面向非凸模型与异构数据的去中心化随机梯度下降统一基于动量的范式

Haizhou Du[1] , **Chendong Ni**[2] ,[1]
上海电力大学

arXiv:2303.00179v1 [cs.LG] 1 Mar 2023

摘要

近年来，新兴的分布式应用推动了去中心化机器学习的发展，尤其是在物联网和边缘计算领域。在现实场景中，非凸性和数据异质性这两个常见问题会导致效率低下、性能下降和发展停滞。大量研究集中于上述问题之一，尚未形成一个经过证明具有最优性的更通用框架。为此，我们提出了一种名为 UMP 的统一范式，该范式包含两种基于动量技术和去中心化随机梯度下降（SGD）的算法：D-SUM 和 GT-DSUM。前者为一般非凸目标函数提供收敛保证；后者通过引入梯度跟踪进行扩展，用于估计全局优化方向，从而缓解数据异质性（即分布漂移）。在 UMP 中，通过设置不同参数，可以涵盖大多数基于经典重球法或 Nesterov 加速的基于动量的变体。在理论方面，我们严格给出了这两种方法针对非凸目标函数的收敛性分析，并开展了广泛的实验。实验结果表明，在实际应用中，与其他方法相比，模型准确率最高提升了 57.6%。

1 引言

分布式机器学习（DML）已成为大规模机器学习中的重要范式 [Wan 等，2022；Zhang 等，2022；Qu等，2022]。根据在工作节点之间聚合模型参数/梯度的方式，研究人员将系统架构分为两大类：参数服务器（PS）和去中心化架构。前者通常被视为集中式范式，其中中央服务器作为协调者以提供便利；后者则允许通过底层拓扑以点对点方式进行通信，从而能够保证所有工作节点之间的模型一致性，并具备更好的可扩展性。

与此同时，多项互补性研究 [Fang 等人, 2018；Yu 等人, 2019；Hsieh 等人，2020] 主要从以下两个关键方面关注了DML的问题。

•非凸目标函数的性质在深度学习中相当复杂，尤其是在分布式场景中 *[Karim-ireddy* 等人，*2020*；*Lian* 等人， *2017]*。尽管针对凸模型已获得一些标准理论结果 [Tao 等人，2022；Deng 和 Gao，2021；Tao 等人，2021],但由于它们可能存在损失并导致严重障碍（例如，计算复杂度高和泛化能力差），在非凸设置中可适用的结果要少得多 [Ghadimi 等人, 2015年；Mai 和 Johansson， 2020]。 • 众所周知，数据异质性是分布式训练中的关键挑战之一，会导致收敛缓慢且不稳定，以及较差的模型泛化能力。令人失望的经验性能与数据异质性程度之间仍然存在差距 [Shang 等人, 2022；Lin 等人, 2021；Esfandiari等人，2021]。遗憾的是，目前尚无研究尝试从综合视角改进现实世界中的去中心化训练，同时将非凸性和数据异质性纳入考量。因此，应对这些挑战并非易事，它们严重阻碍了现实应用的发展。

受动量对最优收敛复杂度的影响及其在实证评估中的成功启发[Koloskova等，2019；Yu 等，2019；Han 和 Gao，2021；Lin 等， 2021], 我们提出了**UMP**，即一种统一的、基于**U**nified、**M**omentum 的**P**aradigm，在全文中不考虑通信开销的去中心化学习范式。该范式包含两种算法，分别称为 D-SUM 和 GT-DSUM。前者 D-SUM 通过维护并缩放动量缓冲区，显著锐化损失景观，探索动量的潜力，并克服非凸性的限制，从而在非凸设置下获得更好的模型性能和更快的收敛速度。后者算法 GT-DSUM 还通过引入梯度跟踪（GT）技术 [Di Lorenzo 和 Scutari, 2016]，旨在减轻数据异质性对局部模型参数差异的影响。其核心洞见是：工作节点之间的方差不断减小，同时本地梯度渐近地与全局优化方向保持一致，而与数据的异质性无关。GT-DSUM 加速了去中心化学习，使其在非凸和不同程度的非独立同分布条件下均能实现更好的泛化性能。

本文的**主要贡献**如下：

• 我们提出了一种统一的基于动量的范式 UMP，其中包含两种算法，用于同时处理非凸问题和

非独立同分布程度。此外，通过指定我们基础算法的参数，还可以得到各种采用动量技术的算法。

- 我们设计了首个算法 D-SUM，其模型性能良好，证明了该算法在效果和效率方面的适用性。我们还给出了其在非凸情形下的收敛结果。
- 我们正在进一步开发第二种算法 GT-DSUM。该算法应用 GT 技术，能够应对分布漂移问题。我们严格证明了其在平滑非凸设置下的收敛界。
- 此外，我们还开展了大量实验，在常用模型、数据集和动态真实场景中评估 UMP 的性能。实验结果表明，在不同的非独立同分布程度下，与知名的去中心化基线方法相比，D-SUM 和 GT-DSUM 分别将模型准确率最多提高了 35.8% 和 57.6%。在受数据偏斜影响的各项训练任务中，GT-DSUM 的模型泛化优于 D-SUM。

2 统一范式：UMP

本节首先介绍符号，并回顾两种动量方法：重球（HB）方法 [Polyak,1964] 和 Nesterov动量 [Nesterov，1983]。受这两种方法启发，我们提出一种统一的基于动量的范式，并给出两种算法 D-SUM 和 GT-DSUM。该范式既涵盖上述两种经典方法，也涵盖其他基于动量的变体，旨在解决现实去中心化学习应用中的非凸性和数据异质性问题。最后，我们给出收敛结果，表明这些算法对于一般的光滑非凸目标函数几乎都能收敛到一个稳定点。

2.1 符号与预备知识

为了更好地展示其在现实复杂场景中的适用效果，我们考虑一种具有网络拓扑结构的去中心化设置，其中 n 个工作节点共同处理一个优化问题。假设每个工作节点 i 都持有从 $\mathcal{D}_i$ 分布中抽取的数据集，这体现了数据异质性。令 $f_i: \mathbb{R}^d \to \mathbb{R}$ 为工作节点 i 的训练数据集损失函数，其随机形式可写为$\mathbb{E}_{\xi_i\sim\mathcal{D}_i}[\nabla F_i(\mathbf{x}\ \xi_i)] = \nabla f_i(\mathbf{x})$，其中 $F_i(\mathbf{x}\ \xi_i)$ 是与小批量样本 $\xi_i \sim \mathcal{D}_i$ 相关的逐数据损失函数。随后，我们将经验风险最小化问题表述为具有求和结构目标函数的形式：

$$f^* = \min_{\mathbf{x}\in\mathbb{R}^d}\left[f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n}[f_i(\mathbf{x}) = \mathbb{E}_{\xi_i\sim\mathcal{D}_i}F_i(\mathbf{x},\xi_i)]\right]. \tag{1}$$

在工作节点之间，存在一种底层拓扑结构$\mathbf{W}$ h $\in \mathbb{R}^{n\times n}$，它便于描述任意两个工作节点之间的通信；即，当且仅当工作节点 i 与 j 不相连时，我们令 $w_{ij} = 0$。

定义 1 (共识矩阵 [Koloskova 等人， 2021])。一个元素均为非负数的矩阵$\mathbf{W} \in [0,1]^{n\times n}$，满足对称（$\mathbf{W} = \mathbf{W}^\top$），且为双随机矩阵（$\mathbf{W}\mathbf{1} = \mathbf{1}$，$\mathbf{1}^\top\mathbf{W} = \mathbf{1}^\top$），其中表示中的全一向量$\mathbf{1} = \mathbb{R}^n$。

全文中，我们用记号$\mathbf{x}_{\cdot,i}^{(t)\ \tau}$表示工作节点 i 在第 t轮次的第 τ次本地更新时的模型参数序列。对于任意向量$\mathbf{a}_i \in \mathbb{R}^d$，我们将其模型平均记为 $\bar{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{a}_i$。令 $\|\cdot\|$ 和 $\|\cdot\|_F$分别表示 l_2向量范数和弗罗贝尼乌斯矩阵范数。

为便于表述，我们在更合适时采用向量和矩阵两种记法。我们用大写字母表示由$\mathbf{a}_i$ 组合得到的矩阵形式，如下所示，

$$\mathbf{A} = [\mathbf{a}_1,\cdots,\mathbf{a}_n] \in \mathbb{R}^{d\times n},\quad \bar{\mathbf{A}} = [\bar{\mathbf{a}},\cdots,\bar{\mathbf{a}}] = \mathbf{A}\tfrac{1}{n}\mathbf{1}\mathbf{1}^\top. \tag{2}$$

引入动量 项是最常见的改进方法之一，它被视为训练最先进深度神经网络的关键组成部分 [Qu等，2022；Lin等， 2021]。与其经验上的成功相对应，动量通过将随机梯度与历史方向相结合来设定优化后的搜索方向，从而试图提升非凸目标函数的收敛速度。

HB 方法（即，也称为 Polyak 动量）最初是针对光滑凸情形提出的，其形式如下：

$$\begin{cases}\mathbf{u}_i^{(t+1)} = \beta\mathbf{u}_i^{(t)} + \mathbf{g}_i^{(t)}\\ \mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} - \eta\mathbf{u}_i^{(t+1)},\end{cases} \tag{3}$$

其中，$\mathbf{u}_i^{(t)}$ 和 $\mathbf{g}_i^{(t)}$ 分别表示动量缓冲区和工作节点 i 在第 t 轮次的随机梯度。η 表示学习率。动量变量 β 会根据过去的信息估计与随机梯度，调整更新方向的幅度，从而指示最速下降的方向。等价地，（3）也可以按以下形式更新：

$$\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} - \eta\mathbf{g}_i^{(t)} + \beta\left(\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t-1)}\right), \tag{4}$$

当 $t \geq 1$时。保留过去的梯度值，这种更新方式在一定程度上具有更好的稳定性，并且相较于一些原始随机梯度下降方法 [CutkoskyandMehta 2020]，能够实现改进。

另一种称为 Nesterov 的技术表明，选择合适的参数后，外推步骤可以从 $\mathcal{O}\left(\frac{1}{t}\right)$ 加速至 $\mathcal{O}\left(\frac{1}{t^2}\right)$，这对于光滑凸问题而言是最优速率。具体来说，其更新步骤如下：

$$\begin{cases}\mathbf{u}_i^{(t+1)} = \beta\mathbf{u}_i^{(t)} + \mathbf{g}_i^{(t)}\\ \mathbf{v}_i^{(t+1)} = \beta\mathbf{u}_i^{(t+1)} + \mathbf{g}_i^{(t)}\\ \mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} - \eta\mathbf{v}_i^{(t+1)}.\end{cases} \tag{5}$$

通过引入动量向量 $\mathbf{u}_i$ 和额外的辅助$\mathbf{v}_i$ 序列来更新模型参数。与 (3) 相比，通过衰减动量缓冲区 $\mathbf{u}_i^{(t)}$，该方法能够有效提高收敛速度，同时不会引发振荡。类似地，上述步骤可以写为

$$\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} - \eta\mathbf{g}_i^{(t)} + \beta\left(\mathbf{x}_i^{(t)} - \eta\mathbf{g}_i^{(t)} - \mathbf{x}_i^{(t-1)} + \eta\mathbf{g}_i^{(t-1)}\right). \tag{6}$$

算法1： 原始随机 JM；彦 中备选变体指示这两种替代变体。

输入： $\forall i$，初始化$\mathbf{x}_i^{(0)\,0} = \cdot\ \mathbf{v}_i^{(0)\,0}\ \cdot = \mathbf{x}_0$；常数参数 η、α 和 β；$\forall i,\ j$，共识矩阵 $\mathbf{W}$ 其元素为 w_{ij}；训练轮次为 T 和本地步数 K。

对于 $t \in \{0, \cdots, T-1\}$，**在工作节点** i **并行执行**

  设置 $\mathbf{x}_{i\cdot}^{(t)\cdot\,0} = \mathbf{x}_i^{(t)}, \mathbf{v}_{i\cdot}^{(t)\cdot\,0} = \mathbf{v}_i^{(t)}$.

  对于 $\tau \in \{0, \cdots, K-1\}$，**执行**

    样本 $\xi_{\cdot i}^{(t)\,\tau}$ 并计算 $\mathbf{g}_i^{(t)\,\tau} = \nabla F_i(\mathbf{x}_{\cdot i}^{(t)\,\tau}, \xi_{\cdot i}^{(t)\,\tau})$。

    $\mathbf{x}_{\cdot i}^{(t)\,\tau+1} = \mathbf{x}_{\cdot i}^{(t)\,\tau} - \eta \mathbf{g}_{\cdot i}^{(t)\,\tau}$.

    根据式（7）计算本地模型$\mathbf{x}_{\cdot i}^{(t)\,\tau}$。

  end

  通过式（8）进行八卦平均。

  $\mathbf{v}_i^{(t+1)} = \sum_{j=1}^n w_{ij} \mathbf{v}_j^{(t)\,K}$.

结束

根据 (4) 和 (6)，不难看出，前者可以同时计算梯度并添加动量，而后者则是在计算梯度后再应用动量，直观上会带来更高的计算成本。同时，借鉴 HB 动量的思想，Nesterov 加速法通过引入额外的梯度下降规则，将被减去的梯度 $\eta(\mathbf{g}_i^{(t-1)} - \mathbf{g}_i^{(t)})$ 相加，从而使我们更接近最小值（即，$\mathbf{x}^*$），适用于一般凸情形。上述两种基本的基于动量的方法最初是在凸情形下研究的，显示出其相较于原始随机梯度下降的优势。然而，在常见的现实场景中，对于非凸条件下基于动量的随机梯度下降，仍缺乏全面的分析。

2.2 D-SUM算法

本节介绍 UMP 及其首个算法 D-SUM，该算法用于非凸情形下的去中心化训练。

在每一轮次中，工作节点首先使用不同的优化器（即，随机梯度下降、Adam [Kingma 和 Ba,2015], 等）进行 K 本地更新，可选择使用或不使用动量。本文主要关注基于动量的随机梯度下降变体，例如式 (3) 和式 (5) 中所示的变体。从整体上看，我们采用随机统一动量（SUM）的关键更新，其形式如下

$$\begin{cases} \mathbf{u}_i^{(t),\tau+1} = \mathbf{x}_i^{(t),\tau} - \eta \mathbf{g}_i^{(t),\tau} \\ \mathbf{v}_{i,}^{(t),\tau+1} = \mathbf{x}_i^{(t),\tau} - \alpha\eta \mathbf{g}_i^{(t),\tau} \\ \mathbf{x}_i^{(t),\tau+1} = \mathbf{u}_i^{(t),\tau+1} + \beta \left(\mathbf{v}_i^{(t),\tau+1} - \mathbf{v}_i^{(t),\tau} \right), \end{cases} \tag{7}$$

其中 $\alpha \geq 0$，以及 $\beta \in [0, 1)$。$\mathbf{a}_{\cdot i}^{(t)\,\tau}$（$\mathbf{a}_i$ 可以是$\mathbf{x}_i$、$\mathbf{u}_i$、$\mathbf{v}_i$和$\mathbf{g}_i$的实例）表示工作节点 i 在第 t 轮完成 τ 次本地更新后的相关变量。经过 K 个本地步数后，

算法2： GT-DSUM

输入： $\forall i$，初始化$\mathbf{x}_i^{(0)\,0} = \cdot\ \mathbf{v}_i^{(0)\,0}\ \cdot = \mathbf{x}_0$，$\mathbf{y}_i^{(0)} = \mathbf{g}_i^{(0)\,0} = \nabla F_i(\mathbf{x}_i^{(0)\,0}, \xi_{\cdot i}^{(0)\,0})$，以及 $\mathbf{d}_i^{(-1)} = \mathbf{0}_p$；常数参数 $\alpha \geq 0,\ \beta \in [0, 1)\ \eta,\ \lambda \in [0, 1];\ \forall i,\ j$，共识矩阵 $\mathbf{W}$，其元素为 w_{ij}；训练轮次 T，本地步数 K。

对 $t \in \{0, \cdots, T-1\}$，**在工作节点** i **并行执行**

  对 $\tau \in \{0, \cdots, K-1\}$ **执行**

    采样 $\xi_{\cdot i}^{(t)\,\tau}$，计算 $\mathbf{g}_i^{(t)\,\tau} = \nabla F_i(\mathbf{x}_{\cdot i}^{(t)\,\tau}, \xi_{\cdot i}^{(t)\,\tau})$.

    $\mathbf{m}_{\cdot i}^{(t)\,\tau} = \lambda \mathbf{g}_{\cdot i}^{(t)\,\tau} + (1-\lambda)\ \mathbf{y}_i^{(t)}$。

    替换$\mathbf{g}_i^{(t),\tau}$为$\mathbf{m}_i^{(t),\tau}$作为本地梯度估计，执行 (7)。

  end

  八卦平均$\mathbf{x}_i^{(t+1)} = \sum_{j=1}^n w_{ij} \mathbf{x}_j^{(t)\,K}$。

  $\mathbf{v}_i^{(t+1)} = \sum_{j=1}^n w_{ij} \mathbf{v}_i^{(t)\,K}$。

  $\mathbf{d}_i^{(t)} = \frac{\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t+1)}}{K\eta}$.

  基于梯度跟踪 $\mathbf{y}_i^{(t+1)} = \sum_{j=1}^n w_{ij} \left(\mathbf{y}_j^{(t)} + \mathbf{d}_j^{(t)} - \mathbf{d}_j^{(t-1)} \right)$。

结束

工作节点 i 按照通信模式$\mathbf{W}$ 与其邻居节点通信，以交换各自的局部模型参数。我们将这种同步操作称为八卦平均，可简洁地表示为

$$\mathbf{x}_i^{(t+1)} = \sum_{j=1}^n w_{ij} \mathbf{x}_j^{(t),K}. \tag{8}$$

为展示式（7）中原始随机梯度下降与随机统一动量之间的差异，我们在算法1中总结了训练过程。通过调节超参数 α、β、η 和 K，即可得到具体的算法实例。当分别设置 $\alpha = 0$和 $\alpha = 1$时，我们涵盖了基本的重球法（4）和 Nesterov 动量（6）。此外，当 $K = 1$ 时，该方法退化为采用动量加速的标准小批量随机梯度下降。特别地，为简化理论分析，我们对任意工作节点 i 使用与式（8）相同的 gossip同步来更新辅助变量序列 $\{\mathbf{v}_i\}$，并将其解释为下一训练轮次中的重启。

然而，目前尚无理论或实证分析能够证明，动量可以消除异质性；由于局部激活统计之间存在差异，异质性会损害分布式深度训练 [Hsieh 等人， 2020]。我们不仅考虑了非凸函数，还将一种不受数据异质性影响的技术——梯度跟踪——融入 D-SUM，以减轻异构数据对去中心化训练的影响，从而提升模型泛化能力，具体如下。

2.3 GT-DSUM算法

本小节进一步探讨异质性会阻碍局部动量加速这一事实 [Lin 等人， 2021]，并介绍 UMP 中的第二种算法，称为 GT-DSUM。该算法旨在通过应用梯度跟踪技术，更好地泛化共识模型参数，并减轻异构数据的影响。

考虑到工作节点本地数据划分之间的差异，GT 引入了一个额外的、位于工作节点侧的辅助变量 $\mathbf{y}_i^{(t)}, \forall i$，旨在渐近跟踪 ∇f_i 的平均值，前提是每个轮次 t 都可以获得本地精确梯度。直观而言，GT 不受异质性的影响，而随着轮次 t 增加，$\mathbf{y}_i^{(t)}$ 近似等于全局梯度方向。受此启发，我们将 GT 引入 D-SUM，得到 GT-DSUM。具体而言，我们使用小批量梯度 $\mathbf{g}_{\cdot,i}^{(t)\ \top}$ 对所应用的梯度$\mathbf{m}_{\cdot,i}^{(t)\ \top}$进行归一化，并使用阻尼因子 λ 对$\mathbf{y}_i^{(t)}$ 进行归一化，以凸显本地更新的必要性。详细算法见算法2。在本地更新过程中，第 5 行使用 D-SUM 更新模型参数，但采用了归一化项$\mathbf{m}_{\cdot,i}^{(t)\ \top}$。第 7 行和第 8 行与算法1中的基本 D-SUM 步骤相同。对于 GT-DSUM，我们使用第 9 行所示的两个连续同步模型之差，在第 10 行以类谣言传播方式 [更新梯度跟踪变量 Xin 等，2021b；Xin 等，2021a]。特别地，当 $K=1$、$\lambda=1$ 和 $\beta=0$ 时，算法2可简化为原始 GT 算法[Koloskova等，2021] 实例。

由于算法1和算法2在参数交换过程中采用了多次共识步骤，这会显著增加通信成本，因此我们在第 3 节应用通信压缩技术 GRACE [Xu 等人， 2021] 以在模型泛化与通信开销之间进行权衡。

2.4 理论分析

下面我们将介绍 UMP 中两种算法在一般非凸设置下的收敛性分析。详细证明见附录。首先，我们给出全文所采用的假设。

假设1 (*L*-光滑)。 对于每个函数 f_i： $\mathbb{R}^d \to \mathbb{R}$均可微，并且存在常数 $L>0$，使得对于任意 $\mathbf{x}$、$\mathbf{x}' \in \mathbb{R}^d$： $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \le L\|\mathbf{x}-\mathbf{x}'\|$。

假设2 (有界方差)。 我们假设，对于任意 i、$\mathbf{x} \in \mathbb{R}^d$，存在 $\sigma > 0$ 和 $\zeta > 0$，使得 $\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\|\nabla F_i(\mathbf{x}\ \xi_i) - \nabla f_i(\mathbf{x})\|^2 \le \sigma^2$，以及 $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \le \zeta^2$。

假设 1 和 2 是一般非凸目标函数文献中的标准假设 [Lin 等，2021；Yu 等，2019；Koloskova等， 2020] 旨在确保损失函数的基础连续性，并限制分布式场景中异质性的影响。注意，当 $\zeta = 0$时，我们有$\nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$，即，所有参与工作节点的数据分布均为 IID 时，问题便简化为这种情形。第三个常见假设是随机梯度一致有界，具体如下。

假设3 （随机梯度有界）。 我们假设对于任意i，随机梯度的二阶矩均有界：$\mathbf{x} \in \mathbb{R}^d, \mathbb{E}_{\xi_i\sim\mathcal{D}_i}\|\nabla F_i(\mathbf{x}\ \xi_i)\|^2 \le G^2$。

假设4。 根据定义 *1*，混合矩阵是双随机的。此外，对于任意矩阵 $\mathbf{Z} \in \mathbb{R}^{d\times n}$，定义 $\bar{\mathbf{Z}} = \mathbf{Z}\frac{1}{n}\mathbf{1}\mathbf{1}^\top$。于是，混合矩阵满足 $\mathbb{E}_{\mathbf{W}}\|\mathbf{Z}\mathbf{W} - \bar{\mathbf{Z}}\|_F^2 \le (1-\rho)\|\mathbf{Z} - \bar{\mathbf{Z}}\|_F^2$。

在假设4中，我们假设 $\rho := 1 - \max\{|\lambda_2(\mathbf{W})|, |\lambda_n(\mathbf{W})|\}^2 > 0$，其中令 $\lambda_i(\mathbf{W})$ 表示混合矩阵 $\mathbf{W}$ 的第 i 大特征值，且$-1 \le \lambda_n(\mathbf{W}) \le \cdots \le \lambda_2(\mathbf{W}) \le \lambda_1(\mathbf{W}) \le 1$。例如，当 $\rho=1$ 适用于全网状（完全）通信拓扑时，通常使用 ρ 的值。

D-SUM的收敛性分析

现在，我们给出算法1中 D-SUM（红色标注）的收敛结果。详细证明见附录B

定理1。 考虑上述假设下的问题（*1*），记 $\beta_0 = \max\{1+\beta, 1+\alpha\beta\}$，对于算法*1*中的所有 $T \ge 1$ 和 $K \ge 1$，其学习率为$\eta \le \frac{\rho}{5L}$ 且参数满足 $\frac{4-\rho}{2} < \frac{1}{\beta_0^2}$，我们有

$$\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})\right\|^2 \le \frac{2\left(f(\mathbf{x}_0)-f^*\right)}{\tilde{\eta}KT} + \frac{2\beta^2\hat{\eta}^2L^2G^2}{n(1-\beta)^4} + \frac{2L^2C_1}{n^2(1-Q_1)} + \frac{L}{n}\left(\sigma^2 + 2\tilde{\eta}G^2 + 3\sigma^2\tilde{\eta}\right),$$

其中 $\tilde{\eta} = \frac{\eta}{1-\beta}$、$\hat{\eta} = ((1-\beta)\alpha - 1)\ \eta$、$Q_1 = 2\beta_0^2(1 - \frac{\rho}{4})$，以及 $C_1 = 24\eta^2\beta_0^2\zeta^2/\rho + 4(1-\rho)\left(1+2\alpha\beta+2\alpha^2\beta^2\right)\eta^2\sigma^2$。

备注 1。 定理*1*为一般神经网络的D-SUM 提出了一个非渐近收敛界，因为第二项即，$\mathcal{O}\left(\frac{L^2\eta^2}{n}\right)$源自式（*7*）中的核心随机梯度下降步骤。直观而言，存在一个合适的 α，可在实践中实现最佳训练性能；单节点情形中也观察到了这一点 *[Yan 等， 2018]*。在第 *3* 节中，我们进行了相关实验以验证这一推测。

GT-DSUM 的收敛性分析

下面的定理给出了在 $K=1$ 且工作节点之间采用固定通信拓扑这一方便设定下，算法 2 中 GT-DSUM 的收敛结果，详细证明见附录 C。基于 GT 所要解决的问题是如何应用小批量梯度估计来跟踪全局优化下降方向，我们定义以下命题来阐明这一点。

命题1（梯度平均跟踪器 [Di Lorenzo 和 Scutari， 2016]）。 我们假设一个较宽松的约束，即辅助变量 $\mathbf{y}_i^{(t)}$ 被视为平均值的跟踪器$\frac{1}{n}\sum_{j=1}^n \nabla f_j\left(\mathbf{x}_i^{(t)}\right)$，这意味着对于任意轮次 t，都有$\mathbb{E}\left\|\mathbf{y}_i^{(t)} - \frac{1}{n}\sum_{j=1}^n \nabla f_j\left(\mathbf{x}_i^{(t)}\right)\right\|^2 \le \epsilon^2$。

定理 2。 考虑满足所列特定假设的问题 *(1)*，记 $\beta_0 =$ 最大值 $\{1+\alpha\beta, 1+\beta\}$，并在不含多个本地步数的算法*2* 中设置$T \geq 1$ （即， $K = 1$），学习率 η 选为

$$0 \leq \eta \leq \min\left\{\frac{\rho}{12\lambda}, \frac{1-\beta}{2}, \frac{3+\beta}{2\sqrt{3}\lambda(1+\alpha\beta)}, \frac{1+\beta}{2\sqrt{3}\lambda\alpha\beta}\right\}\frac{1}{L}$$

且参数满足

$$\begin{cases}(1+\alpha\beta)(1-\lambda) \leq \dfrac{1}{2\sqrt{2}},\\ \rho \leq \dfrac{48\lambda^2L^2}{(1-\lambda)^2},\\ 4\beta_0^2\left(1-\dfrac{\rho}{4}\right) < 1,\\ 8(1-\rho)(1+\alpha\beta)^2 < 1,\end{cases}$$

我们有

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2\\ &\leq \frac{2\left(f(\mathbf{x}_0)-f^*\right)}{\tilde{\eta}T} + \frac{8\beta^2\hat{\eta}^2L^2}{n(1-\beta)^4}\left(\sigma^2+3\zeta^2+4G^2\right)\\ &\quad + \frac{12L^2}{n}\left(\frac{Q_3}{T} + \frac{(2-Q_2)Q_4}{(1-Q_2)T} + \frac{V_{\max}}{1-Q_2} + \frac{C_2}{1-Q_2}\right)\\ &\quad + \frac{4\lambda^2\sigma^2+16\zeta^2}{n} + 8(1-\lambda^2)\epsilon^2 + \frac{12\beta^2\hat{\eta}^2L^2\epsilon^2}{(1-\beta)^4}\end{aligned}$$

其中 $\tilde{\eta} = \frac{\eta}{\beta_0^2\sqrt{1-\beta}}$ 且 $\hat{\eta} = (\alpha - \alpha\beta - 1)\eta$。此外，$C_2 = \frac{\beta_0^2(1-\frac{\rho}{2})}{L^2}(\sigma^2 + \zeta^2)(192\lambda^2L^2 + \rho)$，$Q_2 \triangleq \min\left\{4\beta_0^2\left(1-\frac{\rho}{4}\right), 8(1-\rho)(1+\alpha\beta)^2\right\}$，$Q_3 = 6(\zeta^2+\sigma^2)$，$Q_4 = \beta_0^2\left(1-\frac{\rho}{2}\right)(\sigma^2+\zeta^2)\left(48+\frac{\rho}{L^2}+192\lambda^2\right)$。进一步地，我们定义 $V_{\max} \triangleq \max_{0\leq t\leq T-1}\left\{\frac{1}{n}\left(\mathbb{E}\left\|\boldsymbol{X}^{(t)} - \bar{\boldsymbol{X}}^{(t)}\right\|_F^2 + \mathbb{E}\left\|\boldsymbol{Y}^{(t)} - \bar{\boldsymbol{Y}}^{(t)}\right\|_F^2\right)\right\}$.

注释 2。 定理 *2* 右侧的第四项即 $\mathcal{O}\left(\frac{1}{nT} + \frac{\beta_0^2}{nT} + \frac{1}{n\beta_0^2}\right)$ 来自算法*2*第10行中用于搜索全局最优下降估计的额外 *GT* 步骤。然而，由于其阶数较高，当 α 缩放时，这一项可能占主导地位。显然，当 $\alpha > 1$时， $\beta_0 = 1+\alpha\beta$，且其收敛速度为 $\mathcal{O}\left(\frac{\beta_0^2}{nT}\right)$；因此，如果 α 增大，从收敛角度来看，性能会显著恶化。接下来，我们将在第 *3* 节展示 α 的影响。

3 评估

我们的主要评估结果表明，在模型准确率方面，D-SUM 优于其他方法；在不同程度的非独立同分布情况下，GT-DSUM 的性能更优。所有实验均在 CPU/GPU 集群上执行，该集群配备了 Inter(R) Xeon(R) Gold 6126、4 张 GTX 2080Ti 显卡和 12 张 Tesla T4 显卡。我们使用 PyTorch 和 Ray [Moritz等人， 2018] 来实现和训练模型。

3.1 实验方法

基线方法。 我们考虑以下三种带动量的去中心化方法，具体如下： •本地随机梯度下降 [Stich 定期在所有工作节点之间平均模型参数。与原始随机梯度下降相比，每个节点独立运行带重球动量的单节点随机梯度下降。 • *QG-DSGDm* [Lin 等人提出的方法， 2021]模拟全局优化方向，并将准全局动量整合到本地随机梯度中，而不会产生额外的通信成本。该方法在实证上减轻了数据异质性的影响。 • *SlowMo* [Wang 等人,2020] 在执行多步随机梯度下降后，通过全归约模式（模型平均）进行缓慢的周期性动量更新。为简单起见，我们采用常见的小批量梯度作为本地更新方向。

数据集与模型。 我们研究了计算机视觉（CV）和自然语言处理（NLP）任务中的去中心化行为，涉及 MNIST、EMNIST、CIFAR10 和 AG NEWS。对于所有计算机视觉任务，我们训练不同的卷积神经网络模型。对于自然语言处理，我们训练了一个循环神经网络，其中包含一个嵌入层和一个丢弃层，后接一个全连接层。模型描述见附录 D。

超参数。 对于不同基准上的所有算法，默认设置为部署 10 个工作节点。在我们的实验中，CI-FAR10 的本地 mini-batch 大小设为 256，其余数据集设为 128，本地更新次数设为 $K = 10$。为展示去中心化深度训练中数据异质性带来的挑战，我们采用狄利克雷分布参数 [Lin 等， 2021] 以控制不同程度的非独立同分布程度；对于非独立同分布 $= 0.1$、1、10 的情况，该值越小，工作节点越可能只持有来自单一标签类别的样本（即，非独立同分布 $= 0.1$可视为一种极端的数据偏斜情况）。此外，默认情况下，标量 α、动量 β和归一化参数 λ分别设为 2 、 0.9 和 0.8。在实践中考虑的 $\mathbf{W}$ 选项中，我们使用广受欢迎的 Metropolis-Hastings 规则 [Koloskova 等预先构造了一个动态拓扑变化序列，使拓扑结构从全网状拓扑逐步变化为环形拓扑2021] 即， $w_{ij} = w_{ji} = \min\left\{\frac{1}{i\deg()+1}, \frac{1}{\deg(j)+1}\right\}$ 对于任意 i， j、$w_{ii} = 1 - \sum_{j=1}^{n} w_{ij}$。对于每种算法和数据集，我们通过在集合 $\{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{-0.5}\}$ 上进行网格搜索来微调学习率 η。

性能指标。 我们研究不同动量变体对去中心化深度学习的影响，包括

- 模型泛化 通过模型正确预测的数据量与测试数据集中的全部数据量之比进行衡量。我们报告本地模型在测试样本上的平均模型性能。
- 不同超参数的影响 通过调整超参数值进行探究，以研究 D-SUM 和 GT-DSUM 的特性。
- 可扩展性 是处理分布式场景中的任务时一项关键属性。

3.2 评估结果

与基线方法的性能比较。从表1可以看出，我们提出的算法优于所有其他基线方法

表1：不同非独立同分布程度下，各类算法在不同训练基准上的测试准确率。

数据集	算法	测试准确率（%）		
		非独立同分布 = 0.1	非独立同分布 = 1	非独立同分布 = 10
MNIST [LeCun等人， 1998]	带动量的本地 SGD	95.66 ± 0.21	97.99 ± 0.03	98.39 ± 0.03
	QG-DSGDm	96.02 ± 0.19	97.46 ± 1.36	98.21 ± 0.04
	SlowMo	97.32 ± 0.02	97.93 ± 0.07	98.34 ± 0.06
	D-SUM（我们的方法）	**97.89** ± 0.21	**98.77** ± 0.04	**98.94** ± 0.01
	GT-DSUM（我们的方法）	97.51 ± 0.61	98.70 ± 0.01	98.82 ± 0.03
EMNIST [Cohen 等， 2017]	带动量的本地 SGD	45.90 ± 1.21	36.77 ± 0.13	38.29 ± 0.03
	QG-DSGDm	46.03 ± 0.6	46.02 ± 0.12	36.72 ± 0.02
	SlowMo	45.52 ± 0.03	37.11 ± 0.01	37.50 ± 0.0
	D-SUM（我们的方法）	49.68 ± 0.43	49.75 ± 0.05	42.50 ± 0.01
	GT-DSUM（我们的方法）	**50.49** ± 0.82	**50.25** ± 0.07	**51.87** ± 0.02
CIFAR10 [Krizhevsky 等， 2009]	带动量的本地 SGD	22.94 ± 1.11	42.93 ± 0.85	52.82 ± 0.01
	QG-DSGDm	26.34 ± 1.42	49.12 ± 038	54.03 ± 0.24
	SlowMo	31.06 ± 1.27	50.46 ± 0.04	55.50 ± 0.10
	DSUM（我们的方法）	31.16 ± 1.27	54.34 ± 0.11	57.59 ± 1.05
	GT-DSUM（我们的方法）	**36.16** ± 0.74	**56.95** ± 1.56	**59.34** ± 1.55
AG NEWS [Zhang 等人, 2015]	带动量的本地 SGD	75.51 ± 0.44	77.98 ± 0.39	80.66 ± 0.02
	QG-DSGDm	78.82 ± 0.31	79.33 ± 0.38	82.24 ± 0.02
	SlowMo	82.57 ± 0.03	83.17 ± 0.01	83.79 ± 0.01
	DSUM（我们的方法）	84.13 ± 0.55	85.46 ± 0.31	87.52 ± 0.04
	GT-DSUM（我们的方法）	**84.29** ± 0.37	**87.59** ± 0.18	**89.07** ± 0.04

在不同数据偏斜程度下的表现。对于 CIFAR10 和 AG NEWS，我们的算法与基准方法的表现如下：GT-DSUM > D-SUM > SlowMo > QG-DSGDm > 带动量的本地 SGD。我们提出的算法在模型泛化方面优于其他基准方法，并表明 GT 技术能有效减轻数据异质性带来的负面影响。随着非独立同分布程度增加，GT-DSUM 在 CIFAR10 上的准确率比带动量的本地 SGD 高出 57。6%。

本地更新的影响。本地更新的数量 K 是最重要的参数之一，因为它会影响最终的模型泛化和训练时间。与往常一样，K 的数量设定为小于 20 [Reddi 等人，2020；Qu等人，2022]。因此，我们展示了与 $K \in \{1, 5, 10, 15, 20\}$的比较。我们从图1的结果中观察到两点。首先，无论 K如何，我们的算法表现都优于带动量的本地 SGD。对于 CIFAR10 和 AG NEWS，如图 1(c) 、 1(d) 所示，我们观察到，随着本地更新数量增加，GT-DSUM 始终保持竞争力。其中，当 $K = 5$时，其准确率比带动量的本地 SGD 提高了 53.8%。其次，我们发现，工作节点未必能通过增加本地更新数量来显著改善模型泛化 K。此外，当$K = 1$时，所有基准的表现都最差；而当 K 过大时，由于分布式优化中工作节点的本地模型会彼此偏离过远，性能可能会下降 [Qu 等人，2022]。

影响 β。我们通过在集合 $\{0, 01, 0.1, 0.5, 0.9, 0.99\}$ 上进行网格搜索，进一步研究了动量项 β，以评估处理算法的不同策略。第三组模拟评估了模型在不同β上的准确率表现，如图2所示。我们从这些结果中得到了一个重要观察。无论数据集还是模型，较大的 β （例如，0.9、0.99）通常比小的（例如，0.01、0.1）表现更好。此外，可以观察到，在 CIFAR10 和 AG NEWS 上，测试准确率随 β 单调上升。我们注意到，随着 β 增大，GT-DSUM 的模型准确率高于带动量的本地 SGD。在所有任务中，与带动量的本地 SGD 相比，其测试准确率最高提升了 56。8%。

对 α的敏感性。我们将最重要的参数 α 从 0 调整到 15，以分析其对收敛性能的影响。根据表 2，可以得出以下两点观察结果。首先，我们注意到，在具有相同数据异质性水平（例如，非独立同分布 $= 1$）的不同数据集上评估 D-SUM 时，最优的 α值总是不同。由于数据集和模型各自的特性不同，很难且耗时地确定最优的 α。其次，随着 α 增加，GT-DSUM 的模型性能显著下降。这一现象验证了注释 2 中的详细分析，该分析表明，与 D-SUM 相比，GT-DSUM 对 α 的约束要求更严格，才能确保模型有效性。就经验结果而言，α 对基于动量的优化器的脆弱性，与其带来优异性能的鲁棒性之间仍存在差距。

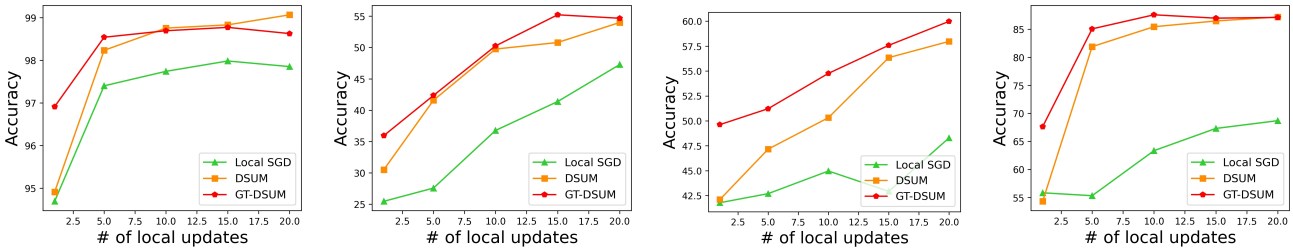
可扩展性。最后，我们针对不同数量的工作节点进行训练，并与带动量的本地 SGD 进行比较，考察非独立同分布= 1的情况。我们通过调整训练所用设备数量 n，将规模扩展至 4、10、16 和 32 个工作节点，以此进行评估。结果如图3所示。随着参与训练的工作节点数量增加，我们的方法优势愈发明显：在这一非独立同分布场景下，我们的 GT-DSUM 方法始终比带动量的本地 SGD 达到更高的模型准确率。

4 结论

本文提出了一种统一的基于动量的范式 UMP，其中包含两种算法：D-SUM 和 GT-DSUM。前者

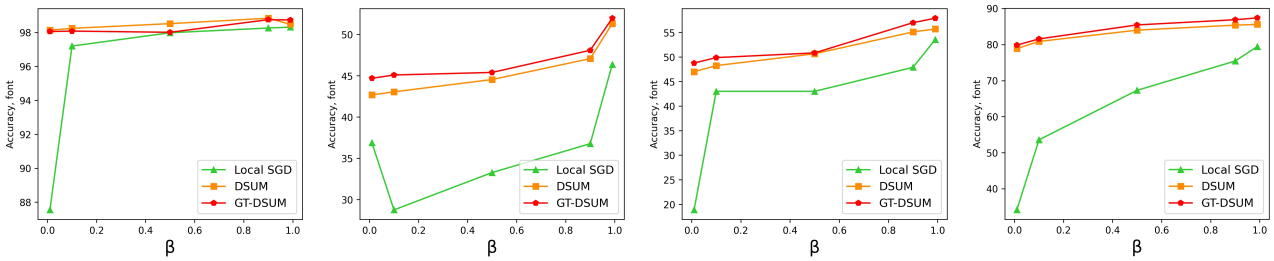
表 2： α 对 D-SUM 和 GT-DSUM 在非独立同分布 $= 1$下测试准确率的影响。“⋆”表示不收敛。

数据集	方法	在非独立同分布 $= 1$ 情况下，使用不同 α 评估的测试准确率（%）									
		$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$	$\alpha = 8$	$\alpha = 10$	$\alpha = 15$
MNIST	D-SUM	98.05	98.40	98.57	98.76	98.85	98.76	98.85	**99.09**	98.87	92.86
	GT-DSUM	98.18	98.50	98.70	98.70	**98.80**	⋆	⋆	⋆	⋆	⋆
EMNIST	D-SUM	37.1	43.58	35.58	49.75	**55.32**	49.80	43.00	48.47	53.04	⋆
	GT-DSUM	33.72	46.06	47.10	**50.25**	39.84	⋆	⋆	⋆	⋆	⋆
CIFAR10	D-SUM	47.10	50.85	51.68	50.32	**54.83**	51.10	51.53	50.68	⋆	⋆
	GT-DSUM	45.98	49.80	50.55	54.77	**57.58**	54.43	⋆	⋆	⋆	⋆
AG NEWS	D-SUM	79.16	81.78	83.72	85.46	86.38	86.36	88.20	87.07	**88.82**	88.56
	GT-DSUM	78.90	83.12	85.34	**87.59**	86.89	77.67	⋆	⋆	⋆	⋆



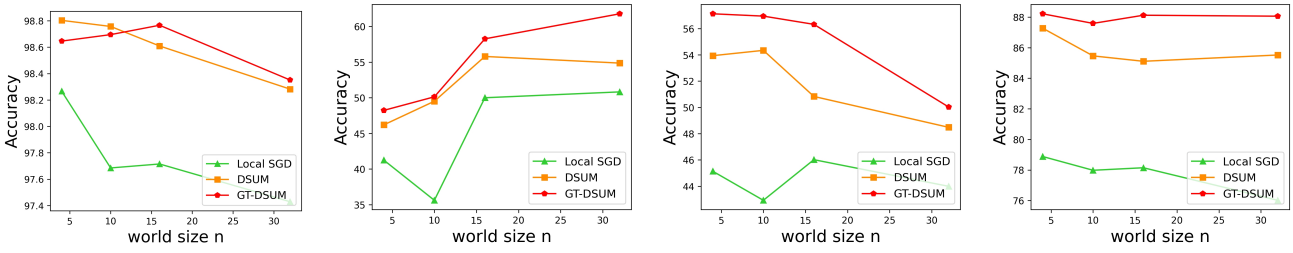
(a) LeNet 在 MNIST 上　(b) 在 EMNIST 上使用卷积神经网络　(c) LeNet 在 CIFAR10 上　(d) 循环神经网络在 AG NEWS 上

图1：本地更新次数 K 对非独立同分布 $= 1$ 情形下动量 $\beta = 0.9$ 时收敛的影响。



(a) LeNet 在 MNIST 上　(b) 卷积神经网络在 EMNIST 上　(c) LeNet 在 CIFAR10 上　(d) AG NEWS 上的循环神经网络

图2：在非独立同分布 $= 1$ 情况下，本地更新次数 $K = 10$ 对动量 β 收敛的影响。



(a) MNIST 上的 LeNet　(b) EMNIST 上的卷积神经网络　(c) CIFAR10 上的 LeNet　(d) AG NEWS 上的循环神经网络

图3：在非独立同分布 $= 1$ 情况下，动量世界大小对收敛的影响，其中本地更新次数为 $K = 10$ 。

可获得良好的模型泛化能力，解决了其在非凸情形下的有效性问题；后者则通过应用 GT 技术进一步发展，以消除异构数据的负面影响。通过推导一般情形下的收敛性

在非凸设置下，这些算法能够取得具有竞争力的性能，且其表现与关键参数 α 密切相关。大量实验结果表明，我们的 UMP 至多可使准确率提升 57.6%。

参考文献

[Cohen 等， 2017] Gregory Cohen、Saeed Afshar、Jonathan Tapson 和 Andre Van Schaik。EMNIST：将 MNIST 扩展至手写字母。*2017*年国际神经网络联合会议（*IJCNN*），第 2921–2926 页，2017年。

[Cutkosky 和 Mehta， 2020] Ashok Cutkosky 和 Harsh Mehta。动量改进了归一化随机梯度下降。国际机器学习会议，第 2260–2268 页，2020 年。

[Deng 和 Gao，2021] Qi Deng 和 Wenzhi Gao。用于随机弱凸优化的小批量和基于动量模型的方法。神经信息处理系统进展，34，2021。[Di Lorenzo 和 Scutari， 2016] Paolo Di Lorenzo 和 Gesualdo Scutari。Next：网络内非凸优化。*IEEE*网络信号与信息处理汇刊，2(2):120–136，2016。[Esfandiari等，2021] Yasaman Esfandiari、Sin Yong Tan、Zhanhong Jiang、Aditya Balu、Ethan Herron、Chinmay Hegde 和 Soumik Sarkar。面向非独立同分布数据去中心化学习的交叉梯度聚合。国际机器学习会议，第 3036–3046 页，2021。[Fang等， 2018] Cong Fang、Chris Junchi Li、Zhouchen Lin 和 Tong Zhang。Spider：通过随机路径积分微分估计器实现近乎最优的非凸优化。神经信息处理系统进展，31，2018。[Ghadimi等，2015] Euhanna Ghadimi、Hamid Reza Feyzmahdavian 和 Mikael Johansson。重球法在凸优化中的全局收敛性。*2015*年欧洲控制会议（*ECC*），第 310–315 页，2015年。[Han 和 Gao， 2021] Andi Han 和 Junbin Gao。用于非凸优化的黎曼随机递归动量法。国际人工智能联合会议，2021年。[Hsieh等，2020] Kevin Hsieh、Amar Phanishayee、Onur Mutlu 和 Phillip Gibbons。去中心化机器学习中的非独立同分布数据困境。国际机器学习会议，第 4387–4398 页，2020年。[Karimireddy等，2020] Sai Praneeth Karimireddy、Satyen Kale、Mehryar Mohri、Sashank Reddi、Sebastian Stich 和 Ananda Theertha Suresh。SCAFFOLD：用于联邦学习的随机控制平均。国际机器学习会议，第 5132–5143 页，2020年。[Kingma 和 Ba， 2015] Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。2015年。[Koloskova等，2019] Anastasia Koloskova、Sebastian Stich 和 Martin Jaggi。采用压缩通信的去中心化随机优化与 Gossip 算法。国际机器学习会议，第 3478–3487 页，2019年。[Koloskova等， 2020] Anastasia Koloskova、Nicolas Loizou、Sadra Boreiri、Martin Jaggi 和 Sebastian Stich。

关于拓扑结构变化和本地更新的去中心化随机梯度下降统一理论。国际机器学习会议，第5381–5393页，2020年。[Koloskova 等人， 2021] Anastasiia Koloskova、Tao Lin 和 Sebastian U Stich。去中心化机器学习中梯度跟踪的改进分析。神经信息处理系统进展，第34卷，2021。[Krizhevsky 等人， 2009] Alex Krizhevsky、Geoffrey Hinton 等人。从微型图像中学习多层特征。2009年。[LeCun 等人， 1998] Yann LeCun、L´eon Bottou、Yoshua Ben- gio 和 Patrick Haffner。应用于文档识别的基于梯度的学习。IEEE会刊，86(11):2278–2324，1998年。[Lian 等人， 2017] Xiangru Lian、Ce Zhang、Huan Zhang、Cho-Jui Hsieh、Wei Zhang 和 Ji Liu。去中心化算法能否优于集中式算法？去中心化并行随机梯度下降案例研究。神经信息处理系统进展，第30卷，2017年。[Lin 等人，2021] Tao Lin、Sai Praneeth Karimireddy、Se bas- tian U Stich 和 Martin Jaggi。准全局动量：加速异构数据上的去中心化深度学习。国际机器学习会议，第6654–6665页，2021。[Mai 和 Johansson，2020] Vien Mai 和 Mikael Johansson。带动量的随机梯度法在非光滑非凸优化中的收敛性。国际机器学习会议，第6630–6639页，2020年。[Moritz 等人，2018] P hilipp Moritz、Robert Nishihara、Stephanie Wang、Alexey Tumanov、Richard Liaw、Eric Liang、Melih Elibol、Zongheng Yang、William Paul、Michael I. Jordan 和 Ion Stoica。Ray：面向新兴人工智能应用的分布式框架。第13届USENIX操作系统设计与实现大会论文集，第561–577页，2018年。[Nesterov， 1983] Yurii E Nesterov。一种求解凸规划问题的方法及其收敛速度$o(1/k^2)$。苏联科学院院刊，269:543–547，1983年。[Polyak，1964] Boris T Polyak。加速迭代方法收敛的一些方法。苏联计算数学与数学物理，4(5):1–17，1964年。[Qu 等人， 2022] Zhe Qu、Xingyu Li、LuZhuo Duan、Rui、Yao Liu 和 Bo Tang。通过锐度感知最小化实现广义联邦学习。国际机器学习会议，2022。[Reddi 等人， 2020] Sashank J Reddi、Zachary Charles、Manzil Zaheer、Zachary Garrett、Keith Rush、Jakub Kone˘cn`y、Sanjiv Kumar 和 Hugh Brendan McMahan。自适应联邦优化。国际学习表征会议，2020年。[Shang 等人， 2022] Xinyi Shang、Yang Lu、Gang Huang 和 Hanzi Wang。通过联邦分类器再训练实现异构和长尾数据上的联邦学习

特征。国际人工智能联合会议，第2218–2224页，2022。[Stich， 2018] Sebastian U Stich。本地随机梯度下降收敛快、通信开销低。国际学习表征会议，2018年。[Tao 等人， 2021] Wei Tao、Sheng Long、Gaowei Wu 和 Qing Tao。动量参数在自适应 Polyak 重球法最优收敛中的作用。第9届国际学习表征会议（ICLR），2021。[Tao 等人， 2022] Youming Tao、Yulian Wu、Xiuzhen Cheng 和 Di Wang。重新审视重尾数据下的私有随机凸优化与稀疏学习。国际人工智能联合会议，2022。[Wan 等人， 2022] Wei Wan、Shengshan Hu、Jianrong Lu、Leo Yu Zhang、Hai Jin 和 Yuanyuan He。为联邦学习提供防护：采用自适应客户端选择的鲁棒聚合。国际人工智能联合会议，2022。[Wang 等人， 2020] Jianyu Wang、Vinayak Tantia、Nicolas Ballas 和 Michael G. Rabbat。Slowmo：通过慢动量改进通信高效的分布式随机梯度下降。第8届国际学习表征会议（ICLR），2020年。[Xin等人，2021a] Ran Xin、Usman Khan 和 Soummya Kar。用于去中心化随机非凸优化的混合方差缩减方法。国际机器学习会议，第11459–11469页，2021。[Xin等人，2021b] Ran Xin、Usman A Khan 和 Soummya Kar。去中心化在线随机非凸优化的改进收敛性分析。IEEE信号处理汇刊，69:1842–1858，2021。[Xu 等人， 2021] Hang Xu、Chen-Yu Ho、Ahmed M. Abdel- moniem、Aritra Dutta、EH Bergou、Konstantinos Karat- senidis、Marco Canini 和 Panos Kalnis。GRACE：面向分布式机器学习的压缩通信框架。载于第41届IEEE国际分布式计算系统会议（ICDCS）论文集，2021。[Yan 等人，2018] Yan Yan、Tianbao Yang、Zhe Li、Qihang Lin 和 Yi Yang。深度学习随机动量方法的统一分析。第27届国际人工智能联合会议论文集，第2955–2961页，2018年。[Yu 等人， 2019] Hao Yu、Rong Jin 和 Sen Yang。分布式非凸优化中通信高效的动量随机梯度下降线性加速分析。国际机器学习会议，第7184–7193页，2019年。[Zhang 等人， 2015] Xiang Zhang、Junbo Zhao 和 Yann Le- Cun。用于文本分类的字符级卷积网络。神经信息处理系统进展，第28卷，2015年。[Zhang 等人， 2022] Hong Zhang、Ji Liu、Juncheng Jia、Yang Zhou、Huaiyu Dai 和 Dejing Dou。FedDUAP：使用动态更新和自适应剪枝的联邦学习服务器上的共享数据。国际人工智能联合会议，2022。

一个前提

为了对所有提出的算法的收敛结果进行理论分析，我们首先给出以下一些预备事实：

- **事实1：** 对于任意随机向量 $\mathbf{a}$，有 $\mathbb{E}\|\mathbf{a}\|^2 = \mathbb{E}\|\mathbf{a} - \mathbb{E}[\mathbf{a}]\|^2 + \|\mathbb{E}[\mathbf{a}]\|^2$。
- **事实2：** 对于任意 $a > 0$，有 $\pm\langle \mathbf{a}, \mathbf{b}\rangle \leq \frac{1}{2a}\|\mathbf{a}\|^2 + \frac{a}{2}\|\mathbf{b}\|^2$。
- **事实3：** $\langle \mathbf{a}, \mathbf{b}\rangle = -\frac{1}{2}\|\mathbf{a}\|^2 + \frac{1}{2}\|\mathbf{b}\|^2 - \frac{1}{2}\|\mathbf{a} - \mathbf{b}\|^2$。
- **事实4：** 给定两个向量 $\mathbf{a}$ 和 $\mathbf{b}$ $\forall a > 0$，有 $\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+a)\|\mathbf{a}\|^2 + (1+\frac{1}{a})\|\mathbf{b}\|^2$ – .
- **事实5：** 对于任意的 n 向量集合 $\{\mathbf{a}_i\}_{i=1}^n$，我们有 $\|\sum_{i=1}^n \mathbf{a}_i\|^2 \leq n\sum_{i=1}^n \|\mathbf{a}_i\|^2$。
- **事实6：** 假设 $\{b_i\}_{i=1}^n$ 和 $\{\mathbf{a}_i\}_{i=1}^n$ 分别为非负标量和向量的集合。我们定义 $s = \sum_{i=1}^n b_i$。根据詹森不等式，我们有 $\|\sum_{i=1}^n b_i\mathbf{a}_i\|^2 = s^2\|\sum_{i=1}^n \frac{b_i}{s}\mathbf{a}_i\|^2 \leq s^2 \cdot \sum_{i=1}^n \frac{b_i}{s}\|\mathbf{a}_i\|^2 = s\cdot\sum_{i=1}^n b_i\|\mathbf{a}_i\|^2$。 – –

事实4 中的不等式对于两个矩阵$\mathbf{A}$、$\mathbf{B}$ 之和的弗罗贝尼乌斯范数也成立。

命题 2. 使用定义 *1* 中定义的混合矩阵$\mathbf{W}$ 进行一步八卦平均，可以保持迭代的平均值不变，即，$\mathbf{XW}\frac{\mathbf{11}^\top}{n} = \mathbf{X}\frac{\mathbf{11}^\top}{n}$。

请注意，我们的方案对动量作用的观察如下，现在我们陈述一个基本引理：

引理 1. 设引入一个辅助变量$\mathbf{y}_i^{(t)\,\tau} = \frac{\beta}{1-\beta}\left(\mathbf{x}^{(t),\,\tau}_{,\,i} - \mathbf{v}_i^{(t)\,\tau}\right)$，，我们定义$\mathbf{z}^{(t),\,\tau}_{,\,i} \triangleq \mathbf{x}^{(t),\,\tau}_{,\,i} + \mathbf{y}^{(t),\,\tau}_{,\,i}$以及$\mathbf{c}^{(t),\,\tau}_{,\,i} \triangleq \frac{1-\beta}{\beta}\mathbf{y}^{(t),\,\tau}_{,\,i}$。那么我们有

$$\mathbf{z}_i^{(t),\tau+1} = \mathbf{z}_i^{(t),\tau} - \frac{\eta}{1-\beta}\mathbf{g}_i^{(t),\tau} \tag{9}$$

and

$$\mathbf{c}_i^{(t),\tau+1} = \beta\mathbf{c}_i^{(t),\tau} + (\alpha - \alpha\beta - 1)\,\eta\mathbf{g}_i^{(t),\tau}. \tag{10}$$

证明。对于（9），从（7）中$\mathbf{v}^{(t)}_{,i}{}^{\tau}$的定义出发，

$$\begin{aligned}
\mathbf{z}_i^{(t),\tau+1} &= \frac{1}{1-\beta}\mathbf{x}_i^{(t),\tau+1} - \frac{\beta}{1-\beta}\left(\mathbf{x}_i^{(t),\tau} - \alpha\eta\mathbf{g}_i^{(t),\tau}\right) \\
&= \mathbf{x}_i^{(t),\tau} + \frac{\beta}{1-\beta}\left(\mathbf{x}_i^{(t),\tau} - \mathbf{x}_i^{(t),\tau-1} + \alpha\eta\mathbf{g}_i^{(t),\tau-1}\right) - \frac{\eta}{1-\beta}\mathbf{g}_i^{(t),\tau} \\
&= \mathbf{x}_i^{(t),\tau} + \mathbf{y}_i^{(t),\tau} - \frac{\eta}{1-\beta}\mathbf{g}_i^{(t),\tau}.
\end{aligned}$$

对于（10），同理，

$$\begin{aligned}
\mathbf{c}_i^{(t),\tau+1} &= \mathbf{x}^{(t),\tau} - \eta\mathbf{g}_i^{(t),\tau} - \mathbf{x}_i^{(t),\tau} + \alpha\eta\mathbf{g}_i^{(t),\tau} + \beta\left(\mathbf{x}^{(t),\tau} - \alpha\eta\mathbf{g}_i^{(t),\tau} - \mathbf{x}^{(t),\tau-1} + \alpha\eta\mathbf{g}_i^{(t),\tau-1}\right) \\
&= \beta\left(\mathbf{x}_i^{(t),\tau} - \mathbf{v}_i^{(t),\tau}\right) + (\alpha - \alpha\beta - 1)\,\eta\mathbf{g}_i^{(t),\tau}.
\end{aligned} \tag{11}$$

□

B D-SUM的证明

由于 f的光滑性，可得

$$\mathbb{E}_{t,\tau} f(\bar{\mathbf{z}}^{(t),\tau+1}) - f(\bar{\mathbf{z}}^{(t),\tau}) \leq -\frac{\eta}{1-\beta}\left\langle \nabla f(\bar{\mathbf{z}}^{(t),\tau}), \mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\rangle + \frac{\eta^2 L}{2(1-\beta)^2}\mathbb{E}_{t,\tau}\left\|\bar{\mathbf{g}}^{(t),\tau}\right\|^2, \tag{12}$$

其中，$\mathbb{E}_{t,\,\tau}$表示在给定所有过去随机变量的条件下，对轮次 t 中第 τ 次本地更新的随机性取条件期望。根据所述因素，对于（12）右侧：

$$\begin{aligned}
&-\left\langle \nabla f(\bar{\mathbf{z}}^{(t),\tau}), \mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\rangle \\
&\overset{(a)}{\leq} \frac{1}{2a}\left\|\nabla f(\bar{\mathbf{z}}^{(t),\tau}) - \nabla f(\bar{\mathbf{x}}^{(t),\tau})\right\|^2 - \frac{1-a}{2}\left\|\mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2 - \frac{1}{2}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})\right\|^2 + \frac{1}{2}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau}) - \mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2 \\
&\overset{(b)}{\leq} \frac{L^2}{2a}\left\|\bar{\mathbf{z}}^{(t),\tau} - \bar{\mathbf{x}}^{(t),\tau}\right\|^2 - \frac{1-a}{2}\left\|\mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2 - \frac{1}{2}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})\right\|^2 + \frac{1}{2}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau}) - \mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2,
\end{aligned} \tag{13}$$

其中，(a)由**事实2** 和**事实3**结合得到；(b)由假设**1**中的光滑性得到。由于我们假设对于初始阶段的任意 i，$\mathbf{x}_i^{(0)} = \mathbf{v}_i^{(0)} = \mathbf{0}$，根据$\mathbf{z}^{(t),\tau}$和$\mathbf{y}^{(t),\tau}$的定义，通过平均化可以证明

$$\left\|\bar{\mathbf{z}}^{(t),\tau} - \bar{\mathbf{x}}^{(t),\tau}\right\|^2 = \frac{\beta^2}{(1-\beta)^2}\|\underbrace{\bar{\mathbf{x}}^{(t),\tau} - \bar{\mathbf{v}}^{(t),\tau}}_{\bar{\mathbf{c}}^{(t),\tau}}\|^2. \tag{14}$$

应用引理 1 中的递推式 (10),

$$\begin{aligned}\left\|\bar{\mathbf{z}}^{(t),\tau} - \bar{\mathbf{x}}^{(t),\tau}\right\|^2 &\overset{(a)}{=} \left\|\bar{\mathbf{z}}^{l} - \bar{\mathbf{x}}^{l}\right\|^2 \\ &\overset{(b)}{=} \frac{\beta^2\hat{\eta}^2 s^2}{(1-\beta)^2}\left\|\sum_{j=0}^{l-1}\frac{\beta^{l-1-j}}{s}\bar{\mathbf{g}}^j\right\|^2 \\ &\overset{(c)}{\leq} \frac{\beta^2\hat{\eta}^2 s^2}{(1-\beta)^2}\sum_{j=0}^{l-1}\frac{\beta^{l-1-j}}{s}\left\|\bar{\mathbf{g}}^j\right\|^2 \\ &\overset{(d)}{\leq} \frac{\beta^2\hat{\eta}^2}{(1-\beta)^3}\sum_{j=0}^{l-1}\beta^{l-1-j}\left\|\bar{\mathbf{g}}^j\right\|^2,\end{aligned} \tag{15}$$

其中，我们省略轮次（即，t）和本地更新（即，τ）的细节，并将它们替换为更一般的术语：**迭代**（即，l）（见 a）；我们在 b 中定义 $s = \sum_{j=0}^{l-1}\beta^{l-1-j}$；（$c$）由**事实6** 和詹森不等式得出；（$d$）成立，因为 $s = \frac{1-\beta^l}{1-\beta} < \frac{1}{1-\beta}$，且 $\beta \in [0,1)$。将 (15) 代入 (13)，并令 $a = \frac{1}{2}$，得到

$$\begin{aligned}&-\left\langle\nabla f(\bar{\mathbf{z}}^{(t),\tau}), \mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\rangle \\ &\leq \frac{1}{2}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau}) - \mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2 + \frac{\beta^2\hat{\eta}^2L^2}{(1-\beta)^3}\sum_{j=0}^{l-1}\beta^{l-1-j}\left\|\bar{\mathbf{g}}^l\right\|^2 - \frac{1}{2}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})\right\|^2 - \frac{1}{4}\left\|\mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2.\end{aligned} \tag{16}$$

此外，对于 (12) 中的第二项，我们对任意向量应用已知的**事实1** 和假设2，得到

$$\mathbb{E}_{t,\tau}\left[\|\bar{\mathbf{g}}^{t,\tau}\|^2\right] \leq \left\|\mathbb{E}_{t,\tau}[\bar{\mathbf{g}}^{t,\tau}]\right\|^2 + \frac{\sigma^2}{n} \tag{17}$$

然后将 (16) 和上述不等式代入 (12),

$$\begin{aligned}\mathbb{E}_{t,\tau}f(\bar{\mathbf{z}}^{(t),\tau+1}) - f(\bar{\mathbf{z}}^{(t),\tau}) \leq{}& \frac{\tilde{\eta}^2\sigma^2L}{2n} + \left(\frac{\tilde{\eta}^2L}{2} - \frac{\tilde{\eta}}{4}\right)\left\|\mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2 - \frac{\tilde{\eta}}{2}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})\right\|^2 \\ &+ \frac{\tilde{\eta}}{2}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau}) - \mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2 + \frac{\beta^2\hat{\eta}^2L^2}{(1-\beta)^3}\sum_{j=0}^{l-1}\beta^{l-1-j}\left\|\bar{\mathbf{g}}^j\right\|^2,\end{aligned} \tag{18}$$

其中 $\tilde{\eta} = \frac{\eta}{1-\beta}$ 取全期望，并将求和范围从 $\tau = 0$ 到 $K-1$，可得

$$\begin{aligned}\mathbb{E}\left[f(\bar{\mathbf{z}}^{(t),K}) - f(\bar{\mathbf{z}}^{(t),0})\right] &= \mathbb{E}\left[f(\bar{\mathbf{z}}^{(t+1),0}) - f(\bar{\mathbf{z}}^{(t),0})\right] \\ &\leq -\frac{\tilde{\eta}}{2}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})\right\|^2 + \left(\frac{\tilde{\eta}^2L}{2} - \frac{\tilde{\eta}}{4}\right)\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2 \\ &\quad + \frac{\tilde{\eta}}{2}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau}) - \mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2 + \frac{\tilde{\eta}^2K\sigma^2L}{2n} \\ &\quad + \frac{\beta^2\hat{\eta}^2L^2}{(1-\beta)^3}\sum_{\tau=0}^{K-1}\mathbb{E}\left[\sum_{j=0}^{l-1}\beta^{l-1-j}\left\|\bar{\mathbf{g}}^j\right\|^2\right].\end{aligned} \tag{19}$$

将求和范围从 $t=0$ 到 $T-1$，并将等式两边除以 KT，

$$\begin{aligned}\frac{1}{KT}\mathbb{E}\left[f(\bar{\mathbf{z}}^{(T),0})-f(\bar{\mathbf{z}}^{(0),0})\right] \leq & -\frac{\tilde{\eta}}{2KT}\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})\right\|^2+\frac{\tilde{\eta}^2\sigma^2 L}{2n} \\ & +\left(\frac{\tilde{\eta}^2 L}{2KT}-\frac{\tilde{\eta}}{4KT}\right)\underbrace{\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2}_{T_1} \\ & +\frac{\tilde{\eta}}{2KT}\underbrace{\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})-\mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2}_{T_2} \\ & +\frac{\beta^2\hat{\eta}^2 L^2}{(1-\beta)^3 KT}\underbrace{\sum_{l=1}^{KT-1}\mathbb{E}\left[\sum_{j=0}^{l-1}\beta^{l-1-j}\left\|\bar{\mathbf{g}}^j\right\|^2\right]}_{T_3}.\end{aligned} \tag{20}$$

现在我们来界定 T_1的上界：

$$\begin{aligned}\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2 &= \sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\pm\bar{\mathbf{g}}^{(t),\tau}\right\|^2 \\ &\overset{(a)}{\leq} 2\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]-\bar{\mathbf{g}}^{(t),\tau}\right\|^2+2\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\bar{\mathbf{g}}^{(t),\tau}\right\|^2 \\ &\overset{(b)}{\leq} \frac{2KT\sigma^2}{n}+2\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\bar{\mathbf{g}}^{(t),\tau}\right\|^2,\end{aligned} \tag{21}$$

其中，(a) 是通过令 $a=1$ 并应用 **事实4** 得出的；(b) 则由假设2得出。然后我们估计 T_2:

$$\begin{aligned}\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})-\mathbb{E}_{t,\tau}\left[\bar{\mathbf{g}}^{(t),\tau}\right]\right\|^2 &= \sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f(\bar{\mathbf{x}}^{(t),\tau})-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t),\tau})\right\|^2 \\ &\overset{(a)}{\leq} \frac{1}{n^2}\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})-\nabla f_i(\mathbf{x}_i^{(t),\tau})\pm\nabla f_i(\bar{\mathbf{x}}^{(t),\tau})\right\|^2 \\ &\overset{(b)}{\leq} \frac{2}{n^2}\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\sum_{i=1}^{n}\left(\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})-\nabla f_i(\bar{\mathbf{x}}^{(t),\tau})\right\|^2+\mathbb{E}\left\|\nabla f_i(\bar{\mathbf{x}}^{(t),\tau})-\nabla f_i(\mathbf{x}_i^{(t),\tau})\right\|^2\right) \\ &\overset{(c)}{\leq} \frac{2KT\zeta^2}{n}+\frac{2L^2}{n^2}\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\bar{\mathbf{x}}^{(t),\tau}-\mathbf{x}_i^{(t),\tau}\right\|^2,\end{aligned} \tag{22}$$

其中，(a) 由 $\|\cdot\|^2$ 的复杂性和詹森不等式得出；(b) 由**事实4** 得出；(c) 是通过在假设2中应用$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{x})-\nabla f(\mathbf{x})\right\|^2\leq\zeta^2$ 得出的。最后，我们对项 T_3 进行界定：

$$\begin{aligned}\sum_{l=1}^{KT-1}\mathbb{E}\left[\sum_{j=0}^{l-1}\beta^{l-1-j}\left\|\bar{\mathbf{g}}^j\right\|^2\right] &\leq \sum_{j=0}^{KT-2}\mathbb{E}\left[\left\|\bar{\mathbf{g}}^j\right\|^2\sum_{u=j+1}^{KT-1}\beta^{u-1-j}\right] \\ &\overset{(a)}{\leq} \frac{1}{1-\beta}\sum_{j=0}^{KT-2}\mathbb{E}\left\|\bar{\mathbf{g}}^j\right\|^2 \\ &\leq \frac{1}{1-\beta}\sum_{j=0}^{KT-1}\mathbb{E}\left\|\bar{\mathbf{g}}^j\right\|^2,\end{aligned} \tag{23}$$

其中，(a) 由注意到 $\sum_{u=j+1}^{TK-1}\beta^{u-1-j}=\frac{1-\beta^{TK-1-j}}{1-\beta}\leq\frac{1}{1-\beta}$ 得出。将 (21)、(22) 和 (23) 代入 (20)，得到

$$\begin{aligned}\frac{1}{KT}\mathbb{E}\left[f(\bar{\mathbf{z}}^{(T),0})-f(\bar{\mathbf{z}}^{(0),0})\right]\leq&-\frac{\tilde{\eta}}{2KT}\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t),\tau})\right\|^2+\frac{\tilde{\eta}^2\sigma^2L}{2n}+\left(\frac{\tilde{\eta}^2L}{KT}-\frac{\tilde{\eta}}{2KT}\right)\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\bar{\mathbf{g}}^{(t),\tau}\right\|^2\\&+\frac{\tilde{\eta}L^2}{KTn^2}\sum_{t=0}^{T-1}\sum_{\tau=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\bar{\mathbf{x}}^{(t),\tau}-\mathbf{x}_i^{(t),\tau}\right\|^2+\frac{\beta^2\hat{\eta}^2L^2}{(1-\beta)^4KT}\sum_{j=0}^{KT-1}\mathbb{E}\left\|\bar{\mathbf{g}}^{j}\right\|^2\\&+\frac{\tilde{\eta}\zeta^2}{n}+\frac{\tilde{\eta}\sigma^2}{2n}\left(3\tilde{\eta}L-1\right).\end{aligned}\tag{24}$$

利用 $\sum_{i=1}^{n}\|\mathbf{a}_i\|^2=\|\mathbf{A}\|_F^2$（其中 $\mathbf{A}:=[\mathbf{a}_1,\cdots,\mathbf{a}_n]$），我们现在尝试界定各节点参数与其平均值之间的共识误差。首先，我们以矩阵形式重述式（7）的更新方案，不考虑轮次 t 和本地更新τ，并将 l 记为更新迭代的索引：

$$\mathbf{X}^{(l+1)}=\mathbf{X}^{(l)}-\eta\mathbf{G}^{(l)}+\beta\left(\mathbf{X}^{(l)}-\alpha\eta\mathbf{G}^{(l)}-\mathbf{X}^{(l-1)}+\alpha\eta\mathbf{G}^{(l-1)}\right).\tag{25}$$

对于在所有节点之间进行模型平均的平均参数，由于$\mathbf{W}$ 是双随机的，我们也可以简化更新，具体如下：

$$\bar{\mathbf{x}}^{(l+1)}=\bar{\mathbf{x}}^{(l)}-\eta\bar{\mathbf{g}}^{(l)}+\beta\left(\bar{\mathbf{x}}^{(l)}-\alpha\eta\bar{\mathbf{g}}^{(l)}-\bar{\mathbf{x}}^{(l-1)}+\alpha\eta\bar{\mathbf{g}}^{(l-1)}\right).\tag{26}$$

根据上述两个方程，我们有

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{(l+1)}-\bar{\mathbf{X}}^{(l+1)}\right\|_F^2=&\frac{1}{n}\mathbb{E}\Big\|\mathbf{X}^{(l)}-\eta\mathbf{G}^{(l)}+\beta\left(\mathbf{X}^{(l)}-\alpha\eta\mathbf{G}^{(l)}-\mathbf{X}^{(l-1)}+\alpha\eta\mathbf{G}^{(l-1)}\right)\\&-\left(\bar{\mathbf{X}}^{(l)}-\eta\bar{\mathbf{G}}^{(l)}+\beta\left(\bar{\mathbf{X}}^{(l)}-\alpha\eta\bar{\mathbf{G}}^{(l)}-\bar{\mathbf{X}}^{(l-1)}+\alpha\eta\bar{\mathbf{G}}^{(l-1)}\right)\right)\Big\|_F^2\\\overset{(a)}{\leq}&\frac{1-\rho}{n}\mathbb{E}\Big\|(1+\beta)\left(\mathbf{X}^{(l)}-\bar{\mathbf{X}}^{(l)}\right)-\beta\left(\mathbf{X}^{(l-1)}-\bar{\mathbf{X}}^{(l-1)}\right)-(1+\alpha\beta)\,\eta\left(\mathbf{G}^{(l)}-\bar{\mathbf{G}}^{(l)}\right)\\&+\alpha\beta\eta\left(\mathbf{G}^{(l-1)}-\bar{\mathbf{G}}^{(l-1)}\right)\Big\|_F^2\\\overset{(b)}{\leq}&\frac{1-\rho}{n}\mathbb{E}\Big\|(1+\beta)\left(\mathbf{X}^{(l)}-\bar{\mathbf{X}}^{(l)}\right)-\beta\left(\mathbf{X}^{(l-1)}-\bar{\mathbf{X}}^{(l-1)}\right)-(1+\alpha\beta)\,\eta\left(\mathbb{E}\left[\mathbf{G}^{(l)}\right]-\mathbb{E}\left[\bar{\mathbf{G}}^{(l)}\right]\right)\\&+\alpha\beta\eta\left(\mathbb{E}\left[\mathbf{G}^{(l-1)}\right]-\mathbb{E}\left[\bar{\mathbf{G}}^{(l-1)}\right]\right)\Big\|_F^2+\underbrace{4\left(1-\rho\right)\left(1+2\alpha\beta+2\alpha^2\beta^2\right)\eta^2\sigma^2}_{:=\Delta}\\\overset{(c)}{\leq}&\frac{2(1-\rho)}{n}\mathbb{E}\Big\|(1+\beta)\left(\mathbf{X}^{(l)}-\bar{\mathbf{X}}^{(l)}\right)-(1+\alpha\beta)\,\eta\left(\mathbb{E}\left[\mathbf{G}^{(l)}\right]-\mathbb{E}\left[\bar{\mathbf{G}}^{(l)}\right]\right)\Big\|_F^2\\&+\frac{2(1-\rho)}{n}\mathbb{E}\Big\|\beta\left(\mathbf{X}^{(l-1)}-\bar{\mathbf{X}}^{(l-1)}\right)-\alpha\beta\eta\left(\mathbb{E}\left[\mathbf{G}^{(l-1)}\right]-\mathbb{E}\left[\bar{\mathbf{G}}^{(l-1)}\right]\right)\Big\|_F^2+\Delta,\end{aligned}\tag{27}$$

其中，(a) 由应用假设4得到；(b) 成立是因为我们加入了 $\mathbf{G}$ 和 $\bar{\mathbf{G}}$的期望项，从而 $\mathbf{G}-\bar{\mathbf{G}}=(\mathbf{G}-\mathbb{E}[\mathbf{G}])-(\bar{\mathbf{G}}-\mathbb{E}[\bar{\mathbf{G}}])+(\mathbb{E}[\mathbf{G}]-\mathbb{E}[\bar{\mathbf{G}}])$，其满足将常数 Δ 推广后的假设2条件；(c) 通过令 $a=1$，由**事实4** 得到。这里我们使用了矩阵 $\mathbf{W}$ 的压缩性和杨氏不等式。我们可以

进一步推导如下

$$
\begin{aligned}
&\frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{(l+1)}-\bar{\mathbf{X}}^{(l+1)}\right\|_F^2 \\
&\overset{(a)}{\leq} \frac{2(1-\rho)(1+\beta)^2}{n}\left(1+\frac{\rho}{2}\right)\mathbb{E}\left\|\mathbf{X}^{(l)}-\bar{\mathbf{X}}^{(l)}\right\|_F^2+\frac{2(1-\rho)(1+\alpha\beta)^2\eta^2}{n}\left(1+\frac{2}{\rho}\right)\mathbb{E}\left\|\mathbb{E}\left[\mathbf{G}^{(l)}\right]-\mathbb{E}\left[\bar{\mathbf{G}}^{(l)}\right]\right\|_F^2 \\
&\quad+\frac{2(1-\rho)\beta^2}{n}\left(1+\frac{\rho}{2}\right)\mathbb{E}\left\|\mathbf{X}^{(l-1)}-\bar{\mathbf{X}}^{(l-1)}\right\|_F^2+\frac{2(1-\rho)}{n}\alpha^2\beta^2\eta^2\left(1+\frac{2}{\rho}\right)\mathbb{E}\left\|\mathbb{E}\left[\mathbf{G}^{(l-1)}\right]-\mathbb{E}\left[\bar{\mathbf{G}}^{(l-1)}\right]\right\|_F^2+\Delta \\
&\leq \frac{2(1-\rho)(1+\beta)^2}{n}\left(1+\frac{\rho}{2}\right)\mathbb{E}\left\|\mathbf{X}^{(l)}-\bar{\mathbf{X}}^{(l)}\right\|_F^2+\frac{2(1-\rho)(1+\alpha\beta)^2\eta^2}{n}\left(1+\frac{2}{\rho}\right)\mathbb{E}\left\|\mathbb{E}\left[\mathbf{G}^{(l)}\right]-\nabla f(\bar{\mathbf{X}}^{l})\right\|_F^2 \\
&\quad+\frac{2(1-\rho)\beta^2}{n}\left(1+\frac{\rho}{2}\right)\mathbb{E}\left\|\mathbf{X}^{(l-1)}-\bar{\mathbf{X}}^{(l-1)}\right\|_F^2+\frac{2(1-\rho)}{n}\alpha^2\beta^2\eta^2\left(1+\frac{2}{\rho}\right)\mathbb{E}\left\|\mathbb{E}\left[\mathbf{G}^{(l-1)}\right]-\nabla f(\bar{\mathbf{X}}^{l-1})\right\|_F^2+\Delta \\
&= \frac{2(1-\rho)(1+\beta)^2}{n}\left(1+\frac{\rho}{2}\right)\mathbb{E}\left\|\mathbf{X}^{(l)}-\bar{\mathbf{X}}^{(l)}\right\|_F^2+\frac{2(1-\rho)(1+\alpha\beta)^2\eta^2}{n}\left(1+\frac{2}{\rho}\right)\sum_{i=1}^n\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(l)})\pm\nabla f_i(\bar{\mathbf{x}}^{(l)})-\nabla f(\bar{\mathbf{x}}^{(l)})\right\|^2+\Delta \\
&\quad+\frac{2(1-\rho)\beta^2}{n}\left(1+\frac{\rho}{2}\right)\mathbb{E}\left\|\mathbf{X}^{(l-1)}-\bar{\mathbf{X}}^{(l-1)}\right\|_F^2+\frac{2(1-\rho)}{n}\alpha^2\beta^2\eta^2\left(1+\frac{2}{\rho}\right)\sum_{i=1}^n\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(l-1)})\pm\nabla f_i(\bar{\mathbf{x}}^{(l-1)})-\nabla f(\bar{\mathbf{x}}^{(l-1)})\right\|^2 \\
&\overset{(b)}{\leq} \frac{2(1-\frac{\rho}{2})}{n}(1+\beta)^2\left\|\mathbf{X}^{(l)}-\bar{\mathbf{X}}^{(l)}\right\|_F^2+\frac{12\eta^2L^2}{\rho n}(1+\alpha\beta)^2\sum_{i=1}^n\mathbb{E}\left\|\mathbf{x}_i^{(l)}-\bar{\mathbf{x}}^{(l)}\right\|^2+\frac{12\eta^2(1+\alpha\beta)^2\zeta^2}{\rho} \\
&\quad+\frac{2(1-\frac{\rho}{2})\beta^2}{n}\mathbb{E}\left\|\mathbf{X}^{(l-1)}-\bar{\mathbf{X}}^{(l-1)}\right\|_F^2+\frac{12\eta^2L^2}{\rho n}\alpha^2\beta^2\sum_{i=1}^n\mathbb{E}\left\|\mathbf{x}_i^{(l-1)}-\bar{\mathbf{x}}^{(l-1)}\right\|^2+\frac{12\eta^2\alpha^2\beta^2\zeta^2}{\rho}+\Delta,
\end{aligned}
\tag{28}
$$

其中，(a)由应用**事实4**和集合 $a=\frac{\rho}{2}$ 得到；(b)是因为正的 $\rho\leq 1$，并且也用到了**事实4**和假设2。选择学习率 $\eta\leq\frac{\rho}{5L}$ 以确保 $6\eta^2L^2\leq\frac{\rho^2}{4}$ 后，由于 $a\geq 0$，可分为两种情况

- **情况一：** $\alpha\in[0,1)$，我们定义 $\hat{\beta}=1+\beta$，于是

$$
\frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{(l+1)}-\bar{\mathbf{X}}^{(l+1)}\right\|_F^2\leq\frac{2\hat{\beta}^2\left(1-\frac{\rho}{4}\right)}{n}\mathbb{E}\left\|\mathbf{X}^{(l)}-\bar{\mathbf{X}}^{(l)}\right\|_F^2+\frac{2\hat{\beta}^2\left(1-\frac{\rho}{4}\right)}{n}\mathbb{E}\left\|\mathbf{X}^{(l-1)}-\bar{\mathbf{X}}^{(l-1)}\right\|_F^2+\frac{24\eta^2\hat{\beta}^2\zeta^2}{\rho}+\Delta.
$$

- **情况二：** $\alpha\in[1,\infty)$，我们定义 $\hat{\beta}=1+\alpha\beta$，只需将 β 项替换为 $\hat{\beta}$。

这里我们记 $\beta_0=\max\left\{\hat{\beta},\beta\right\}$。由于 $\mathbf{x}_i^{(0)}=\bar{\mathbf{x}}^{(0)}=\mathbf{x}_0$，可得 $\frac{1}{n}\mathbb{E}\|\mathbf{X}^{(0)}-\bar{\mathbf{X}}^{(0)}\|_F^2=\mathbf{0}$。此外，我们可以容易地得到 $\frac{1}{n}\mathbb{E}\|\mathbf{X}^{(1)}-\bar{\mathbf{X}}^{(1)}\|_F^2=24\eta^2\beta_0^2\zeta^2/\rho+\Delta$。注意，当 $l\geq 1$时，$\frac{1}{n}\mathbb{E}\|\mathbf{X}^{(l)}-\bar{\mathbf{X}}^{(l)}\|_F^2\leq\frac{C_1l}{(1-Q_1)}$，其中我们记 $Q_1=2\beta_0^2(1-\frac{\rho}{4})$，以确保 $Q_1<1$，且 $C_1=24\eta^2\beta_0^2\zeta^2/\rho+\Delta$。将上述不等式代入(24)，对于任意 $\bar{\mathbf{g}}$，根据假设3有 $\mathbb{E}\left\|\bar{\mathbf{g}}\right\|^2\leq\frac{G^2}{n}$，并且根据定义有 $\bar{\mathbf{z}}^{(0)}=\bar{\mathbf{x}}^{(0)}=\mathbf{x}_0$。整理各项后即可得到定理1。

C GT-DSUM的证明

Next，我们将在非凸性条件下对 GT-DSUM 给出严格证明。这里我们考虑一种特殊情况，即 $K=1$ 采用固定的共识矩阵。然后，我们构造算法2的矩阵形式如下：

$$
\begin{aligned}
\mathbf{M}^{(t)}&=\lambda\mathbf{G}^{(t)}+(1-\lambda)\mathbf{Y}^{(t)} \\
\mathbf{X}^{(t+1)}&=\mathbf{W}\left((1+\beta)\mathbf{X}^{(t)}-\beta\mathbf{W}\mathbf{X}^{(t-1)}-(1+\alpha\beta)\eta\mathbf{M}^{(t)}+\alpha\beta\eta\mathbf{W}\mathbf{M}^{(t-1)}\right) \\
\mathbf{Y}^{(t+1)}&=\mathbf{W}\left(\mathbf{Y}^{(t)}+\frac{2\mathbf{X}^{(t)}-\mathbf{X}^{(t+1)}-\mathbf{X}^{(t-1)}}{\eta}\right).
\end{aligned}
\tag{29}
$$

证明概要。 我们尝试界定工作节点的参数与其平均值之间的共识距离（引理2）。在这一步中，我们执行传播步骤，使各工作节点的参数彼此更加接近。此外，我们还执行额外的梯度跟踪（引理3）及其累积步骤（引理4），这些步骤会使它们之间的距离增大。之后，我们可以直接将引理4应用于式（45）中的单步更新过程。

引理2 (共识距离变化)。 在第 *2* 节给定的假设下，令 $\beta_0 = \text{最大值}\{1+\alpha\beta, 1+\beta\}$，并且由式（*29*）使用学习率生成的更新规则满足 $\eta \leq \frac{\rho}{12\lambda L}$ 当 $t \geq 1$时，

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t+1)}\right\|_F^2 \leq & \frac{2\left(1-\frac{\rho}{4}\right)\beta_0^2}{n}\mathbb{E}\left\|\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2+\frac{2\left(1-\frac{\rho}{4}\right)\left(\beta_0-1\right)^2}{n}\mathbb{E}\left\|\mathbf{X}^{(t-1)}-\bar{\mathbf{X}}^{(t-1)}\right\|_F^2 \\ & +\frac{\rho\left(1-\frac{\rho}{2}\right)(1+\alpha\beta)^2(1-\lambda)^2}{12\lambda^2L^2n}\mathbb{E}\left\|\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right\|_F^2+\frac{\rho\left(1-\frac{\rho}{2}\right)\alpha^2\beta^2(1-\lambda)^2}{12\lambda^2L^2n}\mathbb{E}\left\|\mathbf{Y}^{(t-1)}-\bar{\mathbf{Y}}^{(t-1)}\right\|_F^2 \\ & +\frac{\rho\left(1-\frac{\rho}{2}\right)}{nL^2}(1+\alpha\beta)^2\left(\sigma^2+\zeta^2\right).\end{aligned}$$

证明。 由(29)式出发，且所有共识矩阵均满足命题 2，我们有

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t+1)}\right\|_F^2 \overset{(a)}{\leq} & \frac{1-\rho}{n}\mathbb{E}\left\|(1+\beta)\left(\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right)-\beta\left(\mathbf{W}\mathbf{X}^{(t-1)}-\bar{\mathbf{X}}^{(t-1)}\right)-(1+\alpha\beta)\eta\left(\mathbf{M}^{(t)}-\bar{\mathbf{M}}^{(t)}\right)\right. \\ & \left.+\alpha\beta\eta\left(\mathbf{W}\mathbf{M}^{(t-1)}-\bar{\mathbf{M}}^{(t-1)}\right)\right\|_F^2 \\ \overset{(b)}{\leq} & \frac{1-\rho}{n}\left(1+\frac{\rho}{2}\right)\mathbb{E}\left\|(1+\beta)\left(\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right)-\beta\left(\mathbf{W}\mathbf{X}^{(t-1)}-\bar{\mathbf{X}}^{(t-1)}\right)\right\|_F^2 \\ & +\frac{1-\rho}{n}\left(1+\frac{2}{\rho}\right)\mathbb{E}\left\|(1+\alpha\beta)\eta\left(\mathbf{M}^{(t)}-\bar{\mathbf{M}}^{(t)}\right)-\alpha\beta\eta\left(\mathbf{W}\mathbf{X}^{(t-1)}-\bar{\mathbf{X}}^{(t-1)}\right)\right\|_F^2 \\ \overset{(c)}{\leq} & \frac{2(1-\rho)(1+\beta)^2}{n}\left(1+\frac{\rho}{2}\right)\mathbb{E}\left\|\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2+\frac{2(1-\rho)^2\beta^2}{n}\left(1+\frac{2}{\rho}\right)\mathbb{E}\left\|\mathbf{X}^{(t-1)}-\bar{\mathbf{X}}^{(t-1)}\right\|_F^2 \\ & +\frac{2(1-\rho)(1+\alpha\beta)^2\eta^2}{n}\left(1+\frac{\rho}{2}\right)\mathbb{E}\left\|\mathbf{M}^{(t)}-\bar{\mathbf{M}}^{(t)}\right\|_F^2 \\ & +\frac{2(1-\rho)^2\alpha^2\beta^2\eta^2}{n}\left(1+\frac{2}{\rho}\right)\mathbb{E}\left\|\mathbf{M}^{(t-1)}-\bar{\mathbf{M}}^{(t-1)}\right\|_F^2,\end{aligned} \tag{30}$$

其中，(*a*)由应用假设4得到；通过分别选择 $\rho/2$和1，(*b*)、(*c*)由**事实4** 得到。接下来，我们尝试界定$\mathbf{M}^{(t)}$ 与 $\bar{\mathbf{M}}^{(t)}$之间的距离，

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left\|\mathbf{M}^{(t)}-\bar{\mathbf{M}}^{(t)}\right\|_F^2 \overset{(a)}{=} & \frac{1}{n}\mathbb{E}\left\|\lambda\left(\mathbf{G}^{(t)}-\bar{\mathbf{G}}^{(t)}\right)+(1-\lambda)\left(\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right)\right\|_F^2 \\ \overset{(b)}{\leq} & \frac{2\lambda^2}{n}\mathbb{E}\left\|\mathbf{G}^{(t)}-\bar{\mathbf{G}}^{(t)}\right\|_F^2+\frac{2(1-\lambda)^2}{n}\mathbb{E}\left\|\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right\|_F^2 \\ = & \frac{2\lambda^2}{n}\mathbb{E}\left\|\mathbf{G}^{(t)}\pm\mathbb{E}_t\left[\mathbf{G}^{(t)}\right]-\bar{\mathbf{G}}^{(t)}\pm\mathbb{E}_t\left[\bar{\mathbf{G}}^{(t)}\right]\right\|_F^2+\frac{2(1-\lambda)^2}{n}\mathbb{E}\left\|\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right\|_F^2 \\ \overset{(c)}{\leq} & 12\lambda^2\sigma^2+\frac{6\lambda^2}{n}\mathbb{E}\left\|\mathbb{E}_t\left[\mathbf{G}^{(t)}\right]-\mathbb{E}_t\left[\bar{\mathbf{G}}^{(t)}\right]\right\|_F^2+\frac{2(1-\lambda)^2}{n}\mathbb{E}\left\|\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right\|_F^2 \\ \leq & \frac{6\lambda^2}{n}\sum_{i=1}^n\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(t)})\pm\nabla f_i(\bar{\mathbf{x}}^{(t)})-\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2+12\lambda^2\sigma^2+\frac{2(1-\lambda)^2}{n}\mathbb{E}\left\|\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right\|_F^2 \\ \overset{(d)}{\leq} & \frac{12\lambda^2L^2}{n}\sum_{i=1}^n\mathbb{E}\left\|\mathbf{x}_i^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|^2+12\lambda^2\left(\sigma^2+\zeta^2\right)+\frac{2(1-\lambda)^2}{n}\mathbb{E}\left\|\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right\|_F^2,\end{aligned} \tag{31}$$

其中，(*a*)由(29)式得到；(*b*)由应用**事实4**得到；(*c*)由**事实 5** 并取向量集合$\{\mathbf{G}-\mathbb{E}[\mathbf{G}], \mathbb{E}[\bar{\mathbf{G}}]-\bar{\mathbf{G}}\ \mathbb{E}[\mathbf{G}]-\mathbb{E}[\bar{\mathbf{G}}]\}$得到；(*d*)由**事实4** 和假设2得到。由于正标量 $\rho \leq 1$，在 $\eta \leq \frac{\rho}{12\lambda L}$ 的条件下，将(31)式代入(30)式可确保 $36\lambda^2L^2\eta^2 \leq \rho^2/4$，从而完成证明。 □

引理3 (梯度跟踪器距离变化)。 根据第 *2* 节中的假设，当 $t \geq 1$ 时，令学习率满足$\eta \leq$最小值 $\left\{\frac{3+\beta}{2\sqrt{3\lambda}L(1+\alpha\beta)}, \frac{1+\beta}{2\sqrt{3\lambda}L\alpha\beta}\right\}$，并由超参数假设可得$(1+\alpha\beta)(1-\lambda) \leq \frac{1}{2\sqrt{2}}$，这

得到

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{Y}^{(t+1)}-\bar{\mathbf{Y}}^{(t+1)}\right\|_F^2 \le \frac{16(1-\rho)(3+\beta)^2}{n\eta^2}\mathbb{E}\left\|\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2 + \frac{16(1-\rho)(1+\beta)^2}{n\eta^2}\mathbb{E}\left\|\mathbf{X}^{(t-1)}-\bar{\mathbf{X}}^{(t-1)}\right\|_F^2 + \frac{4(1-\rho)}{n}\mathbb{E}\left\|\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right\|_F^2 + \frac{2(1-\rho)}{n}\mathbb{E}\left\|\mathbf{Y}^{(t-1)}-\bar{\mathbf{Y}}^{(t-1)}\right\|_F^2 + \frac{192\lambda^2}{n}(1-\rho)(1+\alpha\beta)^2(\sigma^2+\zeta^2).$$

证明。 根据式（29）中的更新方案，我们可以得到

$$\mathbf{X}^{(t)}-\mathbf{X}^{(t+1)} = (\mathbf{I}-(1+\beta)\,\mathbf{W})\,\mathbf{X}^{(t)} + \beta\mathbf{W}^2\mathbf{X}^{(t-1)} + (1+\alpha\beta)\,\eta\mathbf{W}\mathbf{M}^{(t)} - \alpha\beta\eta\mathbf{W}^2\mathbf{M}^{(t-1)}, \tag{32}$$

and

$$2\mathbf{X}^{(t)}-\mathbf{X}^{(t+1)}-\mathbf{X}^{(t-1)} = (2\mathbf{I}-(1+\beta)\,\mathbf{W})\,\mathbf{X}^{(t)} + \left(\beta\mathbf{W}^2-\mathbf{I}\right)\mathbf{X}^{(t-1)} + (1+\alpha\beta)\,\eta\mathbf{W}\mathbf{M}^{(t)} - \alpha\beta\eta\mathbf{W}^2\mathbf{M}^{(t-1)}. \tag{33}$$

然后，根据式（33）和式（29），我们有

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{Y}^{(t+1)}-\bar{\mathbf{Y}}^{(t+1)}\right\|_F^2 \le \frac{2(1-\rho)}{n}\mathbb{E}\left\|\frac{2\mathbf{I}-(1+\beta)\mathbf{W}}{\eta}\left(\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right) + \frac{\beta\mathbf{W}^2-\mathbf{I}}{\eta}\left(\mathbf{X}^{(t-1)}-\bar{\mathbf{X}}^{(t-1)}\right) + (1+\alpha\beta)\mathbf{W}\left(\mathbf{M}^{(t)}-\bar{\mathbf{M}}^{(t)}\right) - \alpha\beta\mathbf{W}^2\left(\mathbf{M}^{(t-1)}-\bar{\mathbf{M}}^{(t-1)}\right)\right\|_F^2 + \frac{2(1-\rho)}{n}\mathbb{E}\left\|\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right\|_F^2, \tag{34}$$

由假设4和**事实4**，该不等式成立。由于$\mathbf{W} \prec \mathbf{I}$， $-\mathbf{I} \prec \mathbf{W}$，且两个双随机矩阵的乘积仍为双随机矩阵，因此我们有 $2\mathbf{I}-(1+\beta)\mathbf{W} \prec (3+\beta)\mathbf{I}$； $\beta\mathbf{W}^2-\mathbf{I} \prec (\beta+1)\mathbf{W}^2 \prec (\beta+1)\mathbf{I}$，我们可以继续

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{Y}^{(t+1)}-\bar{\mathbf{Y}}^{(t+1)}\right\|_F^2 \le \frac{8(1-\rho)(3+\beta)^2}{n\eta^2}\mathbb{E}\left\|\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2 + \frac{8(1-\rho)(1+\beta)^2}{n\eta^2}\mathbf{E}\left\|\mathbf{X}^{(t-1)}-\bar{\mathbf{X}}^{(t-1)}\right\|_F^2 + \frac{8(1-\rho)(1+\alpha\beta)^2}{n}\mathbb{E}\left\|\mathbf{M}^{(t)}-\bar{\mathbf{M}}^{(t)}\right\|_F^2 + \frac{8(1-\rho)\alpha^2\beta^2}{n}\mathbf{E}\left\|\mathbf{M}^{(t-1)}-\bar{\mathbf{M}}^{(t-1)}\right\|_F^2 + \frac{2(1-\rho)}{n}\mathbb{E}\left\|\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right\|_F^2, \tag{35}$$

其中，不等式由对同维矩阵应用基本不等式 $\left\|\sum_{i=1}^j \mathbf{A}_i\right\|_F^2 \le j\sum_{i=1}^j \|\mathbf{A}_i\|_F^2$ 并满足 $j=4$ 得出。为便于表述，在条件（i）$\eta \le \frac{3+\beta}{2\sqrt{3}\lambda L(1+\alpha\beta)}$；（$ii$）$\eta \le \frac{1+\beta}{2\sqrt{3}\lambda L\alpha\beta}$；（$iii$）$(1+\alpha\beta)(1-\lambda) \le \frac{1}{2\sqrt{2}}$ 下，将式（31）代入式（35）。□

引理4 (距离步改进**)**。 当 $t \ge 1$ 时，使用学习率 $\eta \ge \max\left\{\frac{2\sqrt{2}(3+\beta)}{\beta_0}, 2\sqrt{2}\left(\frac{1+\beta}{\beta_0-1}\right)\right\}$ 且 $\rho \le \frac{48\lambda^2L^2}{(1-\lambda)^2}$，满足

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t+1)}\right\|_F^2 + \frac{1}{n}\mathbb{E}\left\|\mathbf{Y}^{(t+1)}-\bar{\mathbf{Y}}^{(t+1)}\right\|_F^2 \le \frac{4\beta_0^2\left(1-\frac{\rho}{4}\right)}{n}\mathbb{E}\left\|\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2 + \frac{4\beta_0^2\left(1-\frac{\rho}{4}\right)}{n}\mathbb{E}\left\|\mathbf{X}^{(t-1)}-\bar{\mathbf{X}}^{(t-1)}\right\|_F^2 + \frac{8(1-\rho)(1+\alpha\beta)^2}{n}\mathbb{E}\left\|\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right\|_F^2 + \frac{8(1-\rho)(1+\alpha\beta)^2}{n}\mathbb{E}\left\|\mathbf{Y}^{(t-1)}-\bar{\mathbf{Y}}^{(t-1)}\right\|_F^2 + \frac{\beta_0^2\left(1-\frac{\rho}{2}\right)}{nL^2}(\sigma^2+\zeta^2)(192\lambda^2L^2+\rho).$$

证明。 将引理2和引理3的结果相加，并根据我们对学习率 η 以及超参数 ρ, λ, L 的假设，即可得到所述结果。□

现在我们给出算法2中 GT-DSUM 的收敛结果。与定理1的证明过程类似，利用 f 的光滑性，

$$\mathbb{E}f(\bar{\mathbf{z}}^{(t+1)}) \le \mathbb{E}f(\bar{\mathbf{z}}^{(t)}) + \tilde{\eta}L^2\underbrace{\mathbb{E}\left\|\bar{\mathbf{z}}^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|^2}_{T_1} + \underbrace{\left(\frac{\tilde{\eta}^2L}{2}-\frac{\tilde{\eta}}{4}\right)\mathbb{E}\left\|\bar{\mathbf{m}}^{(t)}\right\|^2}_{T_2} - \frac{\tilde{\eta}}{2}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{\tilde{\eta}}{2}\underbrace{\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})-\bar{\mathbf{m}}^{(t)}\right\|^2}_{T_3}, \tag{36}$$

其中 $\tilde{\eta}=\frac{\eta}{1-\beta^{\circ}}$ 对于项 T_1，我们采用与式 (15) 相同的推导过程，这同样适用于没有多次本地更新的情况，这表明

$$\left\|\bar{\mathbf{z}}^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|^2 \leq \frac{\beta^2\hat{\eta}^2}{(1-\beta)^3}\sum_{l=0}^{t-1}\beta^{t-1-l}\left\|\bar{\mathbf{m}}^{(l)}\right\|^2, \tag{37}$$

其中 $\hat{}\eta=((1-\beta)\alpha-1)\eta_{\circ}$ 我们还可以得到

$$\begin{aligned}
\mathbb{E}\left\|\bar{\mathbf{m}}^{(t)}\right\|^2 &\overset{(a)}{=} \mathbb{E}\left\|\lambda\bar{\mathbf{g}}^{(t)}+(1-\lambda)\,\bar{\mathbf{y}}^{(t)}\right\|^2 \\
&\overset{(b)}{\leq} \mathbb{E}\left\|\bar{\mathbf{g}}^{(t)}+\bar{\mathbf{y}}^{(t)}\right\|^2 \\
&= \mathbb{E}\left\|\bar{\mathbf{g}}^{(t)}\pm\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})+\bar{\mathbf{y}}^{(t)}\right\|^2 \\
&\overset{(c)}{\leq} \underbrace{2\mathbb{E}\left\|\bar{\mathbf{g}}^{(t)}+\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2}_{T_4}+\underbrace{2\mathbb{E}\left\|\bar{\mathbf{y}}^{(t)}-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2}_{T_5},
\end{aligned} \tag{38}$$

其中 (a) 由 (29) 得出，　　通过模型平均；我们省略了 (b) 中的系数；(c) 由应用 **事实4** 得出。对于 T_4　　，

$$\begin{aligned}
T_4 &\overset{(a)}{=} \mathbb{E}\left\|\frac{2}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})+\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})-\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})\right\|^2 \\
&\overset{(b)}{\leq} \frac{8}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})\right\|^2+\frac{2}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(t)})-F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})\right\|^2 \\
&\overset{(c)}{\leq} \frac{8G^2}{n}+\frac{2\sigma^2}{n},
\end{aligned} \tag{39}$$

其中 (a) 根据 $\bar{}\mathbf{g}_i$ 的定义得出；(b) 由詹森不等式和 **事实4** 得出；而 (c) 则是因为假设2和3。估计 T_5,

$$\begin{aligned}
T_5 &\overset{(a)}{=} \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i^{(t)}-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{y}_i^{(t)}-\frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\mathbf{x}_i^{(t)})\right)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\nabla f_j(\mathbf{x}_i^{(t)})-\nabla f(\mathbf{x}_i^{(t)})\right)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\nabla f(\mathbf{x}_i^{(t)})-\nabla f_i(\mathbf{x}_i^{(t)})\right)\right\|^2 \\
&\overset{(b)}{\leq} \frac{3}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{y}_i^{(t)}-\frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\mathbf{x}_i^{(t)})\right\|^2+\frac{3}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left\|\nabla f_j(\mathbf{x}_i^{(t)})-\nabla f(\mathbf{x}_i^{(t)})\right\|^2+\frac{3}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left\|\nabla f(\mathbf{x}_i^{(t)})-\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 \\
&\overset{(c)}{\leq} 3\epsilon^2+\frac{6\zeta^2}{n^2},
\end{aligned} \tag{40}$$

其中，(a)成立是因为 $\bar{}\mathbf{y}^{(t)}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i^{(t)}$；$(b)$是应用**事实 5**得到的；$(c)$是根据命题1和假设2得到的。将(37)、(38)、(39)和(40)代入(36)中的 T_1，得到

$$T_1 \leq \frac{\beta^2\hat{\eta}^2}{(1-\beta)^3}\sum_{l=0}^{t-1}\beta^{t-1-l}\left(\frac{16G^2}{n}+\frac{4\sigma^2}{n}+\frac{12\zeta^2}{n^2}+6\epsilon^2\right). \tag{41}$$

假设 $\eta \leq \frac{1-\beta}{2L}$ 保证 $\tilde{\eta}^2 L - \tilde{\eta} < 0$，我们可以省略(36)中的 T_2。接着，我们估计(36)中 T_3 的上界，

$$
\begin{aligned}
\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)}) - \bar{\mathbf{m}}^{(t)}\right\|^2 &\overset{(a)}{=} \mathbb{E}\left\|\lambda\left(\nabla f(\bar{\mathbf{x}}^{(t)}) - \bar{\mathbf{g}}^{(t)}\right) + (1-\lambda)\left(\nabla f(\bar{\mathbf{x}}^{(t)}) - \bar{\mathbf{y}}^{(t)}\right)\right\|^2 \\
&\overset{(b)}{\leq} 2\lambda^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)}) - \bar{\mathbf{g}}^{(t)}\right\|^2 + 2(1-\lambda)^2\,\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)}) - \bar{\mathbf{y}}^{(t)}\right\|^2 \\
&= 2\lambda^2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\left(\nabla f(\bar{\mathbf{x}}^{(t)}) \pm \nabla f_i(\mathbf{x}_i^{(t)}) - \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right)\right\|^2 + 2(1-\lambda)^2\,\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)}) - \bar{\mathbf{y}}^{(t)}\right\|^2 \\
&\overset{(c)}{\leq} \frac{4\lambda^2}{n^2}\sum_{i=1}^n\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 + \frac{4\lambda^2\sigma^2}{n} + 2(1-\lambda)^2\,\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)}) - \bar{\mathbf{y}}^{(t)}\right\|^2 \\
&= \frac{4\lambda^2}{n^2}\sum_{i=1}^n\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)}) \pm \nabla f(\mathbf{x}_i^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 + \frac{4\lambda^2\sigma^2}{n} + 2(1-\lambda)^2\,\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)}) - \bar{\mathbf{y}}^{(t)}\right\|^2 \\
&\overset{(d)}{\leq} \frac{8L^2\lambda^2}{n^2}\sum_{i=1}^n\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|^2 + \frac{4\lambda^2}{n}\left(\sigma^2 + 2\zeta^2\right) + 2(1-\lambda)^2\underbrace{\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)}) - \bar{\mathbf{y}}^{(t)}\right\|^2}_{T_6},
\end{aligned}
\tag{42}
$$

其中，(a) 根据 $\bar{\mathbf{m}}^{(t)}$ 的定义可得；(b) 根据 **事实4** 可得；(c) 和 (d) 通过应用 **事实4**、詹森不等式和假设2可得。对于 T_6，我们可以估计如下

$$
\begin{aligned}
T_6 &= \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\nabla f(\bar{\mathbf{x}}^{(t)}) \pm \frac{1}{n}\sum_{i=1}^n\nabla f(\mathbf{x}_i^{(t)}) - \bar{\mathbf{y}}^{(t)}\right\|^2 \\
&\overset{(a)}{\leq} \frac{2}{n^2}\sum_{i=1}^n\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)}) - \nabla f(\mathbf{x}_i^{(t)})\right\|^2 + 2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\nabla f(\mathbf{x}_i^{(t)}) - \frac{1}{n}\sum_{i=1}^n\mathbf{y}_i^{(t)}\right\|^2 \\
&\overset{(b)}{\leq} \frac{2L^2}{n^2}\sum_{i=1}^n\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|^2 + 2\mathbb{E}\left\|\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n\nabla f(\mathbf{x}_i^{(t)}) \pm \frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n\nabla f_j(\mathbf{x}_i^{(t)}) - \frac{1}{n}\sum_{i=1}^n\mathbf{y}_i^{(t)}\right\|^2 \\
&\overset{(c)}{\leq} \frac{2L^2}{n^2}\sum_{i=1}^n\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|^2 + \frac{2}{n^2}\sum_{i=1}^n\left(\frac{2}{n^2}\sum_{j=1}^n\mathbb{E}\left\|\nabla f(\mathbf{x}_i^{(t)}) - \nabla f_j(\mathbf{x}_i^{(t)})\right\|^2 + 2\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_i^{(t)}) - \mathbf{y}_i^{(t)}\right\|^2\right) \\
&\overset{(d)}{\leq} \frac{2L^2}{n^2}\sum_{i=1}^n\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|^2 + \frac{4\zeta^2}{n^2} + 4\epsilon^2,
\end{aligned}
\tag{43}
$$

其中，(a) 根据 **事实4** 可得；(b) 是因为不等式右侧第一项满足假设1；(c) 通过结合詹森不等式应用 **事实** 4 可得；(d) 是因为假设 2 和命题 1。将 (43) 代入 (42) 得

$$
T_3 \leq \left(\frac{8L^2\lambda^2}{n^2} + \frac{4L^2(1-\lambda)^2}{n^2}\right)\mathbb{E}\left\|\bar{\mathbf{X}}^{(t)} - \mathbf{X}^{(t)}\right\|_F^2 + \frac{4\lambda^2\sigma^2 + 16\zeta^2}{n} + 8(1-\lambda)^2\epsilon^2. \tag{44}
$$

将（41）和（44）代入（36），可得

$$
\begin{aligned}
\mathbb{E}f(\bar{\mathbf{z}}^{(t+1)}) &\leq \mathbb{E}f(\bar{\mathbf{z}}^{(t)}) + \frac{\beta^2\hat{\eta}^2\tilde{\eta}L^2}{(1-\beta)^3}\sum_{l=0}^{t-1}\beta^{t-1-l}\left(\frac{16G^2}{n} + \frac{4\sigma^2}{n} + \frac{12\zeta^2}{n^2} + 6\epsilon^2\right) - \frac{\tilde{\eta}}{2}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{6L^2\tilde{\eta}}{n^2}\mathbb{E}\left\|\bar{\mathbf{X}}^{(t)} - \mathbf{X}^{(t)}\right\|_F^2 \\
&\quad + \frac{2\tilde{\eta}\lambda^2\sigma^2 + 8\tilde{\eta}\zeta^2}{n} + 4\tilde{\eta}(1-\lambda)^2\epsilon^2 \\
&\leq \mathbb{E}f(\bar{\mathbf{z}}^{(t)}) + \frac{\beta^2\hat{\eta}^2\tilde{\eta}L^2}{(1-\beta)^4}\left(\frac{16G^2}{n} + \frac{4\sigma^2}{n} + \frac{12\zeta^2}{n} + 6\epsilon^2\right) - \frac{\tilde{\eta}}{2}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{2\tilde{\eta}\lambda^2\sigma^2 + 8\tilde{\eta}\zeta^2}{n} + 4\tilde{\eta}(1-\lambda)^2\epsilon^2 \\
&\quad + \frac{6L^2\tilde{\eta}}{n^2}\left(\mathbb{E}\left\|\bar{\mathbf{X}}^{(t)} - \mathbf{X}^{(t)}\right\|_F^2 + \mathbb{E}\left\|\bar{\mathbf{Y}}^{(t)} - \mathbf{Y}^{(t)}\right\|_F^2\right),
\end{aligned}
\tag{45}
$$

最后一个不等式成立是因为 $\sum_{l=0}^{t-1}\beta^{t-1-l}\le\frac{1}{1-\beta}$，并且我们加上了非负项 $\mathbb{E}\left\|\bar{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2$。

命题3. 设 $\{V_t\}_{t\ge0}$ 为一个非负序列，$C_2\ge0$ 为某个常数，满足 $\forall t\ge1, V_{t+1}\le Q_2V_t+Q_2V_{t-1}+C_2$，其中 $Q_2\in(0,1)$。则当 $T\ge1$时，以下不等式成立，

$$\begin{aligned}V_{t+1}&\le Q_2V_t+Q_2V_{t-1}+C_2\\&\le Q_2V_t+V_{t-1}+C_2\\&\le Q_2^2V_{t-1}+(Q_2V_{t-2}+V_{t-1})+(Q_2+1)C_2\\&\cdots\\&\le Q_2^tV_1+\sum_{l=0}^{t-1}Q_2^{t-1-l}V_l+C_2\sum_{l=0}^{t-1}Q_2^l.\end{aligned}\tag{46}$$

将 t 从0 累加到 $T-1$,

$$\begin{aligned}\sum_{t=0}^{T-1}V_t&=V_0+V_1+\sum_{t=2}^{T-1}V_t\\&\le V_0+V_1+V_1\sum_{t=2}^{T-1}Q_2^{t-1}+\sum_{t=2}^{T-1}\sum_{l=0}^{t-2}Q_2^{t-2-l}V_l+C_2\sum_{t=2}^{T-1}\sum_{l=0}^{t-2}Q_2^l\\&\le V_0+V_1+V_1\sum_{t=0}^{\infty}Q_2^t+\sum_{t=0}^{T-1}\sum_{l=0}^{\infty}Q_2^lV_l+C_2\sum_{t=0}^{T-1}\sum_{l=0}^{\infty}Q_2^l\\&\le V_0+\frac{(2-Q_2)V_1}{1-Q_2}+\frac{TV_{\max}}{1-Q_2}+\frac{C_2T}{1-Q_2},\end{aligned}\tag{47}$$

其中 $V_{\max}=$最大值$_{0\le t\le T-1}\{V_t\}$。

将 (45) 中的 t 从 0 累加到 $T-1$，然后将两边除以 $2/\tilde{\eta}$ 并整理各项。最后，将引理4应用于命题3。具体而言，我们考虑 $C_2=\beta_0^2(1-\frac{\rho}{2})L^2(\sigma^2+\zeta^2)(192\lambda^2L^2+\rho)$, $V_t\triangleq\mathbb{E}\left\|\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2+\mathbb{E}\left\|\mathbf{Y}^{(t)}-\bar{\mathbf{Y}}^{(t)}\right\|_F^2$,$Q_2\triangleq\max\left\{4\beta_0^2\left(1-\frac{\rho}{4}\right),8(1-\rho)(1+\alpha\beta)^2\right\}$ 确保 $\begin{cases}4\beta_0^2\left(1-\frac{\rho}{4}\right)<1\\8(1-\rho)(1+\alpha\beta)^2<1\end{cases}$。由于 $\mathbf{x}_i^{(0)}=\bar{\mathbf{x}}^{(0)}=\mathbf{0}$，且$\mathbb{E}\left\|\mathbf{Y}^{(0)}-\bar{\mathbf{Y}}^{(0)}\right\|_F^2=\mathbb{E}\left\|\mathbf{G}^{(0)}-\bar{\mathbf{G}}^{(0)}\right\|_F^2\le\frac{6L^2}{n}\mathbb{E}\left\|\mathbf{X}^{(0)}-\bar{\mathbf{X}}^{(0)}\right\|+6(\zeta^2+\sigma^2)=6(\zeta^2+\sigma^2)$，因此 $V_0\le Q_3=6(\zeta^2+\sigma^2)$。此外，我们假设 $\forall i$，$\mathbf{x}_i^{(-1)}=\mathbf{y}_i^{(-1)}=\mathbf{0}$，并在引理2的(3)和(4)中将其扩展至初始阶段（即 $t=0$），因此 $V_1\le Q_4=\beta_0^2\left(1-\frac{\rho}{2}\right)(\sigma^2+\zeta^2)\left(48+\frac{\rho}{L^2}+192\lambda^2\right)$。证毕。

D 实验设置

D.1 数据集和模型描述

MNIST 是一个包含 10 个类别的手写数字图像分类数据集，共有 70,000 28×28 个样本，其中 60,000 个为训练数据集，其余 10,000 个为测试数据集。其扩展版本 EMNIST 包含数字以及大写和小写英文字母的图像，共有 62 个类别。CIFAR10 是 8000万图像数据集的一个带标签子集，与之共享相同的 60,000 张输入图像，并包含 10 个不同标签。对于自然语言处理，AG NEWS 是一个包含 4 个类别的新闻文章分类数据集，其中有 120,000 个训练样本和 7,600 个测试样本。总体描述见表3。

表3：数据集和模型

数据集	Task	训练样本	测试样本	类别	模型
MNIST [LeCun 等人，1998]	手写字符识别（计算机视觉）	60，000	10,000	10	表4中描述的LeNet
EMNIST [Cohen 等人，2017]	手写字符识别（计算机视觉）	731,668	82，587	62	表5中描述的卷积神经网络
CIFAR10 [Krizhevsky等人，2009]	图像分类（计算机视觉）	50，000	10，000	10	表4中描述的 LeNet
AG NEWS [Zhang 等，2015]	文本分类（自然语言处理）	120，000	7，600	4	表6中所述的循环神经网络

表4：MNIST 和 CIFAR10 上的 LeNet 模型。

层	输出形状	超参数	激活函数
二维卷积	(28，28，6)	卷积核大小 $= 5$	ReLU
二维最大池化	(14，14，6)	池化大小 $= 2$	
二维卷积	(10，10，16)	卷积核大小 $= 5$	ReLU
二维最大池化	(5，5，16)	池化大小 $= 2$	
展平层	400		
全连接层	120		
全连接层	84		
丢弃层	84	$p = 0.5$	
全连接层	10		

表5：EMNIST上的卷积神经网络模型。

层	输出形状	超参数	激活函数
二维卷积	(26，26，32)	卷积核大小 $= 3$，步幅 $= 1$	
二维卷积	(24，24，64)	卷积核大小 $= 3$，步幅 $= 1$	ReLU
二维最大池化	(12，12，64)	池化大小 $= 2$	
丢弃层	(12,12, 64)	$p = 0.25$	
展平层	$9,216$		
全连接层	128		
丢弃层	128	$p = 0.5$	
全连接层	62		softmax

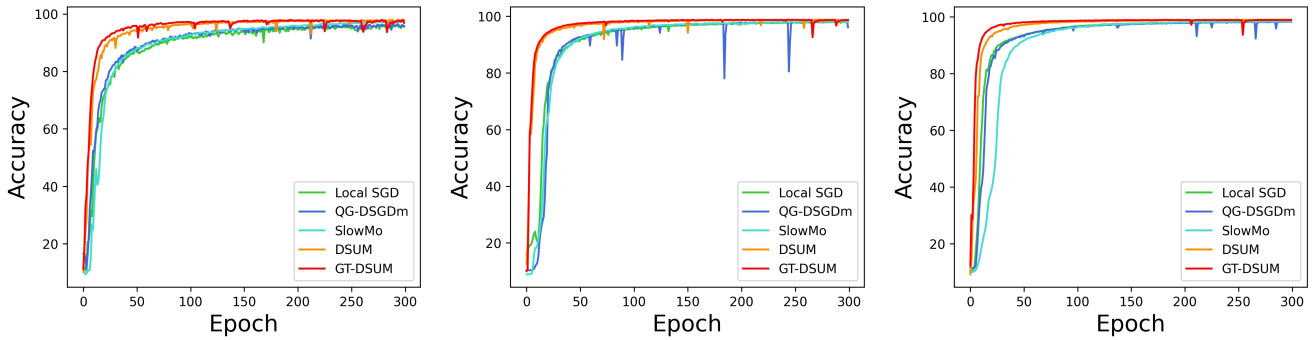
表6：AG NEWS 上的循环神经网络模型。

层	超参数
嵌入袋	嵌入 $= 95$，812，维度 $= 64$
全连接层	输入特征 $= 64$, 输出特征 $= 4$
丢弃层	$p = 0.5$

E 其他评估

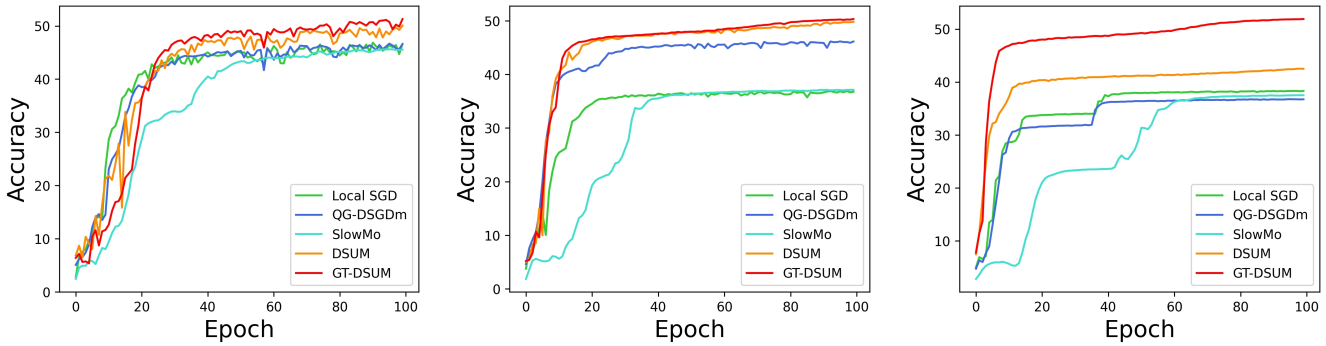
E.1 模型泛化

图4、5、6和7展示了不同基准上模型训练过程的实验结果。我们注意到，所提出方法的优势在加速收敛方面体现得更为明显。例如，在 MNIST 上使用 LeNet 时，D-SUM 和 GT-DSUM 均需约 50 轮次才能收敛，这使训练轮次减少了 40%。在更大的数据集上，即CIFAR10 和 AG NEWS，我们提出的算法在训练轮次方面比其他基线方法收敛得更快。



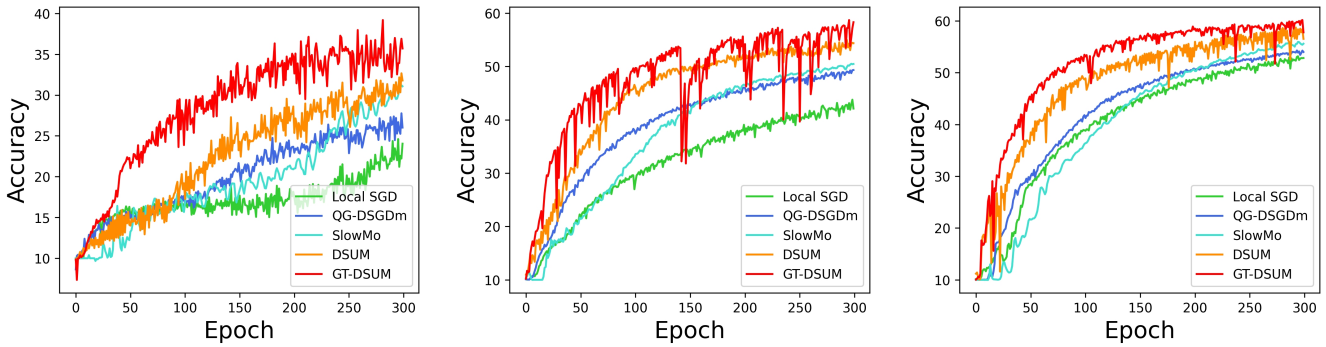
（a）非独立同分布 = 0.1　（b）非独立同分布 = 1　（c）非独立同分布 = 10

图4：在**LeNet** 上使用**MNIST**训练不同任务的测试准确率。



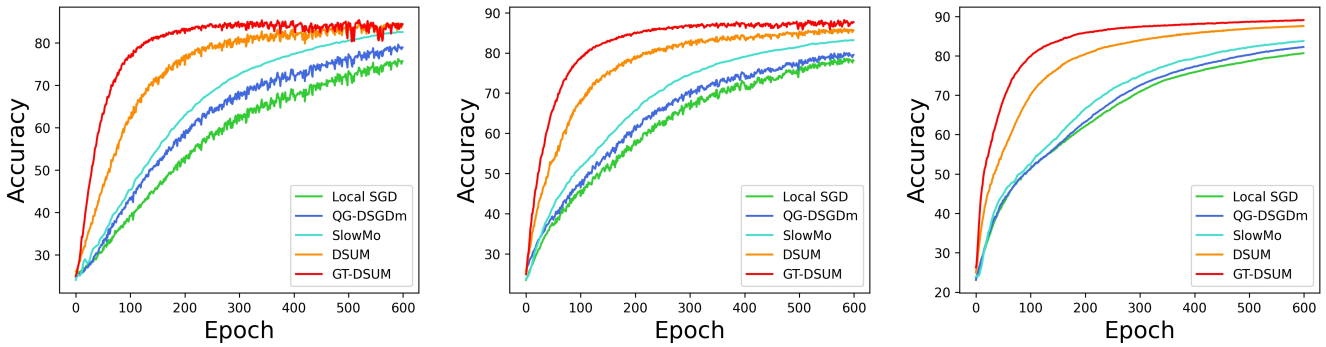
(a) 非独立同分布 = 0.1　(b) 非独立同分布 = 1　(c) 非独立同分布 = 10

图5：在**卷积神经网络** 上使用**EMNIST**训练不同任务的测试准确率。



（a）非独立同分布 = 0.1　（b）非独立同分布 = 1　（c）非独立同分布 = 10

图6：在**LeNet** 上使用**CIFAR10**进行各种任务训练的测试准确率。



（a）非独立同分布 = 0.1　（b）非独立同分布 = 1　（c）非独立同分布 = 10

图7：在**循环神经网络** 上针对**AG NEWS**进行各种任务训练的测试准确率。