中, 节点首先对本地模型

进行更新 $\boldsymbol{\theta}_i^{(t+1/2)} = \boldsymbol{\theta}_i^{(t)} - \eta \boldsymbol{g}_i^{(t)}$, 随后在更新后的参数 $\boldsymbol{\theta}_i^{(t+1/2)}$ 上叠加 Gauss 噪声 $\boldsymbol{\xi}_i^{(t)}$, 并将扰动后的参数发送给其邻居节点. 随后, 每个节点利用接收到的邻居扰动参数更新自身模型:

$$\boldsymbol{\theta}_i^{(t+1)} = \boldsymbol{\theta}_i^{(t+1/2)} + \sum_{j \in \mathcal{N}_i} (W_{ij} - I_{ij})(\boldsymbol{\theta}_j^{(t+1/2)} + \boldsymbol{\xi}_j^{(t)}),$$

其中 $\boldsymbol{W}$ 为双随机矩阵, $\boldsymbol{I}$ 为单位矩阵. 进一步地, 他们证明了 DP^2-SGD 的收敛速率为 $\mathcal{O}(1/\sqrt{mT})$, 其中 T 表示总迭代次数, 该收敛速率达到了与非隐私情形下相同的最优速率. 同时, 他们还指出只要 Gauss 噪声满足 $\sigma \geqslant CT\eta\sqrt{2\log(1.25/\delta)}/\epsilon$, 则整个信息交换过程满足 (ϵ, δ)- 差分隐私, 其中 $C = \max_{i,t} \boldsymbol{g}_i^{(t)}$. 类似地, Xu 等 [142] 提出了一种结构相近的方法, 区别在于该方法并非直接在交换的模型参数上添加噪声, 而是在梯度上注入噪声, 并同样证明了算法满足 (ϵ, δ)- 差分隐私. 然而, 上述两类方法通常依赖于无向且固定的通信拓扑, 这一假设在实际应用中往往过于理想化. 为了适应更为复杂的场景, 已有学者开始研究更一般的通信结构, 如有向图 [69, 129] 和动态图 [167] 等.

总体来看, 保护隐私的分布式统计学习并不只是研究如何在算法中加入随机噪声, 更关键的是刻画隐私约束、通信机制与统计效率之间的权衡关系.

3.3 公平的分布式统计学习

公平性是可信分布式学习的重要组成部分, 其核心目标是避免模型在决策过程中对特定群体产生系统性的不公正结果. 导致不公平决策的因素主要来自两个方面. 一方面, 数据偏差可能在数据收集和预处理阶段被引入. 例如, 在电子产品消费数据中, 由于性别差异, 男性购买比例可能显著高于女性; 在护理学校录取数据中, 性别和年龄的影响可能导致年轻女性比例偏高. 这类数据分布的偏斜揭示了潜在的社会与人口学偏见, 从而影响数据的客观性. 此外, 数据预处理过程也可能进一步放大偏差, 例如, Simpson 悖论表明, 当总体数据被划分为子群体时, 变量之间的相关关系可能出现、消失或反转 (参见文献 [11]). 另一方面, 算法本身的设计选择也可能成为偏差的来源. 若模型在训练过程中受到优化策略、正则化方式或有偏估计等因素的影响 (参见文献 [5]), 则在类别分布不均衡的情形下, 往往会进一步放大主流类别与少数类别之间的差异. 因此, 公平的分布式统计学习所关注的核心问题在于: 在多节点、异质数据和有限通信的条件下, 分布式估计能否在保持预测精度的同时满足给定的公平性准则, 并进一步刻画公平性约束对统计效率、泛化性能与群体差异的影响.

在诸多公平性定义中, 群体公平性 (group-based fairness) 是最常见的一类. 具体而言, 假设训练数据的形式为 $(\boldsymbol{X}, S, Y)$, 其中 $\boldsymbol{X}$ 为特征, S 为敏感属性, Y 为标签. 群体公平性的典型定义包括:

(1) 人口平等 (demographic parity, DP$_{\text{fair}}$), 要求模型预测 $\widehat{Y}$ 与敏感属性 S 统计独立, 即

$$\mathbb{P}(\widehat{Y} = y \mid S) = \mathbb{P}(\widehat{Y} = y);$$

(2) 均等机会 (equalized odds, EO), 给定标签 Y 时, 要求模型的预测 $\widehat{Y}$ 与敏感属性 S 统计独立, 即

$$\mathbb{P}(\widehat{Y} = y \mid S, Y) = \mathbb{P}(\widehat{Y} = y \mid Y).$$

这些定义为公平性提供了清晰的数学度量, 并为分布式公平学习的研究奠定了理论基础.

现有的分布式公平学习研究主要集中在中心化架构中. 在该框架下, 研究的核心在于如何将公平性约束或公平性度量有效融入分布式优化过程, 从而在维持全局模型预测性能的同时控制不同群体之间的决策差异. 现有方法大体可分为约束优化方法 [28]、惩罚方法 [55, 112, 123] 和重加权方法 [32, 39] 三类.

(1) 第一类方法将公平性直接作为约束纳入优化模型, 其目标是在优化风险函数的同时显式控制公平性偏差. 典型问题可表述为

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} f(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\boldsymbol{\theta}) \quad \text{s.t.} \ \ \boldsymbol{g}(\boldsymbol{\theta}) \leqslant \boldsymbol{\epsilon},$$

其中 $\boldsymbol{g}(\boldsymbol{\theta})$ 表示公平性度量函数, 其具体形式取决于所采用的公平性准则. 这类方法所解决的核心问题是, 在给定公平约束下, 分布式模型能否在所有节点上同时实现可接受的预测性能与公平表现. 例如, Cui 等 [28] 提出了多客户端公平性 (multi-client fairness, MCF), 并基于 $\mathrm{DP}_{\mathrm{fair}}$ 和 EO 两类公平准则, 从约束多目标优化的角度构建了能够在所有节点上同时满足公平约束并保持一致性能的模型. (2) 第二类方法采用惩罚思想, 通过 Lagrange 乘子法将公平约束转化为目标函数中的惩罚项, 从而将约束优化问题改写为如下无约束优化问题:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{m}\sum_{i=1}^{m} f_i(\boldsymbol{\theta}) + \boldsymbol{\lambda}^{\top}\boldsymbol{g}(\boldsymbol{\theta}),$$

其中 $\boldsymbol{\lambda}$ 为 Lagrange 乘子. 这类方法的核心在于刻画精度与公平之间的权衡关系, 即通过调节惩罚强度控制模型在风险最小化与公平性之间的平衡. 例如, Su 等 [112] 通过优化上述带惩罚项的目标函数并缓解节点间的梯度冲突, 实现了同时兼顾群体公平性 ($\mathrm{DP}_{\mathrm{fair}}$ 和 EO) 与客户端公平性 (client fairness) 的方法; Wang 等 [123] 研究了全局公平性与局部公平性之间的关系, 并证明了全局公平性指标仅依赖于各节点的统计量, 据此提出了能够确保全局 $\mathrm{DP}_{\mathrm{fair}}$ 与 EO 的公平算法; Hu 等 [55] 则针对有界群体损失 (bounded group loss, BGL) 这一公平准则提出了相应的优化方法. (3) 第三类方法采用重加权思想, 通过对节点或样本赋予差异化权重, 使模型在训练过程中更关注弱势群体或表现较差的群体, 从而降低群体之间的性能差距. 这类方法主要解决的是, 当不同节点或不同群体的数据质量、样本规模和群体分布存在显著差异时, 如何通过调整训练贡献来提升整体公平性. 例如, Ezzeldin 等 [39] 提出了一种基于节点层面的加权聚合策略. 在每轮迭代中, 中心节点将上一轮参数 $\boldsymbol{\theta}^{(t)}$ 广播至所有节点, 节点 i 完成局部更新并得到 $\boldsymbol{\theta}_i^{(t)}$ 后将其发送回中心节点. 随后, 服务器根据节点上传的局部公平性指标与全局公平目标之间的关系, 自适应地分配聚合权重 $w_i^{(t)}$, 从而调整各节点在全局模型更新中的贡献, 更新形式为

$$\boldsymbol{\theta}^{(t+1)} = \sum_{i=1}^{m} w_i^{(t)}\boldsymbol{\theta}_i^{(t)} \Big/ \sum_{i=1}^{m} w_i^{(t)}.$$

另一类重加权方法则基于样本层面. Du 等 [32] 引入参数 $\boldsymbol{\alpha}$ 与基函数 $\boldsymbol{B}(\boldsymbol{x})$, 定义样本 $\boldsymbol{x}$ 的权重为

$$w_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \boldsymbol{B}(\boldsymbol{x})^{\top}\boldsymbol{\alpha},$$

并基于此修正损失函数与公平性度量, 构造了如下受约束的最小最大优化问题:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\max_{0\leqslant\boldsymbol{\alpha}\leqslant C}\frac{1}{m}\sum_{i=1}^{m} f_i(\boldsymbol{\theta},\boldsymbol{\alpha}) \quad \text{s.t.} \ \ \boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{\alpha}) \leqslant \boldsymbol{\epsilon}, \quad \frac{1}{m}\sum_{i}^{m}\sum_{j}^{n_i} w_{\boldsymbol{\alpha}}(\boldsymbol{x}_{ij}) = 1.$$

该方法通过在各节点固定 $\boldsymbol{\theta}$ 以求解内层问题, 在中心节点固定 $\boldsymbol{\alpha}$ 以优化外层问题, 从而实现对公平性与预测性能的联合调节.

总体来看, 公平的分布式统计学习并不只是研究如何在分布式训练中加入公平约束, 更重要的是刻画公平性与统计效率之间的关系. 无论是约束优化、惩罚方法还是重加权方法, 其本质都在于通过不同机制控制群体之间的差异与预测偏差, 从而在保持模型精度的同时满足给定的公平性标准.

4 其他相关研究

4.1 半监督分布式统计学习

在大数据时代, 如何分析海量未标注数据 (如电子病历、文本数据等) 成为了许多实际应用中的核心问题. 由于标注数据的获取成本高昂且耗时, 标注数据的稀缺性使得利用未标注数据成为解决这一问题的关键. 半监督学习 (semi-supervised learning, SSL) 作为一种能够有效结合有限标注数据与丰富未标注数据的学习策略, 已在提升机器学习模型的预测精度方面取得显著成效. 半监督分布式学习所关注的核心问题在于: 在标注样本有限而未标注样本丰富的条件下, 分布式估计能否充分利用未标注数据来提升统计效率, 并进一步刻画未标注样本和通信轮次对估计精度和渐近性质的影响. 随着数据规模的日益增长, 传统的半监督学习方法面临着计算和存储瓶颈, 因此, 分布式半监督学习 (distributed semi-supervised learning, DSSL) 成为了研究的热点, 主要分为中心化方法 [18, 118, 120, 130] 和去中心化方法 [44, 103, 137, 138].

在中心化架构下, 一个基本问题是, 如何在分治或多轮通信的框架下有效引入未标注数据, 使分布式估计达到比纯监督方法更高的统计效率. 针对这一问题, Chang 等 [18] 基于分治策略研究了分布式半监督场景下的核岭回归, 并提供了该方法的误差分析. 该研究表明, 半监督分治方法能够显著受益于未标注数据, 并且相较于普通的分治方法, 能够支持更灵活的节点量配置. 除了一次方法外, 允许适当迭代的半监督分布式方法也受到了广泛关注. Tu 等 [120] 探讨了在无标签数据分布于中心节点、其余节点仅具有有标签数据的情形下的高维 M- 估计和高维广义线性模型的半监督分布式问题. 他们提出了一种高效的多轮分布式去偏估计量, 并证明了高维 M- 估计在 t 轮迭代后参数的收敛性, 即

$$\|\widehat{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\|_2=\mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{s\log p}{mn}}+r_n s^{\frac{3t}{2}}\left(\frac{\log p}{n^*}\right)^{\frac{(1-q)t}{2}}+s^{-\frac{1}{2}}(\sqrt{s}r_n)^{2^t}\right),$$

其中 n 为每个节点上的有标签样本数, n^* 为中心节点上的总样本数, s, r_n 和 q 均由相应假设给出. 当 t 足够大时, 后两项可被第一项控制, 此时存在 σ_i, 使得

$$\frac{\sqrt{mn}}{\sigma_i}(\widehat{\boldsymbol{\theta}}_i^{(t)}-\boldsymbol{\theta}_i^*)\xrightarrow{d}\mathcal{N}(0,1).$$

针对广义线性模型, Tu 等 [120] 也得到了类似结论. 特别地, 该方法的统计精度达到

$$\sqrt{\frac{1}{mn}}+\frac{1}{\sqrt{n}}\left(\frac{1}{n^*}\right)^{t/2},$$

而不利用未标注样本的分治方法仅能达到

$$\sqrt{\frac{1}{mn}}+\frac{1}{n},$$

因此, 半监督方法能够有效放松分治框架下对节点数量 m 的限制; 当 $t=1$ 时, 仅需要满足 $m\leqslant n^*$ 即可实现 $\mathcal{O}(1/\sqrt{N})$ 的收敛效率. 进一步地, Tu 等 [118] 放宽了未标记数据仅存储于中心节点的要求, 假设这些数据分布在一个节点集合 $\mathcal{U}$ 上, 并在此基础上系统研究了分布式半监督广义线性模型. Tu 等提出了基于方差校正技术的代理损失函数, 并说明由此得到的估计量 $\widehat{\boldsymbol{\theta}}^{ss}$ 满足如下渐近正态性: 对任意 $\boldsymbol{v}\in\mathbb{S}^{p-1}$, 存在 $\sigma(\boldsymbol{v})$, 使得

$$\frac{\sqrt{mn}}{\sigma(\boldsymbol{v})}\boldsymbol{v}^\top(\widehat{\boldsymbol{\theta}}^{ss}-\boldsymbol{\theta}^*)\xrightarrow{d}\mathcal{N}(0,1).$$

此外, Tu 等 [118] 进一步提出了一种迭代求解方法, 并证明了相应迭代估计量满足

$$\|\widehat{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\|_2=\mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}}+r_n\left(\frac{p\log n^*}{n^*}\right)^t+r_n\left(\frac{p\log n^*}{|\mathcal{U}|n^*}\right)^{\frac{t}{2}}+\frac{1}{\sqrt{p}}(\sqrt{p}r_n)^{2^t}\right),$$

其中 n^* 为每个具有未标记样本节点上的总样本数, r_n 来源于相应假设. 在上述结果中, 第一项反映了标记数据下几乎最优的统计收敛率, 而第二和三项则体现了未标记样本和多轮迭代所带来的额外增益. 当 $n^* \gg n$ 时, 该方法的收敛速度明显快于不利用未标记样本的算法. 此外, Wang 等 [130] 首次在分块缺失协变量 (block-wise missing covariates) 的设定下研究了广义线性模型的分布式半监督问题, 并提出了相应的去偏估计量. 还有很多文献也研究了半监督分布式统计学习问题, 包含文献 [29, 58, 71, 77, 158, 164] 等.

在去中心化架构下, 除了利用未标注数据本身, 还需要同时处理局部通信与参数共识所带来的额外约束, 因而其核心问题在于如何在网络一致性要求下充分发挥未标注数据的统计价值. 针对这一问题, Scardapane 等 [103] 研究了半监督支持向量机, 并将无标签数据的损失函数近似为 $\exp(-5g(\boldsymbol{x}))$, 其中 $g(\boldsymbol{x})=\boldsymbol{w}^\top\boldsymbol{x}+b$ 为支持向量机的决策超平面. 基于这一代理损失函数, Scardapane 等 [103] 采用去中心化梯度下降方法对总体目标函数进行优化. Fierimonte 等 [44] 则提出了分布式 Laplace 正则化最小二乘方法. 基于流形正则化 (manifold regularization, MR) 的思想, Fierimonte 等 [44] 设计了基于 Euclid 距离矩阵补全的技术, 从而实现去中心化半监督学习. 然而, 由于该方法需要估计所有节点样本的全局 Euclid 距离矩阵, 在大规模数据场景下其计算代价较高. 为克服这一局限, Xie 等 [138] 提出了一种新的基于流形正则化的去中心化半监督学习算法, 将小波神经网络 (wavelet neural network, WNN) 与零梯度和 (zero-gradient-sum, ZGS) 分布式优化策略相结合. 在此基础上, Xie 等 [137] 又将该思想推广到单层前馈神经网络 (single layer feed-forward neural network, SLFNN).

总体来看, 半监督分布式统计学习的核心并不只是如何在分布式环境中引入未标注样本, 更重要的是刻画未标注数据、通信机制与统计效率之间的关系. 在中心化框架下, 现有研究主要表明, 通过适当的迭代修正与去偏技术, 未标注数据能够显著提升统计效率, 放宽节点数量限制, 并支持渐近正态性与统计推断; 在去中心化框架下, 相关研究则进一步讨论了如何在局部通信和参数共识约束下发挥未标注数据的作用.

4.2 分布式在线统计学习

传统的批量学习方法要求在学习任务开始之前提供完整的训练数据集, 并且训练过程通常是在离线环境中进行. 因此该方法存在一些关键缺点: (1) 时间和空间成本较高, 效率低下; (2) 对大规模应用的扩展性差, 因为每当有新的训练数据时, 模型通常需要从头开始重新训练 (参见文献 [52, 72]). 与之相对, 在线学习是一种针对按顺序到达的数据进行学习的方法. 每当新的数据到达时, 节点会在当前步骤中更新模型, 逐步优化对未来数据的预测. 通过这种方式, 在线学习能够克服批量学习的缺点, 因为预测模型可以实时更新, 及时适应新的数据实例. 因此, 在线学习算法在大规模机器学习任务中比传统的批量学习方法更具效率和可扩展性, 尤其适用于现实世界中的数据分析任务, 这些任务不仅数据量巨大, 而且数据到达速度非常快. 从统计角度看, 分布式在线统计学习所关注的核心问题在于: 在数据持续到达、损失函数随时间变化且节点仅掌握局部信息的条件下, 分布式算法能否实现与集中式在线方法相当的学习性能, 并进一步刻画网络结构、通信机制与时间变化环境对遗憾界和动态适应能力的影响.

由于在线学习具有上述优势, 近年来分布式在线学习得到了广泛关注. 分布式在线学习的主要研究目标是如何在分布式系统中高效处理和分析持续到达的大规模数据流, 尤其是在每个节点仅拥有局部数据并且需要依赖邻居节点信息进行决策的情形下. 具体而言, 分布式在线学习通常旨在解决如下问题:

$$\min_{\{\boldsymbol{\theta}_i\}_{i=1}^m} \frac{1}{m}\sum_{t=1}^{T}\sum_{i=1}^{m} f_{i,t}(\boldsymbol{\theta}_i) \quad \text{s.t. } \boldsymbol{\theta}_i = \boldsymbol{\theta}_j, \quad \forall\,(i,j)\in\mathcal{E}, \tag{4.1}$$

其中 $f_{i,t}$ 表示节点 i 在时间步 t 对应的数据流所诱导的局部损失函数. 在线学习算法的性能通常通过遗憾 (regret) 来衡量. 遗憾刻画的是算法在给定时间区间内的实际决策与理想决策之间的差距, 最常用的指标包括静态遗憾 (static regret) 和动态遗憾 (dynamic regret) [168], 分别定义为

$$\text{S-Reg} = \sum_{t=1}^{T}\sum_{i=1}^{m} f_{i,t}(\boldsymbol{\theta}_i^{(t)}) - \sum_{t=1}^{T} f_t(\boldsymbol{\theta}^*),$$
$$\text{D-Reg} = \sum_{t=1}^{T}\sum_{i=1}^{m} f_{i,t}(\boldsymbol{\theta}_i^{(t)}) - \sum_{t=1}^{T} f_t(\boldsymbol{\theta}_t^*),$$

其中 $\boldsymbol{\theta}_i^{(t)}$ 为第 i 个节点在时刻 t 的估计, $f_t = \sum_{i=1}^m f_{i,t}$ 为时刻 t 的总损失, $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^T f_t(\boldsymbol{\theta})$ 为全局静态最优决策, $\boldsymbol{\theta}_t^* = \arg\min_{\boldsymbol{\theta}} f_t(\boldsymbol{\theta})$ 为时刻 t 的全局动态最优决策. 根据这些定义, 一个有效的在线学习算法通常要求其遗憾为次线性阶 $o(T)$. 因而, 现有研究主要围绕三个层面展开: 一是理论保障, 即不同算法所能达到的遗憾上界 (参见文献 [2, 38, 53, 64, 88, 104, 145]); 二是网络通信机制, 即有限通信、反馈压缩或特殊拓扑对在线学习性能的影响 (参见文献 [14, 54, 66, 67, 70, 73, 144, 154]); 三是方法在具体问题中的应用表现 (参见文献 [16, 84, 104, 153]).

首先, 分布式在线学习的主要困难在于各节点的损失函数会随时间变化, 并且在任意时刻 t, 每个节点只能观察到截至 t 为止的历史损失, 而对未来时刻 $(t, T]$ 内的损失函数一无所知. 因此, 理论研究的核心在于建立静态遗憾和动态遗憾的上界, 以说明算法在时变环境下仍具有有效的学习能力. 例如, Yan 等 [145] 基于分布式梯度下降方法提出了如下参数更新策略:

$$\boldsymbol{\theta}_i^{(t+1)} = \mathcal{P}\Big(\sum_{j\in\mathcal{N}_i} W_{ij}\boldsymbol{\theta}_j^{(t)} - \eta(t)\nabla f_{i,t}(\boldsymbol{\theta}_i^{(t)})\Big),$$

其中 $\mathcal{P}$ 表示投影算子, 用于将参数投影到可行域中. Yan 等 [145] 证明了, 在强凸条件下, 该方法的静态遗憾为 $\mathcal{O}(\log T)$, 而在一般凸条件下为 $\mathcal{O}(\sqrt{T})$, 均达到相应情形下的最优阶. 在此基础上, Shahrampour 和 Jadbabaie [104] 进一步引入相邻时刻最优参数之间的变化限制, 提出了去中心化在线镜像下降方法 (decentralized online mirror descent, DOMD), 并证明了在凸条件下动态遗憾达到 $\mathcal{O}(\sqrt{T(1+C_T)})$, 其中 C_T 为共同知识路径长度. 进一步地, Eshraghi 和 Liang [38] 将多重平均决策与梯度一致策略引入 DOMD, 并证明了在强凸条件下可实现 $\mathcal{O}(1+P_T)$ 阶的动态遗憾, 其中 P_T 为路径长度.

其次, 分布式在线学习可以根据节点间交换信息的类型和通信拓扑进行进一步分类. 在信息交换类型方面, 前述文献通常假设节点间交换完整的梯度或参数信息. 然而, 也有一系列研究考虑更弱的信息反馈机制, 如单点强盗反馈 (one-point bandit feedback) 和压缩通信. 在单点强盗反馈的情形下, 每个节点仅向其邻居提供某一决策对应的损失值, 而非完整梯度信息. 例如, Yuan 等 [154] 提出了一种基于随机方向导数估计梯度的方法, 并设计了带长期约束的算法, 证明了在强凸和凸条件下, 该方法的静态遗憾分别为 $\mathcal{O}(T^{2/3}\log T)$ 和 $\mathcal{O}(T^{3/4})$. 在信息压缩的情形下, 由于通信容量受到物理条件限制,

每个节点只能传递压缩后的局部信息. 例如, Cao 和 Başar [14] 提出了去中心化在线次梯度算法的压缩扩展版本, 在该方法中, 每个节点仅需要传递与邻居相对参数的符号信息 (即 1 比特压缩策略). 他们指出, 即使只交换符号信息, 仍可达到与完整梯度通信相同的 $\mathcal{O}(\sqrt{T})$ 遗憾上界. 此外, 根据不同的网络结构, 分布式在线学习的研究还可分为无向图 [2,64,145]、平衡/不平衡固定有向图 [53,70,144]、平衡时变/切换有向图 [149,152]、不平衡时变有向图 [54,66,73] 和随机切换图 [67] 等情形. 总体而言, 相较于无向图, 时变图和有向图上的在线学习问题更复杂, 因为网络不对称性和拓扑变化会进一步影响信息传播与遗憾分析.

最后, 分布式在线学习已被广泛应用于多种动态统计与控制问题中. 现有工作包括线性回归问题 [153]、分布式多目标跟踪问题 [104]、分布式目标包围问题 [16] 和在线纳什古诺博弈 [84] 等. 这些研究表明, 分布式在线学习不仅具有完善的理论基础, 也在动态环境下的统计建模、估计与决策问题中展现出广泛的应用潜力.

总而言之, 分布式在线统计学习的核心并不只是将传统在线学习算法直接推广到网络环境, 更重要的是刻画时变数据流、局部信息、通信约束与学习性能之间的关系. 现有研究主要表明, 通过合适的分布式更新规则和通信机制设计, 在线学习方法能够在静态和动态环境下获得次线性遗憾, 并在一定条件下达到最优或接近最优的理论上界.

4.3 分布式非参数统计学习

本小节主要关注分布式非参数统计学习方法. 与参数模型不同, 非参数模型通常不预设有限维参数形式, 而是通过未知函数来刻画变量之间的关系, 因而参数空间往往是无限维的. 具体而言, 考虑回归模型 $Y = g^*(X_i) + \epsilon$, 其中 g^* 为未知但光滑的函数, ϵ 为均值为 0 的随机噪声. 非参数方法的目标是在不对函数形式作强假设的情形下, 对回归函数 g^* 进行有效估计. 从统计角度看, 分布式非参数统计学习所关注的核心问题在于: 在模型复杂度较高、参数空间无限维以及通信和计算资源受限的条件下, 分布式方法能否在降低计算与存储开销的同时保持与集中式方法相当的估计精度, 并进一步刻画节点数量、通信轮次与函数复杂度对统计收敛率的影响.

核岭回归是一类经典的非参数统计方法, 其目标是在再生 Hilbert 空间的框架下估计回归函数 g^*. 然而, 在大规模数据场景下, 传统核方法通常需要计算并存储 $N \times N$ 的核矩阵, 计算复杂度可达 $\mathcal{O}(N^3)$, 因而难以直接应用于海量数据环境. 基于这一计算瓶颈, 研究者开始探索在分布式环境下实现高效核岭回归估计的方法, 同时尽可能保持良好的统计性质.

针对上述挑战, 在中心化架构下, 一类重要思路是基于分治策略的分布式核岭回归方法. 其基本思想是将数据划分到多个节点上, 在每个节点上分别构造局部核岭回归估计, 再由中心节点对各局部估计结果进行聚合, 从而形成全局估计量. 这类方法主要解决的问题是, 如何在显著降低单节点计算复杂度和存储需求的同时, 保持非参数估计的统计效率. 在这一框架下, Zhang 等 [163] 研究了基于一次平均的分布式核岭回归方法, 并证明了在适当条件下该方法能够达到最优收敛速率. 随后, 一系列工作从不同角度进一步拓展了这一框架, 如文献 [106, 139–141] 等. 然而, 已有研究指出, 这类一次通信的分治方法通常要求机器数量的增长速度不能过快, 否则难以保持与集中式方法相同的全局收敛速率. 针对这一问题, Lin 等 [78] 研究了带通信的分布式核岭回归方法, 通过适当的节点间信息交互扩大了机器数量的允许范围, 从而在更一般的机器规模条件下仍能获得良好的统计性质.

相比中心化架构, 去中心化架构会给分布式核岭回归带来额外挑战. 具体而言, 由于各节点的局部目标函数依赖于各自持有的数据, 在不交换原始样本的情形下, 难以直接在去中心化共识框架下对

这些数据依赖的决策变量进行优化. 针对这一问题, Xu 等 [143] 提出了基于随机特征 (random feature, RF) 近似的去中心化核学习方法. 该方法利用随机特征将核函数映射到有限维特征空间, 从而使不同节点能够在具有数据无关参数的随机特征空间中实现共识优化. 理论分析表明, 该方法能够以线性速度收敛到集中式最优解. 在此基础上, Yang 等 [146] 进一步研究了去中心化核岭回归中的数据依赖随机特征问题, 使不同节点能够使用适应本地数据分布的随机特征表示, 从而更好地适应节点间的数据异质性.

总体来看, 分布式非参数统计学习的研究主要围绕如何在降低计算复杂度和通信开销的同时保持非参数估计的统计效率展开. 现有研究表明, 在中心化框架下, 通过分治或适当通信的策略, 分布式核岭回归能够在一定条件下达到与集中式方法相当的最优收敛速率; 在去中心化框架下, 随机特征等近似技术为在局部通信约束下实现非参数学习提供了可行路径. 除了核岭回归之外, 分布式非参数统计学习还包括其他类型的方法, 如分布式局部平均回归 [17]、分布式最近邻方法 [97] 和分布式样条方法 [106] 等. 这些方法从不同角度拓展了非参数学习在分布式环境中的应用范围.

5 未来研究

通过上述对分布式统计学习现有文献分析, 我们发现, 尽管该领域已经取得了一定的进展, 但仍面临许多待解决的挑战. 本节深入探讨这些挑战, 并提出未来的研究方向, 以推动分布式学习的理论发展与实际应用.

在分布式统计学习中, 通信开销问题一直是一个关键挑战. 分布式学习算法通常要求节点之间, 或者节点与中心节点之间, 频繁地交换高维稠密实值向量. 由于这些信息需要通过无线信道传输, 而无线资源 (如能量和带宽) 在实际应用中是有限的, 尤其是在移动设备和传感器网络等资源受限的环境中 (参见文献 [15]). 因此, 如何降低通信开销, 充分利用有限的无线资源, 成为推动分布式学习发展的关键研究方向. 为应对通信开销问题, 未来的研究可从以下几个方向展开探索. (1) 减少通信轮次 [127]. 减少分布式学习算法中节点之间的通信频率, 可以显著降低通信开销. 例如, 设计自适应通信策略并优化通信周期, 仅在模型性能显著提升时节点才交换信息, 这样可以有效减小通信负担并加速算法收敛. (2) 数据压缩技术. 采用高效的数据压缩技术 (如量化 [99]、稀疏化 [1] 和误差补偿压缩 [63] 等) 减少需要传输的数据量. 通过将信息压缩为低比特数或稀疏向量, 显著减小数据传输量, 从而降低通信成本. 未来的研究可以探讨如何设计更加高效的压缩算法, 在压缩比率和性能之间找到合适的平衡.

在可信的分布式统计学习方面, 现有研究仍存在一定的局限. (1) Byzantine 稳健性. 目前, 大多数 Byzantine 稳健的分布式学习研究集中于中心化架构下, 且主要采用一阶优化方法. 未来的研究可以重点关注去中心化架构中的 Byzantine 稳健性问题, 并深入探索二阶优化方法在此领域的应用, 以填补这两方面的空白 (参见文献 [12]). (2) 公平性. 在公平性方面, 当前关于分布式公平学习的研究较多集中于中心化架构下, 且主要针对二分类任务, 对于其他任务 (如回归问题 [55] 和更为复杂的视觉语言模型 [156] 等) 的研究仍较为有限. 因此, 未来研究可以扩展现有的分布式公平学习框架, 适应多样化的学习任务, 并探索如何在去中心化框架下实现公平性, 以应对更复杂和真实的分布式环境.

在分布式半监督和在线统计学习领域, 现有研究也存在局限性. (1) 半监督学习. 大多数研究集中于中心化框架 [111], 且侧重于参数估计. 对于去中心化框架及从统计推断角度 (如估计量的一致性和置信区间构造等) 的系统性分析仍显不足. 未来的研究可以更多地关注去中心化环境下的半监督学习方法, 并探讨如何在此情境下进行统计推断. (2) 在线学习. 非凸损失函数在实际应用中经常遇到, 尽管目

前的研究主要集中于凸情形. 在这方面, 已有文献研究了单机情形下的在线学习, 例如, Lesage-Landry 等[68] 研究了局部 Lipschitz 和非凸损失函数的情形, 并指出动态遗憾为 $\mathcal{O}(1+P_T)$. 然而, 关于非凸情形下的分布式在线学习的研究仍较为稀缺 (参见文献 [59, 85]). 因此, 未来的研究方向之一是进一步探讨非凸情形下的分布式在线学习方法.

致谢 感谢审稿人对本文提出宝贵意见.

参考文献

1 Aji A F, Heafield K. Sparse communication for distributed gradient descent. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017, 440–445
2 Akbari M, Gharesifard B, Linder T. Individual regret bounds for the distributed online alternating direction method of multipliers. IEEE Trans Automat Control, 2018, 64: 1746–1752
3 Alistarh D, Allen-Zhu Z, Li J. Byzantine stochastic gradient descent. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2018, 4618–4628
4 Aybat N S, Wang Z, Lin T, et al. Distributed linearized alternating direction method of multipliers for composite convex consensus optimization. IEEE Trans Automat Control, 2017, 63: 5–20
5 Baeza-Yates R. Bias on the web. Commun ACM, 2018, 61: 54–61
6 Bajović D, Jakovetić D, Krejić N, et al. Newton-like method with diagonal correction for distributed optimization. SIAM J Optim, 2017, 27: 1171–1203
7 Battey H, Fan J, Liu H, et al. Distributed testing and estimation under sparse high dimensional models. Ann Statist, 2018, 46: 1352–1382
8 Bellet A, Guerraoui R, Taziki M, et al. Personalized and private peer-to-peer machine learning. In: International Conference on Artificial Intelligence and Statistics. Stockholm: PMLR, 2018, 473–481
9 Belloni A, Chernozhukov V. ℓ_1-penalized quantile regression in high-dimensional sparse models. Ann Statist, 2011, 39: 82–130
10 Blanchard P, El Mhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2017, 118–128
11 Blyth C R. On Simpson's paradox and the sure-thing principle. J Amer Statist Assoc, 1972, 67: 364–366
12 Bouhata D, Moumen H, Mazari J A, et al. Byzantine fault tolerance in distributed machine learning: A survey. J Exp Theor Artif Intell, 2025: 37, 1331–1389
13 Boyd S, Ghosh A, Prabhakar B, et al. Randomized gossip algorithms. IEEE Trans Inform Theory, 2006, 52: 2508–2530
14 Cao X, Başar T. Decentralized online convex optimization based on signs of relative states. Automatica, 2021, 129: 109676
15 Cao X, Başar T, Diggavi S, et al. Communication-efficient distributed learning: An overview. IEEE J Sel Areas Commun, 2023, 41: 851–873
16 Carnevale G, Camisa A, Notarstefano G. Distributed online aggregative optimization for dynamic multirobot coordination. IEEE Trans Automat Control, 2022, 68: 3736–3743
17 Chang X, Lin S B, Wang Y. Divide and conquer local average regression. Electron J Stat, 2017, 11: 1326–1350
18 Chang X, Lin S B, Zhou D X. Distributed semi-supervised learning with kernel ridge regression. J Mach Learn Res, 2017, 18: 1–22
19 Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization. J Mach Learn Res, 2011, 12: 1069–1109
20 Chen C, Gu Y, Zou H, et al. Distributed sparse composite quantile regression in ultrahigh dimensions. Stat Sinica, 2023, 33: 1143–1167
21 Chen S X, Peng L. Distributed statistical inference for massive data. Ann Statist, 2021, 49: 2851–2869
22 Chen X, Liu W, Mao X. Robust reduced rank regression in a distributed setting. Sci China Math, 2022, 65: 1707–1730
23 Chen X, Liu W, Mao X, et al. Distributed high-dimensional regression under a quantile loss function. J Mach Learn Res, 2020, 21: 1–43
24 Chen X, Liu W, Zhang Y. Quantile regression under memory constraint. Ann Statist, 2019, 47: 3244–3273
25 Chen Y, Su L, Xu J. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. In:

Proceedings of the ACM on Measurement and Analysis of Computing Systems, vol. 1. New York: ACM, 2017, 1–25
26 Cheng H P, Yu P, Hu H, et al. Towards decentralized deep learning with differential privacy. In: International Conference on Cloud Computing. Cham: Springer, 2019, 130–145
27 Colin I, Bellet A, Salmon J, et al. Gossip dual averaging for decentralized optimization of pairwise functions. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2016, 1388–1396
28 Cui S, Pan W, Liang J, et al. Addressing algorithmic disparity and performance inconsistency in federated learning. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2021, 26091–26102
29 Diao E, Ding J, Tarokh V. SemiFL: Semi-supervised federated learning for unlabeled clients with alternate training. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2022, 17871–17884
30 Dimakis A G, Kar S, Moura J M F, et al. Gossip algorithms for distributed signal processing. Proc IEEE, 2010, 98: 1847–1864
31 Du J, Wang Q. Penalized empirical likelihood over decentralized networks. J R Stat Soc Ser B Stat Methodol, 2026, in press
32 Du W, Xu D, Wu X, et al. Fairness-aware agnostic federated learning. In: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). Alexandria: SIAM, 2021, 181–189
33 Duan R, Ning Y, Chen Y. Heterogeneity-aware and communication-efficient distributed statistical inference. Biometrika, 2022, 109: 67–83
34 Duchi J C, Agarwal A, Wainwright M J. Dual averaging for distributed optimization: Convergence analysis and network scaling. IEEE Trans Automat Control, 2011, 57: 592–606
35 Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography Conference. New York: Springer, 2006, 265–284
36 Eisen M, Mokhtari A, Ribeiro A. Decentralized quasi-Newton methods. IEEE Trans Signal Process, 2017, 65: 2613–2628
37 El-Mhamdi E M, Farhadkhani S, Guerraoui R, et al. Collaborative learning in the jungle (decentralized, Byzantine, heterogeneous, asynchronous and nonconvex learning). In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2021, 25044–25057
38 Eshraghi N, Liang B. Improving dynamic regret in distributed online mirror descent using primal and dual information. In: Learning for Dynamics and Control Conference. Stanford: PMLR, 2022: 637–649
39 Ezzeldin Y H, Yan S, He C, et al. Fairfed: enabling group fairness in federated learning. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence, vol. 37. Palo Alto: AAAI Press, 2023, 7494–7502
40 Fan J, Guo Y, Wang K. Communication-efficient accurate statistical estimation. J Amer Statist Assoc, 2023, 118: 1000–1010
41 Fan J, Wang D, Wang K, et al. Distributed estimation of principal eigenspaces. Ann Statist, 2019, 47: 3009–3031
42 Fang C, Yang Z, Bajwa W U. BRIDGE: Byzantine-resilient decentralized gradient descent. IEEE Trans Signal Inf Process Netw, 2022, 8: 610–626
43 Fang M, Zhang Z, Hairi Khanduri P, et al. Byzantine-robust decentralized federated learning. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City: ACM, 2024, 2874–2888
44 Fierimonte R, Scardapane S, Uncini A, et al. Fully decentralized semi-supervised learning via privacy-preserving matrix completion. IEEE Trans Neural Netw Learn Syst, 2016, 28: 2699–2711
45 Forero P A, Cano A, Giannakis G B. Consensus-based distributed linear support vector machines. J Mach Learn Res, 2010, 11: 1663–1707
46 Gantz J, Reinsel D. Extracting value from chaos. Technical Report. IDC iView. Http://idcdocserv.com/1142, 2011
47 Gu J, Chen S X. Distributed statistical inference under heterogeneity. J Mach Learn Res, 2023, 24: 1–57
48 Gu J, Chen S X. Statistical inference for decentralized federated learning. Ann Statist, 2024, 52: 2931–2955
49 Guerraoui R, Rouault S. The hidden vulnerability of distributed learning in Byzantium. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2018, 3521–3530
50 Guo S, Zhang T, Yu H, et al. Byzantine-resilient decentralized stochastic gradient descent. IEEE Trans Circuits Syst Video Technol, 2021, 32: 4096–4106
51 He L, Karimireddy S P, Jaggi M. Byzantine-robust decentralized learning via self-centered clipping. arXiv:2202.01545, 2022
52 Hoi S C H, Sahoo D, Lu J, et al. Online learning: A comprehensive survey. Neurocomputing, 2021, 459: 249–289
53 Hosseini S, Chapman A, Mesbahi M. Online distributed optimization via dual averaging. In: Proceedings of the IEEE Conference on Decision and Control. Florence: IEEE, 2013, 1484–1489

54 Hosseini S, Chapman A, Mesbahi M. Online distributed convex optimization on dynamic networks. IEEE Trans Automat Control, 2016, 61: 3545–3550

55 Hu S, Wu Z S, Smith V. Fair federated learning via bounded group loss. In: Proceedings of the IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). Toronto: IEEE, 2024, 140–160

56 Huang C, Huo X. A distributed one-step estimator. Math Program, 2019, 174: 41–76

57 Jayaraman B, Wang L, Evans D, et al. Distributed learning without distress: Privacy-preserving empirical risk minimization. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2018, 6343–6354

58 Jeong W, Yoon J, Yang E, et al. Federated semi-supervised learning with inter-client consistency and disjoint learning. In: Proceedings of the International Conference on Learning Representations. Virtual Event: OpenReview.net, 2021, 1–15

59 Jiang Z, Balu A, Lee X Y, et al. Distributed online non-convex optimization with composite regret. In: Proceedings of the 58th Annual Allerton Conference on Communication, Control, and Computing. Monticello: IEEE, 2022, 1–8

60 Jordan M I, Lee J D, Yang Y. Communication-efficient distributed statistical inference. J Amer Statist Assoc, 2019, 114: 668–681

61 Kempe D, Dobra A, Gehrke J. Gossip-based computation of aggregate information. In: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science. Cambridge: IEEE, 2003, 482–491

62 Koenker R, Bassett G Jr. Regression quantiles. Econometrica, 1978, 46: 33–50

63 Koloskova A, Stich S, Jaggi M. Decentralized stochastic optimization and gossip algorithms with compressed communication. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2019, 3478–3487

64 Koppel A, Jakubiec F Y, Ribeiro A. A saddle point algorithm for networked online convex optimization. IEEE Trans Signal Process, 2015, 63: 5149–5164

65 Lamport L, Shostak R, Pease M. The Byzantine generals problem. In: Concurrency: The works of Leslie Lamport. New York: ACM, 2019, 203–226

66 Lee S, Nedić A, Raginsky M. Stochastic dual averaging for decentralized online optimization on time-varying communication graphs. IEEE Trans Automat Control, 2017, 62: 6407–6414

67 Lei J, Yi P, Hong Y, et al. Online convex optimization over Erdos-Rényi random networks. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2020, 15591–15601

68 Lesage-Landry A, Taylor J A, Shames I. Second-order online nonconvex optimization. IEEE Trans Automat Control, 2021, 66: 4866–4872

69 Li B, Chi Y. Convergence and privacy of decentralized nonconvex optimization with gradient clipping and communication compression. IEEE J Sel Top Signal Process, 2025, 19: 273–282

70 Li J, Li C, Yu W, et al. Distributed online bandit learning in dynamic environments over unbalanced digraphs. IEEE Trans Netw Sci Eng, 2021, 8: 3034–3047

71 Li M, Li Q, Wang Y. Class balanced adaptive pseudo labeling for federated semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023, 16292–16301

72 Li X, Xie L, Li N. A survey on distributed online optimization and online games. Annu Rev Control, 2023, 56: 100904

73 Li X, Yi X, Xie L. Distributed online optimization for multi-agent networks with coupled inequality constraints. IEEE Trans Automat Control, 2020, 66: 3575–3591

74 Li Y, Zhu J. L_1-norm quantile regression. J Comput Graph Statist, 2008, 17: 163–185

75 Lian X, Zhang C, Zhang H, et al. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2017, 30–40

76 Lian X, Zhang W, Zhang C, et al. Asynchronous decentralized parallel stochastic gradient descent. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2018, 3043–3052

77 Liang X, Lin Y, Fu H, et al. Rscfed: Random sampling consensus federated semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022, 10154–10163

78 Lin S B, Wang D, Zhou D X. Distributed kernel ridge regression with communications. J Mach Learn Res, 2020, 21: 1–38

79 Ling Q, Wen Z, Yin W. Decentralized jointly sparse optimization by reweighted ℓ_q minimization. IEEE Trans Signal Process, 2010, 61: 1165–1170

80 Ling Q, Xu Y, Yin W, et al. Decentralized low-rank matrix completion. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto: IEEE, 2012, 2925–2928

81 Liu Q, Ihler A T. Distributed estimation, information loss and exponential families. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2014, 1098–1106

82 Liu W, Mao X, Zhang X. Fast and robust sparsity learning over networks: A decentralized surrogate median regression approach. IEEE Trans Signal Process, 2022, 70: 797–809

83 Liu W, Tu J, Mao X, et al. Majority vote for distributed differentially private sign selection. Ann Statist, 2024, 52: 1671–1690

84 Lu K, Li G, Wang L. Online distributed algorithms for seeking generalized Nash equilibria in dynamic environments. IEEE Trans Automat Control, 2020, 66: 2289–2296

85 Lu K, Wang L. Online distributed optimization with nonconvex objective functions: Sublinearity of first-order optimality condition-based regret. IEEE Trans Automat Control, 2021, 67: 3029–3035

86 Ma C, Li J, Ding M, et al. On safeguarding privacy and security in the framework of federated learning. IEEE Netw, 2020, 34: 242–248

87 Mateos G, Bazerque J A, Giannakis G B. Distributed sparse linear regression. IEEE Trans Signal Process, 2010, 58: 5262–5276

88 Mateos-Núñez D, Cortés J. Distributed online convex optimization over jointly connected digraphs. IEEE Trans Netw Sci Eng, 2014, 1: 23–37

89 Minsker S. Distributed statistical estimation and rates of convergence in normal approximation. Electron J Stat, 2019, 13: 5213–5252

90 Mirzaeifard R, Gogineni V C, Venkategowda N K D, et al. Distributed quantile regression with non-convex sparse penalties. In: Proceedings of the IEEE Statistical Signal Processing Workshop (SSP). Hanoi: IEEE, 2023, 250–254

91 Mokhtari A, Shi W, Ling Q, et al. A decentralized second-order method with exact linear convergence rate for consensus optimization. IEEE Trans Signal Inf Process Networks, 2016, 2: 507–522

92 Nedic A, Olshevsky A, Rabbat M G. Network topology and communication-computation tradeoffs in decentralized optimization. Proc IEEE, 2018, 106: 953–976

93 Nedic A, Ozdaglar A. Distributed subgradient methods for multi-agent optimization. IEEE Trans Automat Control, 2009, 54: 48–61

94 Ozdayi M S, Kantarcioglu M, Gel Y R. Defending against backdoors in federated learning with robust learning rate. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, vol. 35. Palo Alto: AAAI Press, 2021, 9268–9276

95 Pathak M, Rane S, Raj B. Multiparty differential privacy via aggregation of locally trained classifiers. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook, Curran Associates, 2010, 1876–1884

96 Predd J B, Kulkarni S R, Poor H V. A collaborative training algorithm for distributed learning. IEEE Trans Inform Theory, 2009, 55: 1856–1871

97 Qiao X, Duan J, Cheng G. Rates of convergence for large-scale nearest neighbor classification. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2019, 10769–10780

98 Qu G, Li N. Harnessing smoothness to accelerate distributed optimization. IEEE Trans Control Netw Syst, 2017, 5: 1245–1260

99 Rabbat M G, Nowak R D. Quantized incremental algorithms for distributed optimization. IEEE J Sel Areas Commun, 2005, 23: 798–808

100 Richards D, Rebeschini P. Optimal statistical rates for decentralised non-parametric regression with linear speed-up. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2019, 1216–1227

101 Richards D, Rebeschini P, Rosasco L. Decentralised learning with distributed gradient descent and random features. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2020, 1–11

102 Rosenblatt J D, Nadler B. On the optimality of averaging in distributed statistical learning. Inf Inference, 2016, 5: 379–404

103 Scardapane S, Fierimonte R, Di Lorenzo P, et al. Distributed semi-supervised support vector machines. Neural Networks, 2016, 80: 43–52

104 Shahrampour S, Jadbabaie A. Distributed online optimization in dynamic environments using mirror descent. IEEE Trans Automat Control, 2017, 63: 714–725

105 Shamir O, Srebro N, Zhang T. Communication-efficient distributed optimization using an approximate Newton-type method. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2014, 1000–1008

106 Shang Z, Cheng G. Computational limits of a distributed algorithm for smoothing spline. J Mach Learn Res, 2017, 18: 1–37

107 Shi J, Wang Y, Zhu Z, et al. Decentralized learning of quantile regression: A smoothing approach. J Comput Graph Statist, 2025, 34: 1091–1101

108 Shi W, Ling Q, Wu G, et al. Extra: An exact first-order algorithm for decentralized consensus optimization. SIAM

J Optim, 2015, 25: 944–966
109 Shi W, Ling Q, Yuan K, et al. On the linear convergence of the ADMM in decentralized consensus optimization. IEEE Trans Signal Process, 2014, 62: 1750–1761
110 Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models. In: Proceedings of the IEEE Symposium on Security and Privacy. San Jose: IEEE, 2017, 3–18
111 Song Z, Yang X, Zhang Y, et al. A systematic survey on federated semi-supervised learning. In: Proceedings of the International Joint Conference on Artificial Intelligence. Jeju: IJCAI Organization, 2024, 8244–8252
112 Su C, Yu G, Wang J, et al. Multi-dimensional fair federated learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence, vol. 38. Palo Alto: AAAI Press, 2024, 15083–15090
113 Su L, Xu J. Securing distributed gradient descent in high dimensional statistical learning. Proc ACM Meas Anal Comput Syst, 2019, 3: 1–41
114 Tan K M, Battey H, Zhou W X. Communication-constrained distributed quantile regression with optimal statistical guarantees. J Mach Learn Res, 2022, 23: 1–61
115 Tang H, Lian X, Yan M, et al. D^2: Decentralized training over decentralized data. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2018, 4848–4856
116 Truex S, Liu L, Chow K H, et al. LDP-Fed: Federated learning with local differential privacy. In: Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking. Heraklion: ACM, 2020: 61–66
117 Tu J, Liu W, Mao X. Byzantine-robust distributed sparse learning for M-estimation. Mach Learn, 2023, 112: 3773–3804
118 Tu J, Liu W, Mao X. Distributed estimation on semi-supervised generalized linear model. J Mach Learn Res, 2024, 25: 1–41
119 Tu J, Liu W, Mao X, et al. Variance reduced median-of-means estimator for Byzantine-robust distributed inference. J Mach Learn Res, 2021, 22: 1–67
120 Tu J, Liu W, Mao X, et al. Distributed semi-supervised sparse statistical inference. IEEE Trans Inform Theory, 2023, 70: 4197–4217
121 Volgushev S, Chao S K, Cheng G. Distributed inference for quantile regression processes. Ann Statist, 2019, 47: 1634–1662
122 Wang C, Shen Z. Distributed high-dimensional quantile regression: Estimation efficiency and support recovery. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2024, 51415–51441
123 Wang G, Payani A, Lee M, et al. Mitigating group bias in federated learning: Beyond local fairness. Trans Mach Learn Res, 2024: 2835–8856
124 Wang H, Li C. Distributed quantile regression over sensor networks. IEEE Trans Signal Inf Process Networks, 2017, 4: 338–348
125 Wang J, Kolar M, Srebro N, et al. Efficient distributed learning with sparsity. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2017, 3636–3645
126 Wang L, Wu Y, Li R. Quantile regression for analyzing heterogeneity in ultra-high dimension. J Amer Statist Assoc, 2012, 107: 214–222
127 Wang S, Tuor T, Salonidis T, et al. Adaptive federated learning in resource constrained edge computing systems. IEEE J Sel Areas Commun, 2019, 37: 1205–1221
128 Wang X, He J, Cheng P, et al. Privacy preserving collaborative computing: Heterogeneous privacy guarantee and efficient incentive mechanism. IEEE Trans Signal Process, 2018, 67: 221–233
129 Wang Y, Nedić A. Tailoring gradient methods for differentially private distributed optimization. IEEE Trans Automat Control, 2023, 69: 872–887
130 Wang Z, Liu J, Shao J, et al. Distributed semi-supervised inference for generalized linear models with block-wise missing covariates. IEEE Trans Inf Theory, 2025, 71, 7815–7841
131 Wei K, Li J, Ding M, et al. Federated learning with differential privacy: Algorithms and performance analysis. IEEE Trans Inform Forensic Secur, 2020, 15: 3454–3469
132 Wei W, Liu L. Trustworthy distributed AI systems: Robustness, privacy, and governance. ACM Comput Surv, 2025, 57: 1–42
133 Wu S, Huang D, Wang H. Network gradient descent algorithm for decentralized federated learning. J Bus Econom Statist, 2023, 41: 806–818
134 Wu T, Yuan K, Ling Q, et al. Decentralized consensus optimization with asynchrony and delays. IEEE Trans Signal Inf Process Networks, 2017, 4: 293–307
135 Xiao L, Boyd S. Fast linear iterations for distributed averaging. Systems Control Lett, 2004, 53: 65–78
136 Xiao L, Boyd S, Kim S J. Distributed average consensus with least-mean-square deviation. J Parallel Distributed

Comput, 2007, 67: 33–46
137 Xie J, Liu S Y, Chen J X. A framework for distributed semi-supervised learning using single-layer feedforward networks. Mach Intell Res, 2022, 19: 63–74
138 Xie J, Liu S Y, Dai H. A distributed semi-supervised learning algorithm based on manifold regularization using wavelet neural network. Neural Networks, 2019, 118: 300–309
139 Xu C, Zhang Y, Li R, et al. On the feasibility of distributed kernel regression for big data. IEEE Trans Knowl Data Eng, 2016, 28: 3041–3052
140 Xu G, Shang Z, Cheng G. Optimal tuning for divide-and-conquer kernel ridge regression with massive data. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2018, 5483–5491
141 Xu G, Shang Z, Cheng G. Distributed generalized cross-validation for divide-and-conquer kernel ridge regression and its asymptotic optimality. J Comput Graph Statist, 2019, 28: 891–908
142 Xu J, Zhang W, Wang F. A $(DP)^2$2 SGD: Asynchronous decentralized parallel stochastic gradient descent with differential privacy. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 8036–8047
143 Xu P, Wang Y, Chen X, et al. COKE: Communication-censored decentralized kernel learning. J Mach Learn Res, 2021, 22: 1–35
144 Yamashita M, Hayashi N, Hatanaka T, et al. Logarithmic regret for distributed online subgradient method over unbalanced directed networks. IEICE Trans Fund, 2021, 104: 1019–1026
145 Yan F, Sundaram S, Vishwanathan S V N, et al. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. IEEE Trans Knowl Data Eng, 2012, 25: 2483–2493
146 Yang R, He F, He M, et al. Decentralized kernel ridge regression based on data-dependent random feature. IEEE Trans Neural Netw Learn Syst, 2024, 36: 7945–7954
147 Yang X, Huang W, Ye M. Dynamic personalized federated learning with adaptive differential privacy. In: Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2023, 72181–72192
148 Yang Z, Bajwa W U. ByRDiE: Byzantine-resilient distributed coordinate descent for decentralized learning. IEEE Trans Signal Inf Process Networks, 2019, 5: 611–627
149 Yi X, Li X, Xie L, et al. Distributed online convex optimization with time-varying coupled inequality constraints. IEEE Trans Signal Process, 2020, 68: 731–746
150 Yin D, Chen Y, Kannan R, et al. Byzantine-robust distributed learning: Towards optimal statistical rates. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2018, 5650–5659
151 Yu D, Zou Z, Chen S, et al. Decentralized parallel SGD with privacy preservation in vehicular networks. IEEE Trans Veh Technol, 2021, 70: 5211–5220
152 Yuan D, Hong Y, Ho D W C, et al. Distributed mirror descent for online composite optimization. IEEE Trans Automat Control, 2020, 66: 714–729
153 Yuan D, Proutiere A, Shi G. Distributed online linear regressions. IEEE Trans Inform Theory, 2020, 67: 616–639
154 Yuan D, Proutiere A, Shi G. Distributed online optimization with long-term constraints. IEEE Trans Automat Control, 2021, 67: 1089–1104
155 Yuan K, Ling Q, Yin W. On the convergence of decentralized gradient descent. SIAM J Optim, 2016, 26: 1835–1854
156 Zeng H, Yue Z, Zhang Y, et al. Fair federated learning with biased vision-language models. Find Assoc Comput Linguist, 2024: 10002–10017
157 Zhang C, Ahmad M, Wang Y. ADMM based privacy-preserving decentralized optimization. IEEE Trans Inform Forensic Secur, 2018, 14: 565–580
158 Zhang J, Ma X, Guo S, et al. Towards unbiased training in federated open-world semi-supervised learning. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2023, 41498–41509
159 Zhang R, Kwok J. Asynchronous distributed ADMM for consensus optimization. In: Proceedings of the International Conference on Machine Learning. New York: PMLR, 2014, 1701–1709
160 Zhang T, Zhu Q. Dynamic differential privacy for ADMM-based distributed classification learning. IEEE Trans Inform Forensic Secur, 2016, 12: 172–187
161 Zhang X, Liu J, Zhu Z. Learning coefficient heterogeneity over networks: A distributed spanning-tree-based fused-lasso regression. J Amer Statist Assoc, 2024, 119: 485–497
162 Zhang Y, Duchi J C, Wainwright M J. Communication-efficient algorithms for statistical optimization. J Mach Learn Res, 2013, 14: 3321–3363
163 Zhang Y, Duchi J C, Wainwright M J. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. J Mach Learn Res, 2015, 16: 3299–3340
164 Zhang Z, Yang Y, Yao Z, et al. Improving semi-supervised federated learning by reducing the gradient diversity of models. In: Proceedings of the IEEE International Conference on Big Data (Big Data). Orlando: IEEE, 2021,

1214–1225
165 Zheng Q, Peng L, He X. High dimensional censored quantile regression. Ann Statist, 2018, 46: 308–343
166 Zhou L, She X, Song P X K. Distributed empirical likelihood approach to integrating unbalanced datasets. Stat Sinica, 2023, 33: 2209–2231
167 Zhu Z, Huang Y, Wang X, et al. Dyn-D^{2}P: Dynamic differentially private decentralized learning with provable utility guarantee. In: Proceedings of the International Joint Conference on Artificial Intelligence. Montreal: IJCAI Organization, 2025, 7272–7281
168 Zinkevich M. Online convex programming and generalized infinitesimal gradient ascent. In: Proceedings of the 20th International Conference on Machine Learning. Washington: AAAI Press, 2003, 928–936

Recent advances in distributed statistical learning

Weidong Liu, Xiaojun Mao & Zhengrong Yu

Abstract With the rapid growth of massive datasets, traditional statistical learning methods face significant challenges in processing large-scale data, driving the emergence of distributed learning as a key research area. However, as technology evolves, the need for trustworthiness in distributed learning has become progressively more critical, particularly concerning robustness, privacy protection, and fairness. At the same time, the scope of application for distributed learning continues to expand, encompassing emerging fields such as semi-supervised learning and online learning. This paper reviews the core frameworks and optimization algorithms of distributed learning, with a particular focus on recent advances under different trustworthiness requirements and application circumstances.

Keywords **distributed algorithms, robustness, privacy protection, fairness, semi-supervised learning, online learning**

MSC(2020) **68W15, 68T05, 62F35, 68P27, 68W27**

doi: 10.1360/SSM-2025-0292