arXiv:2305.04792v1 [cs.LG] 8 May 2023

全局更新跟踪：一种适用于异构数据的去中心化学习算法

Sai Aparna Aketi Abolfazl Hashemi KaushikRoy电
气与计算机工程系 普渡大学 印第安纳州西拉斐特
47906 {saketi, abolfazl, kaushik}
@purdue.edu

摘要

去中心化学习使得无需中央服务器，即可在不同地点生成的大规模分布式数据集上训练深度学习模型。然而，在实际场景中，不同设备上的数据分布可能存在显著差异，导致模型性能下降。本文着重设计一种不易受设备间数据分布变化影响的去中心化学习算法。我们提出了全局更新跟踪（GUT），这是一种新颖的基于跟踪的方法，旨在不增加任何通信开销的情况下，减轻异构数据对去中心化学习的影响。我们通过在各种计算机视觉数据集（CIFAR-10、CIFAR-100、Fashion MNIST 和 Imagenette）、模型架构及网络拓扑上进行大量实验，证明了该方法的有效性。实验结果表明，与其他现有技术相比，所提出的方法在异构数据上的去中心化学习中实现了最先进水平，测试准确率提高了1 – 6%。

1 引言

去中心化学习是分布式优化的一个分支，专注于从分布在多个智能体上的数据中进行学习，而无需中央服务器。与传统的集中式方法相比，去中心化学习方法在数据隐私、容错性和可扩展性等核心方面具有诸多优势 [18]。研究表明，去中心化学习算法 [15] 在基准视觉数据集上的表现可与集中式算法相媲美。 [15] 提出的去中心化并行随机梯度下降（DSGD）将随机梯度下降与八卦平均算法 [26]相结合。此外，作者通过分析证明，DSGD 的收敛速度与其集中式对应算法相近 [5]。随后， [3]提出了 DSGD 的动量版本，即去中心化动量随机梯度下降（DSGDm）。 [2] 的作者介绍了随机梯度推送（SGP），将 DSGD 扩展到了有向图和时变图。最近， [11]提出了一个统一框架，用于分析基于八卦通信的去中心化随机梯度下降方法，并给出了目前已知的最佳收敛保证。

上述所有去中心化算法要实现最先进水平的性能，其关键假设之一是，智能体之间的数据满足独立同分布（IID）。具体而言，假设数据以均匀且随机的方式分布在各个智能体之间。然而，在大多数现实场景中，这一假设并不成立，因为各智能体之间的数据分布存在显著差异（非独立同分布/异构） [9]。在点对点去中心化设置中，异构数据所产生的影响是一个相对缺乏研究的问题，也是当前一个活跃的研究领域。

预印本。正在接受同行评审。

近年来，已有不少研究致力于缩小去中心化设置 [16, 23, 19, 7, 1, 24]中独立同分布数据与非独立同分布数据之间的性能差距。交叉梯度聚合 [7] 和邻域梯度聚类 [1]算法利用交叉梯度的概念来降低异构数据的影响，并在性能（测试准确率）方面取得了显著提升。然而，与 DSGD 等标准去中心化算法相比，这些技术会产生更高的 $2\times$ 通信成本。 [23] 中提出的 D^2 算法被证明不受数据异质性的影响，并且可应用于深度学习任务。 D^2的一个主要局限是，其收敛性要求混合拓扑结构的负特征值具有不低于 $-\frac{1}{3}$的下界。此外，研究表明，在某些情况下， D^2 的性能不如 DSGD [16]。
梯度跟踪（GT） [6, 19] 和动量跟踪（MT） [22]等跟踪机制已被提出，用于应对去中心化场景中的异构数据。但这些算法是以增加 $2\times$ 通信开销为代价来提升性能的。文献 [16] 的作者提出了准全局动量（QGM），这是一种无需通信的方法，通过模拟动量缓冲区的全局同步，缓解异构数据给去中心化学习带来的困难。近期，文献 [24] 提出了 RelaySGD，以 Relay-Sum 取代 gossip 平均步骤。由于 RelaySGD 针对的是 gossip 平均步骤，因此它与上述算法正交，可以与它们协同使用。将 QGM 融入 DSGDm 的 QG-DSGDm [16] 在不增加通信成本的前提下，使异构数据上的去中心化学习达到最先进水平。本研究探讨以下问题：能否通过跟踪机制提升异构数据上的去中心化学习效果，同时不增加任何通信开销？

为此，我们提出全局更新跟踪（*GUT*），这是一种旨在提升异构数据分布下性能的新型去中心化学习算法。受梯度跟踪机制启发，但又与之不同，我们提出通过跟踪全局/平均模型更新来跟踪共识模型（$\bar{x}^t$），其中 x_i^t 是智能体 i 在时间步 t 的模型参数，而 $\bar{x}$ 是平均模型参数。在传统的跟踪方法 [19, 22] 中，每个智能体都跟踪平均梯度，并与邻居通信模型参数 x_i^t 和跟踪变量 y_i^t，因此会产生 $2\times$ 通信开销。所提出的*GUT* 算法通过允许智能体存储邻居模型参数的副本，并跟踪模型更新而非梯度，克服了这一瓶颈。这样只需通信跟踪变量y_i^t，即可得到模型更新（$x_i^t - x_i^{t-1}$）。我们通过在各种数据集、模型架构和图拓扑上开展大量实验，验证了所提出算法的有效性。我们还给出了详细的收敛分析，表明*GUT* 算法的收敛速度与最先进的去中心化学习算法相当。此外，我们表明，在相同通信成本下，
QG-GUTm - 带准全局动量的全局更新跟踪在异构数据上的表现优于当前最先进的去中心化学习算法（即 QG-DSGDm）。

1.1 贡献

综上所述，我们做出了以下贡献。

- 我们提出了全局更新跟踪（*GUT*），这是一种基于跟踪的新型去中心化学习算法，用于减轻异构数据分布的影响。
- 我们从理论上建立了所提出算法收敛到一阶解的非渐近收敛速率。

- 通过在各种数据集、模型架构和图拓扑上进行全面的实验，我们证明了采用准全局动量的全局更新跟踪（*QG-GUTm*）在各种异构数据上优于当前最先进的去中心化学习算法。

2 背景

In t本节介绍具有点对点连接的去中心化设置背景。 s.
去中心化机器学习的主要目标是利用从 n 个智能体本地存储的数据样本中提取的知识来学习一个全局模型，同时满足隐私约束。在

具体而言，我们解决如式（1）所示的优化问题，即最小化分布在 n 个智能体上的全局损失函数 $f(x)$。注意，F_i 是一个局部损失函数（例如，交叉熵损失），它根据从智能体 i的本地数据集 D_i 中采样得到的数据（d_i）定义。

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad \text{where } \; f_i(x) = \mathbb{E}_{d_i\sim D_i}[F_i(x;d_i)], \; \text{ for all } i. \tag{1}$$

该优化问题通常通过结合随机梯度下降 [4] 与基于全局共识的八卦平均 [26]来求解。通信拓扑建模为一个图$G = ([N], \; E)$，当且仅当智能体 i和 j 通过通信链路直接连接并交换消息时，两者之间才存在边 $\{i, j\} \in E$。我们将 $\mathcal{N}(i)$表示为智能体 i 的邻居集合，其中包括其自身。假设图 G 是强连通的且包含自环，即从每个智能体到其他任一智能体都存在一条路径。图 G 的邻接矩阵称为混合矩阵 W，其中 w_{ij} 是与边 $\{i, j\}$关联的权重。请注意，权重为 0 表示智能体之间不存在直接边，单位矩阵的元素则表示为 I_{ij}。与大多数先前的去中心化学习研究类似，假设混合矩阵是双随机的。此外，初始模型和所有超参数在训练开始时均已同步。假设智能体之间的通信是同步的。

传统去中心化算法（如 DSGD）假设各智能体上的数据是独立同分布（IID）的。在 DSGD 中，每个智能体维护局部参数，并按如下方式更新它们。

$$\text{DSGD:} \;\; x_i^{t+1} = \sum_{j\in\mathcal{N}(i)} w_{ij}(x_j^t - \eta g_j^t); \quad g_i^t = \nabla F_j(x_i^t, d_i^t). \tag{2}$$

我们重点关注具有非独立同分布/异构数据的去中心化设置。具体而言，数据分布的异质性表现为类似于 [9] 的偏斜标签划分。在异构数据分布下，采用 DSGD 算法进行去中心化学习会导致性能下降，这是由于各智能体之间的局部梯度存在巨大差异。为了解决这一问题， [16] 中的作者提出了一种基于动量的优化技术（QG-DSGDm），引入了准全局动量，如公式 (3) 所示。

$$\text{QG-DSGDm:} \;\; x_i^{t+1} = \sum_{j\in\mathcal{N}(i)} w_{ij}[x_j^t - \eta(g_j^t + \beta m_j^{t-1})]; \;\; m_i^t = \mu m_i^{t-1} + (1-\mu)\frac{x_i^{t-1} - x_i^t}{\eta}. \tag{3}$$

QG-DSGDm 在不产生任何通信开销的情况下提升了异构数据上的去中心化学习性能，并在本研究中用作比较基线。

梯度跟踪（GT）机制 [19, 22]还因能够通过减少局部梯度与平均（全局）梯度之间的方差来改善异构数据上的去中心化学习而闻名。为实现这一目标，梯度跟踪算法引入了一个跟踪变量 y_i^t，用于近似总梯度，并用于更新局部参数 x_i^t （参见（4））。

$$\text{GT:} \;\; x_i^{t+1} = \sum_{j\in\mathcal{N}(i)} w_{ij}(x_j^t - \eta y_j^t); \quad y_i^t = \sum_{j\in\mathcal{N}(i)} w_{ij} y_j^{t-1} - g_i^{t-1} + g_i^t. \tag{4}$$

跟踪变量的更新规则是递归地将一个校正项$\big(\sum_{j\in\mathcal{N}(i)} w_{ij} y_j^{t-1} - g_i^{t-1}\big)$ 加到本地梯度 g_i^t上，使 y_i^t更接近全局梯度（$\frac{1}{n}\sum_{j=1}^n g_j^t$）。这要求每个智能体 i 与其邻居通信两组参数 x_i^t 和 y_i^t。因此，梯度跟踪算法以增加 $2\times$ 通信开销为代价，改善了非独立同分布数据上的去中心化学习。

3 全局更新跟踪

我们提出了全局更新跟踪（*GUT*）算法，用于非独立同分布数据分布下的去中心化深度学习。*GUT* 是一种无通信跟踪机制，旨在缓解智能体之间数据分布存在异质性时去中心化训练所面临的困难。

算法1全局更新跟踪（*GUT*）

输入： 每个智能体 $i \in [1,n]$ 初始化模型参数 x_i^0和邻居副本 $\hat{x}_j^0$、步长 η、*GUT* 缩放因子 μ、混合矩阵 $W = [w_{ij}]_{i,\ j\in[1,n]}$, $\mathcal{N}(i)$ 表示包括自身在内的 i 的邻居，并记作 $\hat{x}_i^t = x_i^t$。

每个智能体同时执行 TRAIN( ) 过程 1

. **procedure** TRAIN( )
2. **for** t = $0, 1, \ldots, T-1$ **do**
3. $d_i^t \sim D_i$
4. $g_i^t = \nabla_x F_i(\sum_{j\in\mathcal{N}(i)} w_{ij}\hat{x}_j^t; d_i^t)$
5. $\delta_i^t = g_i^t - \frac{1}{\eta}\sum_{j\in\mathcal{N}(i)} w_{ij}(\hat{x}_j^t - x_i^t)$
6. $y_i^t = \delta_i^t + \mu\Big[\sum\limits_{j\in\mathcal{N}(i)} w_{ij}(y_j^{t-1} - \frac{1}{\eta}(\hat{x}_j^t - x_i^t)) - \delta_i^{t-1}\Big]$
7. SENDRECEIVE(y_i^t)
8. $x_i^{t+1} = x_i^t - \eta y_i^t$
9. $\hat{x}_j^{t+1} = \hat{x}_j^t - \eta y_j^t \ \ \forall \ \ j \in \mathcal{N}(i)\backslash i$
10. **end**
11. **return** $\frac{1}{n}\sum_{i=1}^n x_i^T$

为了在不产生通信开销的情况下获得梯度跟踪的优势，我们提出针对模型更新 $x_i^t - x_i^{t-1}$ 而非梯度 g_i^t应用跟踪机制。首先，为了设计一种无需额外通信成本的跟踪机制，每个智能体 i 向其邻居传递模型更新，而不是模型参数。智能体 i 将其邻居的参数副本存储为 $\hat{x}_j$，并利用接收到的模型更新对其进行更新，以获取邻居参数的当前版本，如算法1第9行所示。算法的一种内存高效实现方式（附录B中的算法4）要求每个智能体存储 $s_i = \sum_{j\in\mathcal{N}(i)} w_{ij}\hat{x}_j$，而不是分别存储每个邻居的副本，因此仅需额外的 $\mathcal{O}(1)$ 内存 [12]。

现在，我们在每个智能体 i 上定义一个变量 δ_i^t，该变量累积局部梯度更新 g_i^t 和 gossip 平均更新 $\sum_j (w_{ij} - I_{ij})\hat{x}_j^t$，如算法1第5行所示。注意，通过在更新规则中使用 δ_i^t，即 $x_i^{t+1} = x_i^t - \eta\delta_i^t$，我们可以恢复式（2）中定义的 DSGD 更新。接下来，我们按照算法1第6行所述，利用 δ_i^t 所反映的组合模型更新（局部梯度部分和 gossip 平均部分）来计算跟踪变量 y_i^t。每个智能体 i 的更新中的 gossip 平均部分 $\sum_j w_{ij}(\hat{x}_j^t - x_i^t)$ 是相对于其自身的模型权重计算的。为了在跟踪变量 y_i^t 的计算中考虑这一点，各智能体必须调整从邻居（即 y_j^t）接收到的信息，以将参考对象转换为自身。这体现为算法1第6行所给更新规则中的一个附加项 $\frac{1}{\eta}(\hat{x}_j^t - x_i^t)$。此外，我们将跟踪变量的校正项乘以一个因子 μ 进行缩放，该因子是一个超参数，通过调节该参数来充分发挥所提出算法的最大优势。

总之，*GUT* 的更新方案可以重新表述为以下矩阵形式，其中$X = [x_1, \ldots, x_n] \in \mathbb{R}^{d\times n}$ 是模型参数， $G = [g_1, \ldots, g_n] \in \mathbb{R}^{d\times n}$ 是随机梯度。

$$\begin{aligned} X^{t+1} &= X^t - \eta Y^t, \\ Y^{t+1} &= G^{t+1} - \frac{1}{\eta}(W-I)X^{t+1} + \mu[WY^t - G^t - \frac{1}{\eta}(W-I)(X^{t+1} - X^t)]. \end{aligned} \tag{5}$$

最后，我们证明，将所提出的*GUT* 算法与准全局动量相结合，无需增加任何通信开销，即可显著提升当前最先进水平。我们算法的动量版本（*QG-GUTm*）的伪代码见附录B。

4 收敛保证

本节对所提出的*GUT* 算法进行收敛性分析。我们假设以下标准假设成立：

假设1 (利普希茨梯度)。 每个函数 $f_i(x)$ 均为 L-光滑，即 $||\nabla f_i(y) - \nabla f_i(x)|| \leq L||y-x||$。

假设2 (有界方差)。 随机梯度是无偏的，并假设其方差有界。

$$\mathbb{E}_{d\sim D_i}||\nabla F_i(x;d) - \nabla f_i(x)||^2 \leq \sigma^2 \quad \forall i \in [1,n], \tag{6}$$

$$\frac{1}{n}\sum_{i=1}^{n}||\nabla f_i(x) - \nabla \mathcal{F}(x)||^2 \leq \zeta^2. \tag{7}$$

假设3 (双随机混合矩阵)。 混合矩阵 W 是一个实双随机矩阵，满足 $\lambda_1(W) = 1$ 并且

$$\max\{|\lambda_2(W)|, |\lambda_N(W)|\} \leq 1-\rho < 1, \tag{8}$$

其中 $\lambda_i(W)$ 是 W 的 i^{th} 最大特征值， ρ 是谱间隙。

上述假设通常用于大多数去中心化学习设置。定理1给出了所提出的*GUT* 算法的收敛性，证明详见附录A。

定理1. （*GUT*算法的收敛性）给定假设 *1*、*2* 和 *3*，令步长为 $\eta \leq \frac{\rho}{7L}$，缩放因子为 $\frac{\mu}{1-\mu} \leq \frac{\rho}{42}$。对于所有 $T \geq 1$，有

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\nabla f(\bar{x}^t)||^2 \leq \frac{4}{\eta T}(f(\bar{x}^0) - f^*) + \eta\frac{4L\sigma^2}{n} + \eta^2\frac{1248L^2}{\rho^2}(\zeta^2 + \sigma^2(2-\mu)), \tag{9}$$

其中 $f(\bar{x}^0) - f^*$ 为次优性间隙， $\bar{x}$ 为平均模型参数/共识模型参数。

定理1的结果表明，平均模型的平均梯度上界由次优性间隙（初始目标函数值与最优值之差）、采样方差（σ）以及各智能体之间反映数据异质性的梯度差异（ζ）共同决定。此外，我们给出一个推论，以迭代次数的形式展示*GUT* 的收敛速度。

推论 1. 假设步长满足 $\eta = \mathcal{O}\left(\sqrt{\frac{n}{T}}\right)$。对于足够大的 T，有：

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\nabla f(\bar{x}^t)||^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{nT}} + \frac{1}{T}\right). \tag{10}$$

推论 1 表明，当 T 足够大时，*GUT* 算法可实现线性加速，其收敛速度为 $\mathcal{O}(\frac{1}{\sqrt{nT}})$，且与通信拓扑无关。换言之，
找到一个 ϵ-一阶解所需的通信复杂度，即 $\mathbb{E}\|\nabla f(\bar{x})\|^2 \leq \epsilon$，为 $\mathcal{O}(\frac{\sigma^2}{n\epsilon^2})$。这一收敛速度与文献中去中心化随机梯度下降算法的著名最优结果相近 [15]。

5 实验

在本节中，我们分析所提出的 *GUT* 和 *QG-GUTm* 技术的性能，并分别将其与基线 DSGD [15] 和当前最先进水平的 QG-DSGDm [16]进行比较。 [1]

[1]我们的 PyTorch 代码可在 github.com/aparna-aketi/global_update_ 上获取跟踪

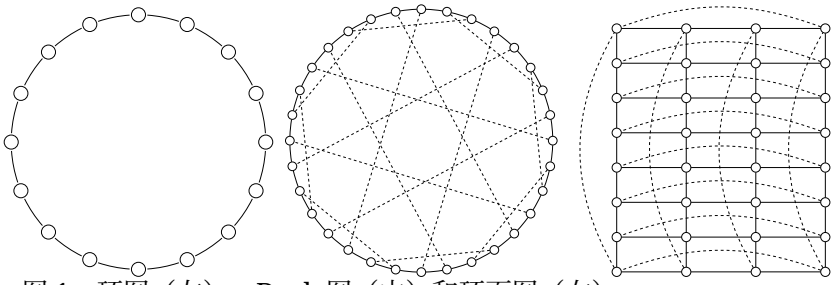


图 1：环图（左）、Dyck 图（中）和环面图（右）。

5.1 实验设置

我们通过在多种数据集、模型架构、图拓扑结构和图规模上开展实验，证明了所提出方法的有效性。我们从以下几个方面进行分析：（a）数据集：CIFAR-10、CIFAR-100、Fashion MNIST 和 Imagenette。（b）模型架构：VGG-11、ResNet-20、LeNet-5 和 MobileNet-V2。所有模型均采用 Evonorm [17] 作为激活归一化层，因为研究表明，该层更适合在异构数据上进行去中心化学习。（c）图拓扑结构：每个智能体拥有 2 个对等节点的环图、每个智能体拥有 3 个对等节点的 Dyck 图，以及每个智能体拥有 4 个对等节点的环面图（参见图 1）。（d）智能体数量：16–40 个智能体。我们使用狄利克雷分布，在各智能体之间生成互不相交的非独立同分布数据。各智能体之间创建的数据划分是固定且互不重叠的，并且在训练过程中不会在不同智能体之间进行重新打乱。异质性程度由 α 的取值控制 —— α 越小，各智能体之间的非独立同分布程度就越高。我们报告共识模型在三个随机选取的种子下取得的平均测试准确率。所有实验的去中心化设置和超参数的详细信息均见附录 C。

5.2 平均共识任务

我们首先考虑一个与随机梯度下降学习过程相分离的平均一致性任务。这里的目标是使所有智能体达成共识，即所有智能体最终达到的值应为每个智能体所持有的初始信息的平均值。以下方程展示了移除梯度更新部分后*GUT* (11) 和 *QG-GUTm* (12) 的简化版本。

$$X^{t+1} = X^t + Y^t; \;\; Y^t = (W-I)X^t + \mu[WY^{t-1} - (W-I)(X^{t-1} - X^t)], \tag{11}$$

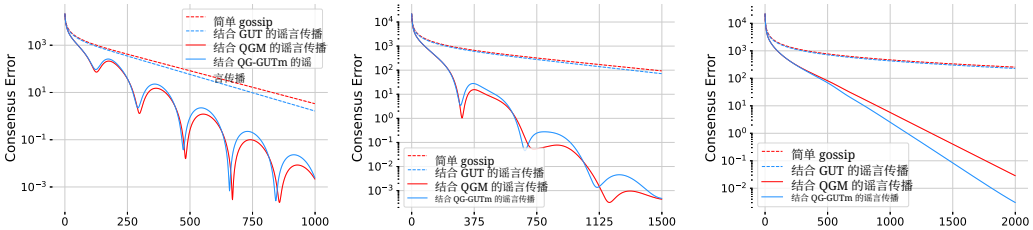
$$\begin{aligned} X^{t+1} &= X^t + \hat{M}^t; \quad M^t = \beta M^{t-1} + (1-\beta)(X^t - X^{t-1}) \\ \hat{M}^t &= \beta M^t + (1-\beta)[(W-I)X^t + \mu(W\hat{M}^{t-1} - (W-I)(X^{t-1} - X^t))]. \end{aligned} \tag{12}$$

请注意，在 (11)、12 中将超参数 μ 设为 0，分别得到简单 gossip[26] 和准全局 gossip [16]，并且所有智能体都会在迭代 t 时与其邻居通信 $X^t - X^{t-1}$。

图 2 展示了在环形拓扑上采用各种算法执行平均一致性任务时，随时间变化的平均共识误差，即 $\frac{1}{n}||X^t - \bar{X}||_F^2$。我们观察到，结合 *GUT* 的 gossip 平均比简单 gossip 平均收敛更快。图 2(c) 说明，对于谱间隙较小的图（对应于更多智能体），所提出的*QG-GUTm* 比准全局 gossip（结合 QGM 的谣言传播）收敛更快，从而实现更好的去中心化优化。

5.3 去中心化深度学习结果

我们通过一系列全面的实验来评估*GUT* 及其准全局动量版本*QG-GUTm* 的效率。我们将*GUT* 与 DSGD 进行比较，并将*QG-GUTm* 与 QG-DSGDm 进行比较，结果表明，所提出的方法优于当前最先进水平。表 1 展示了在 CIFAR-10 数据集上，针对不同程度的非独立同分布数据，在包含 16 个和 32 个智能体的环形拓扑中训练不同模型架构（ResNet-20 和 VGG-11）所获得的平均测试准确率。我们观察到，对于所有模型、图规模和异质性程度，*GUT* 的表现始终优于 DSGD，性能提升幅度显著，变化范围从 $1-18\%$ 起。准全局



(a) 64 个智能体 (b) 128 智能体 (c) 256 个智能体

图2：无向环拓扑上的去中心化平均共识问题

表1：在不同环拓扑上，对采用不同异质性程度（非独立同分布）分布的 CIFAR-10 数据进行评估时，不同去中心化算法的平均测试准确率。结果为三个随机种子的平均值，并标示了标准差。我们还列出了 IID基线的结果，即 DSGDm（独立同分布）；其中本地数据以随机方式划分，且独立于 α。

智能体（n）	方法	ResNet-20		
		$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.01$
16	DSGDm（独立同分布）		89.75 ± 0.29	
	DSGD	84.17 ± 0.32	72.21 ± 2.37	54.66 ± 4.74
	GUT（我们的方法）	84.72 ± 0.20	81.86 ± 1.99	70.16 ± 4.94
	QG-DSGDm	**88.23** ± 0.51	84.21 ± 2.12	79.85 ± 2.11
	QG-GUTm（我们的方法）	88.22 ± 0.36	**86.44** ± 0.36	**81.04** ± 1.66
32	DSGDm（独立同分布）		88.52 ± 0.23	
	DSGD	78.25 ± 0.42	62.97 ± 1.90	42.58 ± 1.84
	GUT（我们的方法）	79.24 ± 0.33	76.07 ± 0.23	60.72 ± 1.03
	QG-DSGDm	87.15 ± 0.33	83.50 ± 1.04	69.99 ± 0.60
	QG-GUTm（我们的方法）	**87.48** ± 0.33	**84.94** ± 0.60	**72.04** ± 3.18

智能体（n）	方法	VGG-11		
		$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.01$
16	DSGDm（独立同分布）		85.76 ± 0.28	
	DSGD	81.78 ± 0.29	76.20 ± 0.81	68.93 ± 1.23
	GUT（我们的方法）	82.12 ± 0.09	81.24 ± 0.95	76.62 ± 1.37
	QG-DSGDm	84.23 ± 0.47	81.70 ± 0.79	77.08 ± 3.19
	QG-GUTm（我们的方法）	**84.46** ± 0.33	**83.05** ± 0.48	**78.32** ± 1.03
32	DSGDm（独立同分布）		84.75 ± 0.30	
	DSGD	79.75 ± 0.56	73.37 ± 1.02	59.93 ± 1.60
	GUT（我们的方法）	80.37 ± 0.33	79.55 ± 1.00	73.59 ± 1.26
	QG-DSGDm	83.67 ± 0.28	80.82 ± 0.19	74.25 ± 2.02
	QG-GUTm（我们的方法）	**84.32** ± 0.11	**83.39** ± 0.38	**77.41** ± 3.44

我们算法的动量版本 *QG-GUTm* 在异质性程度更高的 CIFAR-10 数据集划分情况下，以 $1-3.5\%$ 的提升击败了 QG-DSGDm（$\alpha = 0.1$,0.01）。

我们展示了在各种图拓扑和数据集上的实验结果，以证明*QG-GUTm*的可扩展性和泛化能力。我们在 Dyck 图和环面图上使用 ResNet-20 训练 CIFAR-10 数据集，以说明连通性对所提技术的影响。如表 2 所示，随着连通性（或谱间隙）的变化，我们获得了 $0.5-3.5\%$ 的性能提升。此外，我们在 Fashion MNIST、Imagenette 等多种图像数据集以及 CIFAR-100 等具有挑战性的数据集上评估*QG-GUTm* 。表 3 显示，*QG-GUTm* 在各种数据集上的表现均优于 QG- DSGDm，提升了 $0.2-6.2\%$。因此，在去中心化深度学习设置中，所提出的*GUT* 和*QG-GUTm* 算法对数据分布中的异质性具有更强的鲁棒性，平均性能提升 2%，优于所有对比方法。

表 2：在各种图拓扑上使用 ResNet-20 训练，并在 CIFAR-10 数据集上评估的不同去中心化算法的平均测试准确率

方法	Dyck 图（32 个智能体）		环面（32 个智能体）	
	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.01$
QG-DSGDm	86.49 ± 0.81	81.32 ± 1.50	86.88 ± 0.30	85.20 ± 0.56
QG-GUTm（我们的方法）	$\mathbf{86.93} \pm 0.53$	$\mathbf{84.80} \pm 0.47$	$\mathbf{87.75} \pm 0.42$	$\mathbf{86.20} \pm 0.82$

表 3：在16智能体环形拓扑上，将数据集以不同程度的异质性分布后，对不同去中心化算法进行评估所得的平均测试准确率

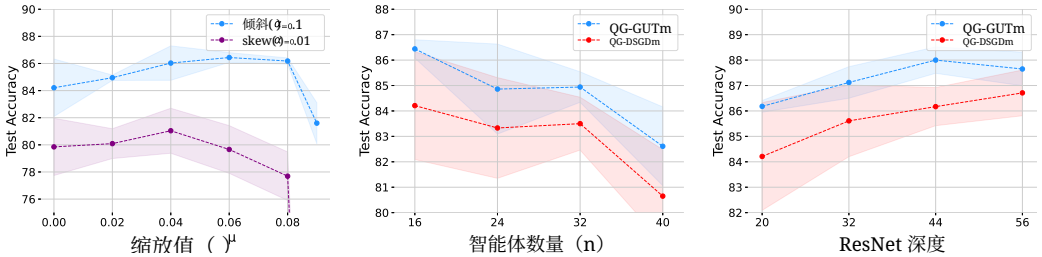
方法	Fashion MNIST（LeNet-5）		CIFAR-100（ResNet-20）		Imagenette（MobileNet-V2）	
	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.01$
QG-DSGDm	89.94 ± 0.44	83.43 ± 0.94	53.19 ± 1.68	44.17 ± 3.64	63.60 ± 4.50	39.49 ± 4.57
QG-GUTm	$\mathbf{90.11} \pm 0.02$	$\mathbf{84.60} \pm 1.00$	$\mathbf{53.40} \pm 1.23$	$\mathbf{50.45} \pm 1.30$	$\mathbf{66.52} \pm 3.68$	$\mathbf{43.85} \pm 8.24$

5.4 消融研究

首先，我们分析表 4 所示的利用或跟踪模型更新信息的不同方式。除*GUT* 和*DSGD*[15] 外，我们还提出了两种不同的更新规则，并将它们与梯度跟踪进行比较 [19]。规则-a 将所提出的跟踪机制应用于模型更新，但不会将从邻居接收的跟踪变量 y_j 的参考对象更改为自身（有关更改参考对象的详情，请参阅第 3 节）。对于规则-b，每个智能体计算邻域平均模型更新（$W(X^t - X^{t-1})$）与自身模型更新（$X^t - X^{t-1}$）之间的差值，并将两者之差作为偏差校正项加上。表 4 显示，这些朴素的跟踪方式或偏差校正更新规则（规则-a、b）并未提升去中心化学习在异构数据上的性能。这证实了我们的发现：*GUT* 技术是一种有效且有理论保证的共识模型跟踪方法，并且无需任何通信开销即可优于梯度跟踪机制。

随后，我们通过实证研究进一步考察了不同动量变体与 *GUT* 结合时的效果。从表 5（更多结果请参见附录 D）可以得出结论：全局更新跟踪的准全局变体始终优于其他方法。这表明，所提出的*GUT* 算法能够加速去中心化优化，并且可以与准全局动量协同使用，从而实现最大的性能提升。

此外，图 3(a) 展示了使用 *QG-GUTm* 时，扩展 μ 对测试准确率的影响，并指出 $\mu = 0$ 表示 QG-DSGDm 的测试准确率。图 3(b) 和图 3(c) 展示了可扩展性
在不同图规模和模型规模下的*QG-GUTm* 。在不同图规模下，*QG-GUTm* 的表现比 QG-DSGDm 高出~ 1.7 %，在不同模型规模下高出 ~ 1.4%。



(a) $n = 16$，ResNet-20　(b) $\alpha = 0.1$，ResNet-20　(c) $\alpha = 0.1$，$n = 16$

图 3：对超参数 μ、智能体数量 n 和模型规模的消融研究。报告了在环形拓扑上使用 ResNet 架构训练 CIFAR-10 数据集的测试准确率。

表4：分析模型更新的不同变体。在包含 $\alpha = 0.1$个智能体的环形拓扑上使用 ResNet-20 训练 CIFAR-10 数据集，并评估测试准确率。

方法	更新 Rule	通信参数（成本）	Test 准确率
DSGD	$X^{t+1}=X^t-\eta Y^t$ $Y^t \quad =G^t-\frac{1}{\eta}(W-I)X^t$	X^t (1×)	72.21 ± 2.37
规则-a	$X^{t+1}=X^t-\eta Y^t$ $Y^t \quad =G^t-\frac{1}{\eta}(W-I)X^t+\mu[WY^{t-1}-(G^{t-1}-\frac{1}{\eta}(W-I)X^{t-1})]$	Y^t (1×)	72.78 ± 0.80
规则-b	$X^{t+1}=X^t-\eta Y^t$ $Y^t \quad =G^t-\frac{1}{\eta}(W-I)X^t+\mu[-\frac{1}{\eta}(W-I)(X^t-X^{t-1})]$	Y^t (1×)	72.62 ± 1.16
GUT	$X^{t+1}=X^t-\eta Y^t$ $Y^t \quad =G^t-\frac{1}{\eta}(W-I)X^t+\mu[WY^{t-1}-G^{t-1}-\frac{1}{\eta}(W-I)(X^t-X^{t-1})]$	Y^t (1×)	$\mathbf{81.86} \pm 1.99$
梯度跟踪	$X^{t+1}=X^t-\eta[Y^t-\frac{1}{\eta}(W-I)X^t]$ $Y^t \quad =G^t+WY^{t-1}-G^{t-1}$	X^t, Y^t (2×)	80.61 ± 2.41

表 5：在 16智能体环形拓扑上使用 ResNet-20 架构训练 CIFAR-10 数据集，评估采用不同动量版本的全局更新跟踪（GUT）

方法	本地动量	内斯特罗夫	准全局动量	全局更新跟踪	测试准确率 $\alpha = 0.1$
DSGD	x	x	x	x	72.21 ± 2.37
DSGDm	✓	x	x	x	79.87 ± 1.73
DSGDm-N	✓	✓	x	x	81.31 ± 0.51
QG-DSGDm	x	x	✓	x	84.21 ± 2.12
QG-DSGDm-N	x	✓	✓	x	85.12 ± 1.11
GUT	x	x	x	✓	81.86 ± 1.99
GUTm	✓	x	x	✓	79.95 ± 1.67
GUTm-N	✓	✓	x	✓	82.08 ± 1.74
QG-GUTm	x	x	✓	✓	86.44 ± 0.36
QG-GUTm-N	x	✓	✓	✓	$\mathbf{86.55} \pm 0.49$

6 讨论与局限性

通过一系列详尽的实验和消融研究，我们展示了全局更新跟踪（*GUT*）算法的优越性。在实验中，我们重点研究了双随机对称图结构。将所提出的*GUT* 算法与随机梯度推送（SGP） [2]相结合，即可轻松扩展到有向图和时变图。此外，*GUT* 引入的额外项也可以解释为一种偏差校正机制：所添加的偏差会推动局部模型趋向共识（平均）模型。*GUT* 这一解释的矩阵表示见式（13），并在引理1中进行了分析。

$$X^{t+1} = WX^t - \eta(G^t + \mu B^t); \quad B^{t+1} = -\frac{1}{\eta}[(2W - I)(X^{t+1} - X^t) + \eta G^t]. \tag{13}$$

引理1。 给定假设*3*，我们定义 $\bar{b}^t = B^t \frac{1}{n}\mathbb{1}\mathbb{1}^T$，其中$\mathbb{1}$是全*1*向量。对于所有 t，我们有：$\bar{b}^t = \mu\bar{b}^{t-1}$。

引理1的完整证明见附录A.2。引理1指出，由于 $\bar{b}^0$ 为零，*GUT* 所引入的平均偏差为零。因此，*GUT* 算法能够保持去中心化系统的平均值。这一特性是我们建立定理1的基础。

GUT 算法存在两个潜在局限：a）内存开销；b）引入额外的超参数。GUT 要求智能体保存邻居的平均模型参数副本，这会增加一个大小相当于模型参数的额外内存缓冲区。存储跟踪变量也会增加内存开销，需要额外占用相当于模型参数大小的内存。此外，我们还引入了一个新的超参数 μ，需要像学习率或动量系数一样进行调优。另外，针对 GUT 算法提出的理论分析没有考虑动量，并假设通信是同步的。我们将对 QG-GUTm 进行理论分析以及制定 GUT 的异步版本留作未来的研究方向。

7 结论

基于异构数据的去中心化学习，是在边缘设备上启动机器学习训练并有效利用海量用户生成的隐私数据的关键。本文提出了一种新颖的去中心化算法——全局更新跟踪（GUT），旨在改善异构数据分布下的学习效果。本文的收敛分析表明，所提出的算法达到了去中心化算法中已知的最佳收敛速率。此外，本文还提出了该算法的准全局动量版本 QG-GUT m，以进一步提升性能。针对不同模型架构、数据集和拓扑结构开展的实验提供了经验证据，表明这两种算法都具有优越的性能。总之，所提出的算法及其准全局动量版本有望促进边缘设备上更具可扩展性和效率的去中心化学习。

参考文献

[1] Sai Aparna Aketi、Sangamesh Kodge 和 Kaushik Roy。邻域梯度聚类：一种针对非独立同分布数据分布的高效去中心化学习方法。*arXiv* 预印本 *arXiv:2209.14390*，2022年。[2] Mahmoud Assran、Nicolas Loizou、Nicolas Ballas 和 Mike Rabbat。用于分布式深度学习的随机梯度推送。载于国际机器学习会议，第344–353页。PMLR，2019年。[3] Aditya Balu、Zhanhong Jiang、Sin Yong Tan、Chinmay Hedge、Young M Lee 和 Soumik Sarkar。采用动量加速共识的去中心化深度学习。载于*ICASSP 2021- 2021 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第3675–3679页。IEEE，2021年。[4] Léon Bottou。使用随机梯度下降进行大规模机器学习。载于*COMPSTAT'2010*论文集，第177–186页。Springer，2010年。[5] Jeffrey Dean、Greg Corrado、Rajat Monga、Kai Chen、Matthieu Devin、Mark Mao、Marc'aurelio Ranzato、Andrew Senior、Paul Tucker、Ke Yang 等。大规模分布式深度网络。神经信息处理系统进展，第25卷，2012年。[6] 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年。[7] 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。载于国际机器学习会议，第3036–3046页。PMLR，2021年。[8] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。[9] 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。载于第*37*届国际机器学习会议论文集，机器学习研究论文集第119卷，第4387–4398页。PMLR，2020年7月13日至18日。[10] Hamel Husain。Imagenette——一个由ImageNet数据集中的10个易分类类别组成的子集。*https://github.com/fastai/imagenette*，2018年。[11] Anastasia Koloskova、Nicolas Loizou、Sadra Boreiri、Martin Jaggi 和 Sebastian Stich。关于变化的拓扑结构与本地更新下去中心化随机梯度下降的统一理论。载于国际机器学习会议，第5381–5393页。PMLR，2020年。[12] Anastasia Koloskova、Sebastian Stich 和 Martin Jaggi。采用压缩通信的去中心化随机优化与八卦算法。载于第*36*届国际机器学习会议论文集，机器学习研究论文集第97卷，第3478–3487页。PMLR，2019年6月9日至15日。[13] Alex Krizhevsky、Vinod Nair 和 Geoffrey Hinton。Cifar（加拿大高等研究院）。*http://www.cs.toronto.edu/kriz/cifar.html*，2014年。[14] Yann LeCun、Léon Bottou、Yoshua Bengio 和 Patrick Haffner。应用于文档识别的基于梯度的学习。*IEEE*论文集，86(11):2278–2324，1998年。[15] 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的案例研究。神经信息处理系统进展，第30卷，2017年。

[16] Tao Lin、Sai Praneeth Karimireddy、Sebastian 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。载于第*38*届国际机器学习会议论文集，机器学习研究论文集第139卷，第6654–6665页。PMLR，2021年7月18日至24日。[17] Hanxiao Liu、Andy Brock、Karen Simonyan 和 Quoc Le。演化归一化激活层。载于 H. Larochelle、M. Ranzato、R. Hadsell、M. F. Balcan 和 H. Lin 编，神经信息处理系统进展，第33卷，第13539–13550页。Curran Associates, Inc.，2020年。[18] Angelia Nedic。网络中凸机器学习问题的分布式梯度方法：分布式优化。*IEEE*信号处理杂志，37(3):92–101，2020年。[19] Shi Pu 和 Angelia Nedi´c。分布式随机梯度跟踪方法。数学规划，187:409–457，2021年。[20] Mark Sandler、Andrew Howard、Menglong Zhu、Andrey Zhmoginov 和 Liang-Chieh Chen。MobileNetV2：倒置残差与线性瓶颈。载于*IEEE*计算机视觉与模式识别会议论文集，第4510–4520页，2018年。[21] Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。*arXiv* 预印本 *arXiv:1409.1556*，2014年。[22] Yuki Takezawa、Han Bao、Kenta Niwa、Ryoma Sato 和 Makoto Yamada。动量跟踪：加速异构数据上的去中心化深度学习。*arXiv* 预印本 *arXiv:2209.15505*，2022年。[23] Hanlin Tang、Xiangru Lian、Ming Yan、Ce Zhang 和 Ji Liu。 d^2：在去中心化数据上进行去中心化训练。载于国际机器学习会议，第4848–4856页。PMLR，2018年。[24] Thijs Vogels、Lie He、Anastasiia Koloskova、Sai Praneeth Karimireddy、Tao Lin、Sebastian U Stich 和 Martin Jaggi。面向异构数据去中心化深度学习的RelaySum。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，神经信息处理系统进展，第34卷，第28004–28015页。Curran Associates, Inc.，2021年。[25] Han Xiao、Kashif Rasul 和 Roland Vollgraf。Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集。*arXiv* 预印本 *arXiv:1708.07747*，2017年。[26] Lin Xiao 和 Stephen Boyd。用于分布式平均的快速线性迭代。《系统与控制快报》，53(1):65–78，2004年。

收敛速率证明

在这项工作中，我们解决了如下所示的优化问题：最小化分布在 n 个智能体上的全局损失函数 $f(x)$。请注意， F_i 是一个局部损失函数（例如交叉熵损失），它根据从智能体 i 的本地数据集 D_i 中采样的数据（d_i）定义。

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x),$$
$$and \;\; f_i(x) = \mathbb{E}_{d_i\in D_i}[F_i(x;d_i)] \;\; \forall i.$$

我们用矩阵形式重述算法 1 中提出的 GUT 更新方案：

$$\begin{aligned} X^{t+1} &= X^t - \eta Y^t \\ Y^t &= \Delta^t + \mu[WY^{t-1} - \frac{1}{\eta}(W-I)X^t - \Delta^{t-1}] \\ \Delta^t &= G^t - \frac{1}{\eta}(W-I)X^t, \end{aligned} \tag{14}$$

其中， W 是混合矩阵， I 是单位矩阵， $X=[x_1,x_2,\ldots,x_n]\in\mathbb{R}^{d\times n}$ 是包含模型参数的矩阵， $x_i\in\mathbb{R}^d$ 是智能体 i 的模型参数， $Y=[y_1,y_2,\ldots,y_n]\in\mathbb{R}^{d\times n}$ 是包含跟踪变量的矩阵， $G=[g_1,g_2,\ldots,g_n]\in\mathbb{R}^{d\times n}$ 是包含局部梯度的矩阵， μ 是 GUT 缩放因子， η 是学习率。现在，我们将上述方程改写为偏差校正更新的形式：

$$\begin{aligned} X^{t+1} &= WX^t - \eta(G^t + \mu B^t) \\ B^t &= -\frac{1}{\eta}[(2W-I)(X^t - X^{t-1}) + \eta G^{t-1}]. \end{aligned} \tag{15}$$

A.1 假设

我们假设以下陈述成立：

Assu假设 1 - 利普希茨梯度：每个函数 $f_i(x)$ 都是 L-光滑的，即 $||\nabla f_i(y) - \nabla f_i(x)|| \leq L||y-x||$。等价地，

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y-x\rangle + \frac{L}{2}||y-x||^2 \tag{16}$$

假设 2 - 有界方差：随机梯度的方差假定为有界。

$$\mathbb{E}_{d\sim D_i}||\nabla F_i(x;d) - \nabla f_i(x)||^2 \leq \sigma^2 \;\; \forall i\in[1,n]$$

$$\frac{1}{n}\sum_{i=1}^{n}||\nabla f_i(x) - \nabla\mathcal{F}(x)||^2 \leq \zeta^2$$

Assu假设 3 - 双随机混合矩阵：混合矩阵 W 是一个实双随机
[illegible]近满足 $\lambda_1(W) = 1$ 且

$$max\{|\lambda_2(W)|, |\lambda_N(W)|\} \leq 1-\rho < 1$$

其中， $\lambda_i(W)$ 是 W 的 i^{th} 最大特征值， ρ 是谱间隙。混合矩阵满足 $\mathbb{E}_W||ZW - \bar{Z}||_F^2 \leq (1-\rho)||ZW - \bar{Z}||_F^2$，其中 $\bar{Z} = Z\frac{1}{n}\mathbb{1}\mathbb{1}^T$。我们还有 $W\mathbb{1}^T = \mathbb{1}$ 和 $W^T\mathbb{1} = \mathbb{1}$。

此外，我们定义平均梯度 $\bar{g}^t = \frac{1}{n}\sum_{i=1}^{n}\nabla F_i(x_i^t, d_i^t)$，其中 d_i^t 是在节点 i 上采样的小批量数据。

A.2 引理1的证明

引理1： 给定假设*3*，我们定义 $\bar{b}^t = B^t \frac{1}{n}\mathbb{1}\mathbb{1}^T$，其中 $\mathbb{1}$ 是一个全*1*向量。对于所有 t，我们有：$\bar{b}^t = \mu\bar{b}^{t-1}$。

证明。从 B^t的定义出发

$$
\begin{aligned}
&B^t = -\frac{1}{\eta}[(2W - I)(X^t - X^{t-1}) + \eta G^{t-1}] \\
&\text{multiply } \frac{1}{n}\mathbb{1}\mathbb{1}^T \text{ on both sides} \\
&\bar{b}^t = -\frac{1}{\eta}[\bar{x}^t - \bar{x}^{t-1} + \eta\bar{g}^{t-1}] \quad (\because (2W - I)\mathbb{1} = \mathbb{1}) \\
&\text{now, multiplying } \frac{1}{n}\mathbb{1}\mathbb{1}^T \text{ to } X^{t+1} = WX^t - \eta(G^t + \mu B^t) \\
&\bar{x}^{t+1} = \bar{x}^t - \eta\bar{g}^t - \eta\mu\bar{b}^t \implies \bar{x}^t - \bar{x}^{t-1} + \eta\bar{g}^{t-1} = -\eta\mu\bar{b}^{t-1} \\
&\implies \bar{b}^t = \mu\bar{b}^{t-1}
\end{aligned}
$$

□

鉴于 $\bar{b}^0 = 0$，每次迭代的平均偏差均为零。这表明，所提出的算法*GUT* 在平均一致性任务中保持系统的平均值不变。

A.3 定理1的证明

本节给出定理1所述*GUT* 算法收敛界的详细证明。首先，我们分析平均模型参数 $\bar{x}$的单步进展。注意，$\bar{X}^t = [\bar{x}^t, \bar{x}^t, \ldots, \bar{x}^t] \in \mathbb{R}^{d\times n}$ 和 $\bar{x}^t = \frac{1}{n}\sum_{i=1}^n x_i^t$

引理2。给定假设*1-3*和 $\eta \le \frac{1}{4L}$，我们有 —
$\mathbb{E}f(\bar{x}^{t+1}) \le \mathbb{E}f(\bar{x}^t) - \frac{\eta}{4}\mathbb{E}||\nabla f(\bar{x}^t)||^2 - \frac{\eta}{4}\mathbb{E}||\frac{1}{n}\sum_{i=1}^n \nabla f(x_i^t)||^2 + \frac{L\eta^2\sigma^2}{n} + \frac{3L\eta^2}{n}||X^t - \bar{X}^t||_F^2.$

证明。根据 X^{t+1}的定义，我们有

$$
\begin{aligned}
X^{t+1} &= WX^t - \eta[G^t + \mu B^t] \\
\implies \bar{x}^{t+1} &= \bar{x}^t - \eta\bar{g}^t \quad (\because \bar{b}^t = 0 \text{from Lemma 1})
\end{aligned}
$$

采用式（16）给出的 L-光滑性假设

$$
\begin{aligned}
\mathbb{E}f(\bar{x}^{t+1}) &\le \mathbb{E}f(\bar{x}^t) + \mathbb{E}\langle\nabla f(\bar{x}^t), \bar{x}^{t+1} - \bar{x}^t\rangle + \frac{L}{2}\mathbb{E}||\bar{x}^{t+1} - \bar{x}^t||^2 \\
&= \mathbb{E}f(\bar{x}^t) + \mathbb{E}\langle\nabla f(\bar{x}^t), -\eta\bar{g}^t\rangle + \frac{L\eta^2}{2}\mathbb{E}||\bar{g}^t||^2 \\
&= \mathbb{E}f(\bar{x}^t) - \eta\mathbb{E}\langle\nabla f(\bar{x}^t), \mathbb{E}[\bar{g}^t]\rangle + \frac{L\eta^2}{2}\mathbb{E}||\frac{1}{n}\sum_{i=1}^n \nabla F_i(x_i^t)||^2 \\
&= \mathbb{E}f(\bar{x}^t) - \eta\frac{1}{n}\sum_{i=1}^n \mathbb{E}\langle\nabla f(\bar{x}^t), \nabla f_i(x_i^t)\rangle + \frac{L\eta^2}{2}\mathbb{E}||\frac{1}{n}\sum_{i=1}^n (\nabla F_i(x_i^t) \pm \nabla f_i(x_i^t))||^2 \\
&\overset{(a)}{\le} \mathbb{E}f(\bar{x}^t) - \eta\mathbb{E}\langle\nabla f(\bar{x}^t), \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^t)\rangle + \frac{L\eta^2}{2}\mathbb{E}||\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^t)||^2 + \frac{L\eta^2\sigma^2}{n} \\
&\overset{(b)}{=} \mathbb{E}f(\bar{x}^t) + \frac{L\eta^2\sigma^2}{n} + \frac{L\eta^2}{2}\mathbb{E}||\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^t)||^2 - \frac{\eta}{2}\mathbb{E}||\nabla f(\bar{x}^t)||^2 \\
&\quad - \frac{\eta}{2}\mathbb{E}||\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^t)||^2 + \frac{\eta}{2}\mathbb{E}||\frac{1}{n}\sum_{i=1}^n (\nabla f_i(x_i^t) - \nabla f(\bar{x}^t))||^2
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(c)}{\leq} \mathbb{E}f(\bar{x}^t) + \frac{L\eta^2\sigma^2}{n} + (\frac{L\eta^2}{2} - \frac{\eta}{2})\mathbb{E}||\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^t)||^2 - \frac{\eta}{2}\mathbb{E}||\nabla f(\bar{x}^t)||^2 \\
&\quad + \frac{\eta}{2n}\sum_{i=1}^{n}\mathbb{E}||\nabla f_i(x_i^t) - \nabla f(\bar{x}^t)||^2 \\
&\overset{(d)}{\leq} \mathbb{E}f(\bar{x}^t) + \frac{L\eta^2\sigma^2}{n} + (\frac{L\eta^2}{2} - \frac{\eta}{2})\mathbb{E}||\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^t)||^2 - \frac{\eta}{2}\mathbb{E}||\nabla f(\bar{x}^t)||^2 \\
&\quad + \frac{L^2\eta}{2n}\sum_{i=1}^{n}\mathbb{E}||x_i^t - \bar{x}^t||^2 \\
&\overset{(e)}{\leq} \mathbb{E}f(\bar{x}^t) - \frac{\eta}{4}\mathbb{E}||\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^t)||^2 - \frac{\eta}{4}\mathbb{E}||\nabla f(\bar{x}^t)||^2 + \frac{L\eta^2\sigma^2}{n} \\
&\quad + \frac{3L^2\eta}{n}\sum_{i=1}^{n}\mathbb{E}||X^t - \bar{X}^t||_F^2
\end{aligned}
$$

(a) 使用假设 2（(6)）。(b) 使用了 $-2\langle a,b\rangle = -||a||^2 - ||b||^2 + ||a-b||^2$ 这一事实。(c) 使用詹森不等式。(d) 使用 L-光滑条件。(e) 根据假设 $\eta \leq \frac{1}{4L}$ 得出。 □

现在，我们通过引理 3 来界定共识误差。

引理 3. 在假设 *1-3* 和 $\eta \leq \frac{\rho}{7L}$ 下，我们有—

$\frac{1}{n}\mathbb{E}||X^{t+1} - \bar{X}^{t+1}||_F^2 \leq \frac{1-\rho/4}{n}\mathbb{E}||X^t - \bar{X}^t||_F^2 + \frac{12\eta^2\zeta^2}{\rho} + 4\eta^2\sigma^2 + \frac{6\eta^2\mu^2}{\rho n}\mathbb{E}||B^t||_F^2.$

证明。从更新步骤 15 开始

$$
\begin{aligned}
\frac{1}{n}\mathbb{E}||X^{t+1} - \bar{X}^{t+1}||_F^2 =& \frac{1}{n}\mathbb{E}||WX^t - \eta[G^t + \mu B^t] - (\bar{X}^t - \eta\bar{G}^t)||_F^2 \\
=& \frac{1}{n}\mathbb{E}||WX^t - \bar{X}^t - \eta(G^t - \bar{G}^t) - \eta\mu B^t||_F^2 \\
\leq& \frac{1}{n}\mathbb{E}||WX^t - \bar{X}^t - \eta(\mathbb{E}[G^t] - \mathbb{E}[\bar{G}^t]) - \eta\mu B^t||_F^2 + 4\eta^2\sigma^2 \\
\overset{(a)}{\leq}& \frac{1+\rho/2}{n}\mathbb{E}||WX^t - \bar{X}^t||_F^2 + \frac{\eta^2(1+2/\rho)}{n}\mathbb{E}||\mathbb{E}[G^t] - \mathbb{E}[\bar{G}^t] - \mu B^t||_F^2 \\
&+ 4\eta^2\sigma^2 \\
\overset{(b)}{\leq}& \frac{(1-\rho)(1+\rho/2)}{n}\mathbb{E}||X^t - \bar{X}^t||_F^2 + \frac{3\eta^2}{n\rho}\mathbb{E}||\mathbb{E}[G^t] - \mathbb{E}[\bar{G}^t] - \mu B^t||_F^2 \\
&+ 4\eta^2\sigma^2 \\
\leq& \frac{(1-\rho)(1+\rho/2)}{n}\mathbb{E}||X^t - \bar{X}^t||_F^2 + 4\eta^2\sigma^2 + \frac{6\eta^2}{n\rho}\mathbb{E}||\mathbb{E}[G^t] - \mathbb{E}[\bar{G}^t]||_F^2 \\
&+ \frac{6\eta^2\mu^2}{n\rho}\mathbb{E}||B^t||_F^2 \\
\leq& \frac{(1-\rho/2)}{n}\mathbb{E}||X^t - \bar{X}^t||_F^2 + 4\eta^2\sigma^2 + \frac{6\eta^2}{n\rho}\mathbb{E}||\mathbb{E}[G^t] - \nabla f(\bar{x}^t)||_F^2 \\
&+ \frac{6\eta^2\mu^2}{n\rho}\mathbb{E}||B^t||_F^2 \\
\leq& \frac{(1-\rho/2)}{n}\mathbb{E}||X^t - \bar{X}^t||_F^2 + 4\eta^2\sigma^2 + \frac{6\eta^2\mu^2}{n\rho}\mathbb{E}||B^t||_F^2 \\
&+ \frac{6\eta^2}{n\rho}\sum_{i=1}^{n}\mathbb{E}||\nabla f_i(x_i^t) \pm \nabla f_i(\bar{x}^t) - \nabla f(\bar{x}^t)||_F^2
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(c)}{\leq}\frac{(1-\rho/2)}{n}\mathbb{E}||X^t-\bar{X}^t||_F^2+4\eta^2\sigma^2+\frac{12\eta^2\zeta^2}{\rho}+\frac{6\eta^2\mu^2}{n\rho}\mathbb{E}||B^t||_F^2\\
&\quad+\frac{12\eta^2}{n\rho}\sum_{i=1}^{n}\mathbb{E}||\nabla f_i(x_i^t)-\nabla f_i(\bar{x}^t)||_F^2\\
&\overset{(d)}{\leq}\frac{(1-\rho/2)}{n}\mathbb{E}||X^t-\bar{X}^t||_F^2+4\eta^2\sigma^2+\frac{12\eta^2\zeta^2}{\rho}+\frac{6\eta^2\mu^2}{n\rho}\mathbb{E}||B^t||_F^2\\
&\quad+\frac{12\eta^2L^2}{n\rho}\sum_{i=1}^{n}\mathbb{E}||x_i^t-\bar{x}^t||_F^2\\
&=\Big(\frac{1-\rho/2}{n}+\frac{12\eta^2L^2}{n\rho}\Big)\mathbb{E}||X^t-\bar{X}^t||_F^2+4\eta^2\sigma^2+\frac{12\eta^2\zeta^2}{\rho}\\
&\quad+\frac{6\eta^2\mu^2}{n\rho}\mathbb{E}||B^t||_F^2\\
&\overset{(e)}{\leq}\frac{1-\rho/4}{n}\mathbb{E}||X^t-\bar{X}^t||_F^2+4\eta^2\sigma^2+\frac{12\eta^2\zeta^2}{\rho}+\frac{6\eta^2\mu^2}{n\rho}\mathbb{E}||B^t||_F^2
\end{aligned}
$$

(a) 由以下事实得出：$||a+b||^2\leq(1+\alpha)||a||^2+(1+\frac{1}{\alpha})||b||^2\ \forall\alpha>0$并令 $\alpha=\frac{\rho}{2}$。(b) 使用 $\mathbb{E}_W||ZW-\bar{Z}||_F^2\leq(1-\rho)||ZW-\bar{Z}||_F^2$和 $1+\frac{2}{\rho}\leq\frac{3}{\rho}$ (c) 使用假设 2 (7)。(d) 使用 L-光滑条件。(e) 由以下假设得出：$\eta\leq\frac{\rho}{7L}$ — □

下一步是为偏差项 $\mathbb{E}||B^t||_F^2$ 求一个上界。

引理 4. 在假设 *1-3* 和 $\frac{\mu}{1-}\leq\frac{\rho}{42}$ 下，我们有 μ—

$\frac{6\eta^2\mu^2}{\rho n(1-\mu)}\mathbb{E}||B^{t+1}||_F^2\leq\Big(\frac{6\eta^2\mu^2}{\rho n(1-\mu)}-\frac{6\eta^2\mu^2}{\rho n}\Big)\mathbb{E}||B^t||_F^2+\frac{\rho}{8n}\mathbb{E}||X^t-\bar{X}^t||_F^2+\frac{\eta^2\zeta^2\rho}{8}+\frac{\eta^2\sigma^2\rho(1-\mu)}{8}$.

证明。 从更新步骤 15 开始

$$
\begin{aligned}
B^{t+1}&=-\frac{1}{\eta}[(2W-I)(X^{t+1}-X^t)+\eta G^t]\\
&=-\frac{1}{\eta}[(2W-I)(WX^t-\eta G^t-\eta\mu B^t-X^t)+\eta G^t]\\
&=-\frac{1}{\eta}[W(2W-I)-I]X^t+2(W-I)G^t+\mu(2W-I)B^t.
\end{aligned}
$$

Now,

$$
\begin{aligned}
\frac{1}{n}\mathbb{E}||B^{t+1}||_F^2=&\frac{1}{n}\mathbb{E}||-\frac{1}{\eta}(W(2W-I)-I)X^t+2(W-I)G^t+\mu(2W-I)B^t||_F^2\\
=&\frac{1}{n}\mathbb{E}||-\frac{1}{\eta}(W(2W-I)-I)X^t+2(W-I)(G^t-\bar{G}^t)+\mu(2W-I)B^t||_F^2\\
=&\frac{1}{n}\mathbb{E}||-\frac{1}{\eta}(W(2W-I)-I)X^t+2(W-I)\mathbb{E}[G^t-\bar{G}^t]+\mu(2W-I)B^t||_F^2\\
&+\frac{1}{n}\mathbb{E}||2(W-I)(G^t-\mathbb{E}[G^t]-(\bar{G}^t-\mathbb{E}[\bar{G}^t]))||_F^2\\
\leq&\frac{1}{n}\mathbb{E}||\frac{1}{\eta}(I-W(2W-I))X^t+2(W-I)\mathbb{E}[G^t-\bar{G}^t]+\mu(2W-I)B^t||_F^2\\
&+8\sigma^2\\
\overset{(a)}{\leq}&\frac{1}{n}\Big(1+\frac{1-\mu}{\mu}\Big)\mathbb{E}||\mu(2W-I)B^t||_F^2+8\sigma^2\\
&+\frac{1}{n}\Big(1+\frac{\mu}{1-\mu}\Big)\mathbb{E}||\frac{1}{\eta}(I-W(2W-I))X^t+2(W-I)\mathbb{E}[G^t-\bar{G}^t]||_F^2
\end{aligned}
$$

$$
\begin{aligned}
\leq & \frac{\mu}{n}\mathbb{E}||B^t||_F^2 + 8\sigma^2 + \frac{2}{n(1-\mu)}\mathbb{E}||\mathbb{E}[G^t - \bar{G}^t]||_F^2 \\
& + \frac{2}{n\eta^2(1-\mu)}\mathbb{E}||(I - W(2W - I))X^t||_F^2 \\
= & \frac{1-(1-\mu)}{n}\mathbb{E}||B^t||_F^2 + 8\sigma^2 + \frac{2}{n(1-\mu)}\mathbb{E}||\mathbb{E}[G^t - \bar{G}^t]||_F^2 \\
& + \frac{2}{n\eta^2(1-\mu)}\mathbb{E}||(2W + I)(I - W)X^t||_F^2 \\
\overset{(b)}{\leq} & \frac{1-(1-\mu)}{n}\mathbb{E}||B^t||_F^2 + 8\sigma^2 + \frac{2}{n(1-\mu)}\mathbb{E}||\mathbb{E}[G^t - \bar{G}^t]||_F^2 \\
& + \frac{18}{n\eta^2(1-\mu)}\mathbb{E}||(I - W)X^t||_F^2 \\
= & \frac{1-(1-\mu)}{n}\mathbb{E}||B^t||_F^2 + 8\sigma^2 + \frac{2}{n(1-\mu)}\mathbb{E}||\mathbb{E}[G^t - \bar{G}^t]||_F^2 \\
& + \frac{18}{n\eta^2(1-\mu)}\mathbb{E}||(I - W)(X^t - \bar{X}^t)||_F^2 \\
\leq & \frac{1-(1-\mu)}{n}\mathbb{E}||B^t||_F^2 + 8\sigma^2 + \frac{36}{n\eta^2(1-\mu)}\mathbb{E}||X^t - \bar{X}^)||_F^2 \\
& + \frac{2}{n(1-\mu)}\mathbb{E}||\mathbb{E}[G^t] \pm \nabla f(\bar{x}^t) - \mathbb{E}[\bar{G}^t]||_F^2 \\
\leq & \frac{1-(1-\mu)}{n}\mathbb{E}||B^t||_F^2 + 8\sigma^2 + \frac{36}{n\eta^2(1-\mu)}\mathbb{E}||X^t - \bar{X}^)||_F^2 \\
& + \frac{8\zeta^2}{1-\mu} + \frac{4L^2}{n(1-\mu)}\mathbb{E}||X^t - \bar{X}^)||_F^2 \\
= & \frac{1-(1-\mu)}{n}\mathbb{E}||B^t||_F^2 + 8\sigma^2 + \frac{4(9+\eta^2L^2)}{n\eta^2(1-\mu)}\mathbb{E}||X^t - \bar{X}^)||_F^2 + \frac{8\zeta^2}{1-\mu}
\end{aligned}
$$

两边同时乘以 $\frac{6\eta^2\mu^2}{\rho\ 1-\mu}$ () ——

$$
\begin{aligned}
\frac{6\eta^2\mu^2}{n\rho(1-\mu)}\mathbb{E}||B^{t+1}||_F^2 \leq & \Big(\frac{6\eta^2\mu^2}{n\rho(1-\mu)} - \frac{6\eta^2\mu^2}{n\rho}\Big)\mathbb{E}||B^t||_F^2 + \frac{24(9+\eta^2L^2)\mu^2}{n\rho(1-\mu)^2}\mathbb{E}||X^t - \bar{X}^)||_F^2 \\
& + \frac{48\eta^2\mu^2\sigma^2}{\rho(1-\mu)} + \frac{48\eta^2\mu^2\zeta^2}{\rho(1-\mu)^2} \\
\overset{(c)}{\leq} & \Big(\frac{6\eta^2\mu^2}{n\rho(1-\mu)} - \frac{6\eta^2\mu^2}{n\rho}\Big)\mathbb{E}||B^t||_F^2 + \frac{\rho}{8n}\mathbb{E}||X^t - \bar{X}^)||_F^2 \\
& + \frac{\eta^2\rho\sigma^2(1-\mu)}{8} + \frac{\eta^2\rho\zeta^2}{8}
\end{aligned}
$$

注意, $W - I < I$、 $I - W < 2I$、$(W - I)\bar{X}^t = 0$ 和$(W - I)\bar{G}^t = 0$。(a) 可由以下事实得出: $||a + b||^2 \leq (1 + \alpha)||a||^2 + (1 + \frac{1}{\alpha})||b||^2 \ \forall \alpha > 0$, 并令 $\alpha = \frac{1-\mu}{\mu}$ 。(b) 利用了以下事实: $||AB||_F^2 \leq \sigma_{max}^2(A)||B||_F^2$, 其中 $A = 2W + I$、 $B = (I - W)X^t$ 以及 $\sigma_{max}^2(A) = 9$。(c) 利用了假设 $\frac{\mu}{1-\mu} \leq \frac{\rho}{42}$ 和 $\eta \leq \frac{\rho}{7L}$。这意味着 $\frac{24(9+\eta^2L^2\ \mu^2)()^2}{\rho\ 1-\mu} \leq \frac{\rho}{8}$ 以及 $\frac{48\mu^2}{\rho\ 1-\mu}()^2 \leq \frac{\rho}{8}$ 。 – ——— – □

我们使用引理2、3和4给出定理1的证明。将引理3和4相加并化简，可得

$$\frac{24L^2\eta}{n}\mathbb{E}||X^{t+1}-\bar{X}^{t+1}||_F^2+\frac{144L^2\eta^3\mu^2}{n\rho^2(1-\mu)}\mathbb{E}||B^{t+1}||_F^2\leq\Big(\frac{24L^2\eta}{n}-\frac{3L^2\eta}{n}\Big)\mathbb{E}||X^t-\bar{X}^t||_F^2 + \frac{144L^2\eta^3\mu^2}{n\rho^2(1-\mu)}\mathbb{E}||B^t||_F^2+\frac{312L^2\eta^3}{\rho^2}(\zeta^2+\sigma^2(2-\mu)) \tag{17}$$

最后，定义如下函数 Φ^t

$$\Phi^t=\frac{24L^2\eta}{n}\mathbb{E}||X^t-\bar{X}^t||_F^2+\frac{144L^2\eta^3\mu^2}{n\rho^2(1-\mu)}\mathbb{E}||B^t||_F^2+\mathbb{E}[f(\bar{x}^t)-f^*] \tag{18}$$

现在将引理2与（17）相加，可得:

$$\begin{aligned}\Phi^{t+1}\leq&\Phi^t-\frac{\eta}{4}\mathbb{E}||\frac{1}{n}\sum_{i=1}^n\nabla f_i(x_i^t)||^2-\frac{\eta}{4}\mathbb{E}||\nabla f(\bar{x}^t)||^2+\frac{L\eta^2\sigma^2}{n}+\frac{312L^2\eta^3}{\rho^2}(\zeta^2+\sigma^2(2-\mu))\\ \leq&\Phi^t-\frac{\eta}{4}\mathbb{E}||\nabla f(\bar{x}^t)||^2+\frac{L\eta^2\sigma^2}{n}+\frac{312L^2\eta^3}{\rho^2}(\zeta^2+\sigma^2(2-\mu))\\ \Longrightarrow\frac{\eta}{4}\mathbb{E}||\nabla f(\bar{x}^t)||^2\leq&(\Phi^t-\Phi^{t+1})+\frac{L\eta^2\sigma^2}{n}+\frac{312L^2\eta^3}{\rho^2}(\zeta^2+\sigma^2(2-\mu))\end{aligned}$$

对 t 求和

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\nabla f(\bar{x}^t)||^2\leq\frac{4}{\eta T}(f(\bar{x}^0-f^*)+\eta\frac{4L\sigma^2}{n}+\eta^2\frac{1248L^2}{\rho^2}(\zeta^2+\sigma^2(2-\mu)). \tag{19}$$

至此，定理1的证明完成。

A.4 推论1的证明

在定理1的证明中，我们假设学习率 η 和缩放因子 μ满足以下约束:

$$\begin{aligned}(i)\ \ &\eta\leq\min\Big\{\frac{1}{4L},\frac{\rho}{7L}\Big\}\\ (ii)\ \ &\frac{\mu}{1-\mu}\leq\frac{\rho}{42}.\end{aligned}$$

我们假设步长 η为 $\mathcal{O}(\sqrt{\frac{n}{T}})$，其中 n为智能体总数， T为迭代次数。在此假设下，定理1中的（19）式各项的阶如下。

$$\frac{4}{\eta T}(f(\bar{x}^0-f^*)=\mathcal{O}\Big(\frac{1}{\sqrt{nT}}\Big).$$

对于其余各项，我们有:

$$\eta\frac{4L\sigma^2}{n}=\mathcal{O}\Big(\frac{1}{\sqrt{nT}}\Big),\quad\eta^2\frac{1248L^2}{\rho^2}(\zeta^2+\sigma^2(2-\mu))=\mathcal{O}\Big(\frac{n}{T}\Big).$$

因此，在高阶项的语境下省略常数 n 后，存在一个常数 $C>0$，使得整体收敛速度如下：

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\nabla f(\bar{x}^t)||^2 \leq C\left(\frac{1}{\sqrt{nT}}+\frac{1}{T}\right),$$

这表明，当 T 足够大时，*GUT* 可实现 $\mathcal{O}(\frac{1}{\sqrt{nT}})$的收敛速度。

B 算法细节

在本节中，我们给出算法2中的伪代码，用于介绍将所提出的*GUT*算法与准全局动量相结合的*QG-GUTm*；其 PyTorch 实现版本见算法3。我们还在算法4中总结了*GUT* 的内存高效实现。

算法2带动量的全局更新跟踪（*QG-GUTm*）

输入： 每个智能体 $i\in[1,n]$ 初始化模型权重 $x_i^{(0)}$ 和邻居的副本 $\hat{x}_j^{(0)}$、动量缓冲区 $m_i^{(0)}$、步长 η、动量系数 β、混合矩阵 $W=[w_{ij}]_{i,\ j\in[1,n]}$，*GUT* 缩放因子 μ，I_{ij} 是 $n\times n$ 单位矩阵的元素，$\mathcal{N}(i)$表示 i 的邻居（包括其自身），并注意 $\hat{x}_i^t=x_i^t$。

每个智能体同时执行 TRAIN( ) 程序 1
p过程dure TRAIN( )
for t=$0,1,\ldots,T-1$ **do**
  $d_i^t\sim D_i$
  $g_i^t=\nabla_x f_i(d_i^t;\sum_{j\in\mathcal{N}(i)}w_{ij}*\hat{x}_j^t)$
  $\delta_i^t=g_i^t-\frac{1}{\eta}\sum_{j\in\mathcal{N}(i)}(w_{ij}-I_{ij})*\hat{x}_j^t$
  $y_i^t=\delta_i^t+\mu\big[\sum_{j\in\mathcal{N}(i)}w_{ij}(m_j^{t-1}-\frac{1}{\eta}(\hat{x}_j^t-x_i^t))-\delta_i^{t-1}\big]$
  $m_i^t=\beta m_i^{t-1}+(1-\beta)y_i^t$
  SENDRECEIVE(m_i^t)
  $x_i^{t+1}=x_i^t-\eta m_i^t$
  $\hat{x}_j^{t+1}=\hat{x}_j^t-\eta m_j^t\ \ \forall\ \ j\in N(i)\backslash i$
end
return

C 去中心化学习设置

对于去中心化设置，我们采用具有均匀混合矩阵的无向环、无向 Dyck 图和无向环面图拓扑。对于任意图规模的无向环拓扑，每个智能体有 3 个对等节点（包括自身），且每条边的权重为 $\frac{1}{3}$。对于包含 32 个智能体的无向 Dyck 图拓扑，每个智能体有 4 个对等节点（包括自身），且每条边的权重为 $\frac{1}{4}$。对于包含 32 个智能体的无向环面图拓扑，每个智能体有 5 个对等节点（包括自身），且每条边的权重为 $\frac{1}{5}$。除 Imagenette 仿真实验外，我们的所有实验均在配备 Nvidia GTX 1080ti 显卡和 4 个 GPU 的系统上进行。Imagenette 仿真实验使用了配备 4 个 GPU 的 NVIDIA A40 显卡。

C.1 数据集

本节简要介绍我们实验中使用的数据集。我们使用一组多样化的数据集，每个数据集均源自不同的图像分布，以展示所提出技术的泛化能力。

CIFAR-10： CIFAR-10 [13]是一个包含 10 个类别的图像分类数据集。图像样本为彩色（3 个输入通道），分辨率为 32×32。共有 50,000 个训练样本，每个类别 5000 个样本，以及 10,000 个测试样本，每个类别 1000 个样本。

算法 3带动量的全局更新跟踪（*QG-GUTm*）– Pytorch 实现**输入：** 每个智能体 $i \in [1,n]$ 初始化模型权重 $x_i^{(0)}$ 和邻居的副本 $\hat{x}_j^{(0)}$、动量缓冲区 $m_i^{(0)}$、步长 η、动量系数 β、混合矩阵 $W=[w_{ij}]_{i,j\in[1,n]}$，*GUT* 缩放因子 μ，I_{ij} 是 $n\times n$ 单位矩阵的元素，$\mathcal{N}(i)$表示 i 的邻居（包括其自身），并记 $\hat{x}_i^t = x_i^t$。

每个智能体同时执行 TRAIN( ) 过程 1

过程 dure TRAIN( )
for t=0, 1, . . . , $T-1$ **do**
  $d_i^t \sim D_i$
  $g_i^t = \nabla_x f_i(d_i^t; \sum_{j\in\mathcal{N}(i)} w_{ij} * \hat{x}_j^t)$
  $\delta_i^t = g_i^t - \frac{1}{\eta}\sum_{j\in\mathcal{N}(i)}(w_{ij} - I_{ij}) * \hat{x}_j^t$
  $y_i^t = \delta_i^t + \mu\big[\sum_{j\in\mathcal{N}(i)} w_{ij}(m_j^{t-1} - \frac{1+\beta}{\eta}(\hat{x}_j^t - x_i^t)) - \delta_i^{t-1}\big]$
  $m_i^t = \beta m_i^{t-1} + y_i^t$
  SENDRECEIVE(m_i^t)
  $x_i^{t+1} = x_i^t - \eta m_i^t$
  $\hat{x}_j^{t+1} = \hat{x}_j^t - \eta m_j^t \ \ \forall \ \ j \in N(i)\backslash i$
end
return

算法4全局更新跟踪（内存高效实现）

输入： 每个智能体 $i \in [1,n]$ 初始化模型参数 x_i^0和邻域的加权模型参数 s_i^0、步长 η、*GUT*缩放因子 μ、混合矩阵 $W=[w_{ij}]_{i,\ j\in[1,n]}$, $\mathcal{N}(i)$ 表示 i 的邻居（包括其自身），并记为 $\hat{x}_i^t = x_i^t$。

每个智能体同时执行 TRAIN( ) 过程 1

Proce dure TRAIN( )
for t = $0, 1, \ldots, T-1$ **do**
  $d_i^t \sim D_i$
  $g_i^t = \nabla_x F_i(s_i^t; d_i^t)$
  $\delta_i^t = g_i^t - \frac{1}{\eta}(s_i^t - x_i^t)$
  $y_i^t = \delta_i^t + \mu\Big[\sum_{j\in\mathcal{N}(i)} w_{ij} y_j^{t-1} - \frac{1}{\eta}(s_i^t - x_i^t) - \delta_i^{t-1}\Big]$
  SENDRECEIVE(y_i^t)
  $x_i^{t+1} = x_i^t - \eta y_i^t$
  $s_i^{t+1} = s_i^t - \eta\sum_{j\in\mathcal{N}(i)} w_{ij} y_j$
end
return $\frac{1}{n}\sum_{i=1}^n x_i^T$

CIFAR-100： CIFAR-100 [13] 是一个包含 100 个类别的图像分类数据集。图像样本为彩色图像（3 个输入通道），分辨率为 32×32。该数据集包含 50,000 个训练样本，每个类别有 500 个样本，以及 10,000 个测试样本，每个类别有 100 个样本。与 CIFAR-10 相比，CIFAR-100 分类是一项更具挑战性的任务，因为它包含 100 个类别，且每个类别可用于学习的样本数量非常少。

Fashion MNIST： Fashion MNIST [25] 是一个包含 10 个类别的图像分类数据集。图像样本为灰度图（1 个输入通道），分辨率为 28×28。共有 60，000 个训练样本，每个类别有 6000 个样本；另有 10，000 个测试样本，每个类别有 1000 个样本。

Imagenette： Imagenette [10] 是 ImageNet 数据集的一个包含 10 个类别的子集。图像样本为彩色图像（3 个输入通道），分辨率为 224×224。训练样本共 9469 个，每个类别约有 950 个样本；测试样本共 3925 个。

C.2 网络架构

我们将所有模型架构中的 ReLU+BatchNorm 层替换为 EvoNorm-S0，因为已有研究表明，它更适用于非独立同分布数据分布下的去中心化学习。

VGG-11： 我们通过将每个卷积层中的滤波器数量减少 $4\times$，并且只使用一个包含 128 个单元的全连接层，对标准 VGG-11 [21] 架构进行了修改。每个卷积层后都接有 EvoNorm-S0，作为激活归一化层。VGG-11 有 $0.58M$ 个可训练参数。

ResNet-20： 对于 ResNet-20 [8], ，我们采用标准架构，其具有 $0.27M$ 个可训练参数，但将 BatchNorm+ReLU 层替换为 EvoNorm-S0。

LeNet-5： 对于 LeNet-5 [14], ，我们采用具有 61,706 个可训练参数的标准架构。 .

MobileNet-V2： 我们使用用于 CIFAR 数据集的标准 MobileNet-V2 [20] 架构，参数为 $2.3M$，但将 BatchNorm+ReLU 层替换为 EvoNorm-S0。

C.3 超参数

本节详细介绍了我们实验中使用的超参数。所有实验均使用随机选取的三个种子运行。除非另有说明，我们在训练进行到 50% 和 75% 后，将步长缩小 10 倍。对于所有使用 *GUT* 优化器的实验，超参数 μ 均设为 0.9。我们使用网格搜索为*QG-GUTm*选择超参数 μ 。

表1中实验所用的超参数： 所有实验的停止条件均设置为200个训练轮次。初始学习率设置为0.1。我们在 100th 和 150th 训练轮次时分多个阶段将步长缩小 $10\times$ 。表6列出了实验中使用的缩放因子 μ 的取值。对于所有实验，每个智能体使用的小批量大小均为32。停止条件为固定的训练轮次数。所有QG-DSDm和*QG-GUTm* 实验均使用了0.9的动量。

表6：表1中所示使用 ResNet-20 和 VGG-11 模型架构，在非独立同分布数据上训练 CIFAR-10 时所用的缩放因子 μ 的值

智能体（n）	方法	ResNet-20		
		$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.01$
16	DSGD	0.0	0.0	0.0
	GUT（我们的方法）	0.9	0.9	0.9
	QG-DSGDm	0.0	0.0	0.0
	QG-GUTm（我们的方法）	0.04	0.06	0.04
32	DSGD	0.0	0.0	0.0
	GUT（我们的方法）	0.9	0.9	0.9
	QG-DSGDm	0.0	0.0	0.0
	QG-GUTm（我们的方法）	0.04	0.04	0.04
智能体（n）	方法	VGG-11		
		$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.01$
16	DSGD	0.0	0.0	0.0
	GUT（我们的方法）	0.9	0.9	0.9
	QG-DSGDm	0.0	0.0	0.0
	QG-GUTm（我们的方法）	0.06	0.08	0.09
32	DSGD	0.0	0.0	0.0
	GUT（我们的方法）	0.9	0.9	0.9
	QG-DSGDm	0.0	0.0	0.0
	QG-GUTm（我们的方法）	0.08	0.08	0.08

实验的超参数（见表 2）： 所有实验的停止条件均设为 200 个训练轮次。初始学习率设为 0.1。我们在第 100th 和第 150th 个训练轮次将步长分阶段衰减 $10\times$ 。表 7 列出了实验中使用的缩放因子 μ 的取值。所有实验中，每个智能体的小批量大小均为 32。停止条件为固定的训练轮次数。对于所有 QG-DSDm 和 *QG-GUTm* 实验，我们使用了 0.9 的动量，并将 Nesterov 设为 False。在非独立同分布数据的实验中，我们未使用任何正则化

，即权重衰减设为零。对于独立同分布数据的实验（表 1 中的 DSGDm 情况），我们将权重衰减设为 $1e^{-4}$。

表 7：在表 2 所示的不同图拓扑上，使用 ResNet-20 模型架构训练非独立同分布 CIFAR-10 数据时所用的缩放因子 μ 值

	Dyck 图（32 个智能体）		环面（32 个智能体）	
	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.01$
QG-DSGDm	0.0	0.0	0.0	0.0
QG-GUTm	0.05	0.05	0.05	0.05

表 3 中实验的超参数：所有使用 Fashion-MNIST 和 Imagenette 数据集的实验，停止条件均设为 100 个训练轮次；而 CIFAR-100 实验的停止条件设为 200 个训练轮次。
CIFAR-100 和 Fashion MNIST 数据集实验的初始学习率设为 0.1。Imagenette 数据集实验的初始学习率设为 0.01。我们在 50th 和 75th 训练轮次处，通过 $10\times$ 多次衰减步长。表 8 列出了实验中使用的缩放因子 μ 的取值。所有实验中，每个智能体的小批量大小均为 32。停止条件为固定的训练轮次数。所有 QG-DSDm 和*QG-GUTm* 实验均使用 0.9 的动量。

表 8：表 3 中所示的 16智能体环形拓扑上用于训练不同数据集的缩放因子 μ 值

方法	Fashion MNIST (LeNet-5)		CIFAR-100（ResNet-20）		Imagenette (MobileNet-V2)	
	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.01$
QG-DSGDm	0.0	0.0	0.0	0.0	0.0	0.0
QG-GUTm	0.01	0.005	0.005	0.005	0.03	0.04

D 附加结果

表 9 显示，即使异质性程度更高（即 $\alpha = 0.01$），全局更新跟踪的准全局变体仍优于其他现有方法。我们还观察到，当数据分布的异质性较高时，Nesterov 动量会降低性能。表 10 将 DSGDm 基线与所提出的 全局更新跟踪 和 *QG-GUTm*进行了比较。结果表明，当异质性程度较高时，全局更新跟踪 甚至优于带动量的 DSGD。

表 9：使用 CIFAR-10 数据集和 ResNet-20 架构，在 16智能体环形拓扑上训练，评估全局更新跟踪（GUT）的各种动量版本，$\alpha = 0.01$。

方法	本地动量	Nesterov	准全局动量	全局更新跟踪	测试准确率 $\alpha = 0.01$
DSGD	x	x	x	x	54.66 ± 4.74
DSGDm	✓	x	x	x	65.62 ± 4.95
DSGDm-N	✓	✓	x	x	63.66 ± 4.44
QG-DSGDm	x	x	✓	x	79.85 ± 2.13
QG-DSGDm-N	x	✓	✓	x	78.64 ± 2.14
GUT	x	x	x	✓	70.16 ± 4.94
GUTm	✓	x	x	✓	64.25 ± 5.31
GUTm-N	✓	✓	x	✓	63.42 ± 2.74
QG-GUTm	x	x	✓	✓	$\mathbf{81.04} \pm 1.66$
QG-GUTm-N	x	✓	✓	✓	80.09 ± 3.82

表10：在环拓扑上，对采用不同异质性程度（非独立同分布）分布的 CIFAR-10 数据集进行评估时，不同去中心化算法在各种模型上的平均测试准确率。

智能体（n）	方法	ResNet-20		
		$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.01$
16	*GUT*（本文方法）	84.72 ± 0.20	81.86 ± 1.99	70.16 ± 4.94
	DSGDm	86.60 ± 0.54	79.87 ± 1.73	65.62 ± 4.95
	QG-GUTm（我们的方法）	$\mathbf{88.22} \pm 0.36$	$\mathbf{86.44} \pm 0.36$	$\mathbf{81.04} \pm 1.66$
32	*GUT*（我们的方法）	79.24 ± 0.33	76.07 ± 0.23	60.72 ± 1.03
	DSGDm	86.12 ± 0.32	77.43 ± 1.71	52.82 ± 4.02
	QG-GUTm（我们的方法）	$\mathbf{87.48} \pm 0.33$	$\mathbf{84.94} \pm 0.60$	$\mathbf{72.04} \pm 3.18$
智能体 (n)	方法	VGG-11		
		$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.01$
16	*GUT*（我们的方法）	82.12 ± 0.09	81.24 ± 0.95	76.62 ± 1.37
	DSGDm	81.77 ± 0.38	74.20 ± 1.89	58.44 ± 14.58
	QG-GUTm（我们的方法）	$\mathbf{84.46} \pm 0.33$	$\mathbf{83.05} \pm 0.48$	$\mathbf{78.32} \pm 1.03$
32	*GUT*（我们的方法）	80.37 ± 0.33	79.55 ± 1.00	73.59 ± 1.26
	DSGDm	81.89 ± 0.29	74.73 ± 0.73	61.60 ± 2.80
	QG-GUTm（我们的方法）	$\mathbf{84.32} \pm 0.11$	$\mathbf{83.39} \pm 0.38$	$\mathbf{77.41} \pm 3.44$