arXiv:2302.01537v1 [math.OC] 3 Feb 2023

用于去中心化非凸学习的多次本地随机梯度下降与梯度、变量跟踪

Songyang Ge 和 Tsung-Hui Chang *†

2023年2月6日

摘要

在多智能体网络上求解优化问题的随机分布式优化方法，在各种大规模信号处理和机器学习应用中发挥着重要作用。在现有方法中，与智能体拥有异构数据分布时收敛速度会减慢的分布式随机梯度下降（SGD）方法相比，梯度跟踪（GT）方法能够有效应对智能体本地数据分布之间的差异。然而，GT 方法需要大量迭代才能收敛，因此通信成本可能较高。本文旨在通过结合本地 SGD 技术，降低 GT 方法的通信成本。具体而言，我们提出一种新的本地随机梯度跟踪（LSGT）算法：在每个通信轮次内，智能体在本地执行多次 SGD 更新。从理论上，我们建立了 LSGT 算法的收敛条件，并证明其可达到改进的收敛速度 $\mathcal{O}(1/\sqrt{ET})$，其中 E 为本地 SGD 更新次数，T 为通信轮次。我们进一步将 LSGT 算法扩展到一种更复杂的学习问题，该问题的目标函数中包含线性耦合变量。实验结果表明，即使在数据分布异构的情况下，所提出的算法也能显著提升收敛速度。

1 引言

随着人类社会数字化进程的持续加速，海量数据正呈爆炸式增长。这一现象使得难以将所有数据存储在一台设备中并由一个处理器进行处理。鉴于此，分布式优化方法在信号处理和机器学习等多个领域引起了广泛关注，参见例如 [1–7]。特别地，考虑一个由 N 个智能体通过连通图相连的多智能体网络，现有的大多数分布式优化方法都致力于求解以下问题

$$\min_{\boldsymbol{y}\in\mathbb{R}^p} F(\boldsymbol{y}) \triangleq \frac{1}{N}\sum_{n=1}^{N} f_n(\boldsymbol{y}), \tag{1}$$

* Songyang Ge 和 Tsung-Hui Chang 任职于深圳大数据研究院、香港中文大学（深圳）理工学院（中国深圳，邮编 518172）。电子邮箱：songyangge@link.cuhk.edu.cn；tsunghui.chang@ieee.org

† Tsung-Hui Chang 是通讯作者。

表 1： 不同算法的比较

算法	函数	梯度	步长	计算	通信	λ_w	$\boldsymbol{W}$
梯度跟踪 [11]	非光滑凸。	full	cst.	$\mathcal{O}(\frac{1}{\epsilon})$	$\mathcal{O}(\frac{1}{\epsilon})$	$(-1,\ 1)$	ds.
DSGT [12]	cvx.	随机	dimi.	$\mathcal{O}(\frac{1}{\epsilon})$	$\mathcal{O}(\frac{1}{\epsilon})$	$(-1,\ 1)$	ds.
GNSD [13]	非凸。	随机	cst.	$\mathcal{O}(\frac{1}{\epsilon^2})$	$\mathcal{O}(\frac{1}{\epsilon^2})$	$(-1,\ 1)$	分布式且对称。
GT-DSGD [14]	非凸	随机,	cst.	$\mathcal{O}(\frac{1}{\epsilon^2})$	$\mathcal{O}(\frac{1}{\epsilon^2})$	$(-1,\ 1)$	ds.
D^2 [15]	非凸。	随机	cst.	$\mathcal{O}(\frac{1}{\epsilon^2})$	$\mathcal{O}(\frac{1}{\epsilon^2})$	$(-\frac{1}{3},\ 1)$	随机响应。和对称。
LU-GT [16]	非凸	E步 全梯度	cst.	✗	✗	$(-1,\ 1)$	确定性采样。和对称。
LSGT（提出的方法）	非凸	E步随机	cst.	$\mathcal{O}(\frac{1}{\epsilon^2})$	$\mathcal{O}(\frac{1}{E\epsilon^2})$	$(-1,\ 1)$	ds.
MUST（提出的方法）	非凸	E步随机	cst.	$\mathcal{O}(\frac{1}{\epsilon^2})$	$\mathcal{O}(\frac{1}{E\epsilon^2})$	$(-1,\ 1)$	ds.

“cvx.”、“n-cvx.”和“ns-cvx.”分别表示凸、非凸和非强凸；“comp.”表示计算复杂度，“comm.”表示通信复杂度，“cst.”和“dimi.”分别是常数和递减的缩写，“ds.”表示双随机，“rs.”表示右随机，“sym.”表示对称，$\boldsymbol{W}$ 是混合矩阵，λ_w 是 $\boldsymbol{W}$的第二大特征值，ϵ 是解精度，E 是本地更新次数。

其中，$\boldsymbol{y} \in \mathbb{R}^p$ 是待优化的参数向量，每个 $f_n: \mathbb{R}^p \to \mathbb{R}$ 都是智能体 n的光滑且可能非凸的本地成本函数，$\forall n \in [N] \triangleq \{1, \ldots, N\}$。对于统计学习问题，可以假设 $f_n(\boldsymbol{y}) = \mathbb{E}_{\xi \sim \mathcal{D}_n}[\ell_n(\boldsymbol{y}, \xi)]$ 其中，ℓ_n 是 $\boldsymbol{y}$的损失函数，数据样本 ξ 是从本地数据集 $\mathcal{D}_n$中随机抽取的。

1.1 文献综述

文献中已有大量关于求解非凸分布式问题（1）的研究成果 [4,6,7]。考虑全梯度 ∇f_n时，已有一些确定性一阶方法被提出。去中心化梯度下降（DGD）算法 [8] 通过适当的混合矩阵和递减步长，采用共识梯度下降（GD）来求得问题（1）的驻点解。基于原始-对偶的方法，例如近端梯度原始-对偶算法（Prox-GPDA） [9] ，可以在常数步长下收敛到驻点解。然而，Prox-GPDA 只有在星形网络和全连接网络中才能达到通信和计算下界。对于一般网络拓扑，文献 [10]提出了一种近乎最优的方案 xFilter。然而，由于使用了全梯度，上述方法仅适用于数据集规模较小至中等的学习问题。

基于小批量随机梯度的随机分布式方法已被用于降低梯度计算的复杂度。一种随机变体

DGD方法，即去中心化随机梯度下降（DSGD）算法，在 [8, 17], 中通过用随机梯度替代全梯度而提出。除了需要采用递减步长外， [8, 15] 中的研究发现，DSGD类方法对智能体数据分布的方差较为敏感。特别是，当各智能体的数据分布存在异质性，即所谓的非独立同分布数据时，DSGD算法的收敛性能会显著下降。为缓解这一问题， [15] 的作者提出了一种名为D^2的新算法。尽管D^2 对异质数据更具鲁棒性，但其收敛依赖于对混合矩阵的严格假设（见 [4]和第2.1节）。基于原始-对偶的随机分布式方法（例如 [18], 中的带动量随机近端原始-对偶算法（SPPDM））对数据分布也表现出更好的鲁棒性，但涉及更多可调参数。

近年来，梯度跟踪（GT）方法 [11, 19] 已被提出，成为求解问题（1）的一类颇具前景的分布式方法。具体而言，与 D^2 相比，梯度跟踪方法对混合矩阵的要求更宽松，且只涉及一个步长参数，便可收敛至驻点解的邻域。梯度跟踪方法基于一个简单而有效的思路：引入辅助变量，帮助分布式智能体通过共识平均跟踪全局梯度。这样，智能体便能模拟集中式GD方案，因此比现有方法具有更强的鲁棒性。基于随机梯度下降的梯度跟踪方法也已得到研究；例如， [12] 针对强凸问题，而 GNSD [13] 和 GT-DSGD [20]则针对非凸问题。然而，这些随机 GT 方法要求智能体之间频繁进行消息交换。由于通信成本通常很高，且受到有限带宽的限制，因此在通信资源有限或时延约束严格的应用中，实现上述算法可能并不容易。

因此，降低随机 GT 方法的通信成本迫在眉睫。一种可行的方式是允许智能体在每个通信轮次内执行多次本地 SGD 更新 [21, 22], ，而不是像现有 GT 方法那样仅执行一次 SGD 更新。这种局部随机梯度下降技术已在联邦学习场景中得到成熟且成功的应用 [23], 在这些场景中，分布式智能体由中心服务器协调。具体而言，理论和实证研究 [21, 22] 均表明，在适当条件下，本地 SGD 可将通信成本降低 E 倍，其中 E 表示每个通信轮次中的本地 SGD 更新次数。然而，针对完全去中心化网络中的 GT 方法，局部随机梯度下降技术尚未得到深入研究。例如，尽管近期工作 [16] 提出了一种 GT 方法，允许智能体在每个通信轮次中执行多次本地更新，但该方法考虑的是全梯度，而非随机梯度。此外，其中的收敛性分析未能充分刻画局部更新对收敛速度的影响，而且其算法还需要额外的步长和对称混合矩阵才能保证适当收敛。

1.2 贡献

本文通过将局部随机梯度下降技术与随机 GT 方法相结合，提出了一种新的分布式算法，称为本地随机梯度跟踪（LSGT）算法。LSGT 算法既不引入额外的步长，也不要求混合矩阵为对称矩阵。本文的主要贡献是进行收敛性分析，证明采用常数步长的 LSGT 算法能够以次线性速度收敛到问题（1）的驻点解邻域。此外，我们还给出了使 LSGT

算法能够受益于局部随机梯度下降技术带来的收敛加速的条件，这一点在以往的研究中从未得到证明。表 1 总结了所提出的算法与现有方法之间的比较结果。可以看出，所提出的 LSGT 算法的通信复杂度为 $\mathcal{O}(\frac{1}{E\epsilon^2})$，而其他方法的通信复杂度为 $\mathcal{O}(\frac{1}{\epsilon^2})$。

作为第二项贡献，我们将 LSGT 算法的思想扩展到另一类优化问题，在这类问题中，智能体的局部变量在线性耦合的形式下出现在目标函数中（见（38）和（39））。一个具有代表性的例子是针对混合数据的分布式学习问题：每个分布式智能体仅持有一部分数据样本，且只掌握部分特征信息。针对这一具有挑战性的混合数据学习问题，目前鲜有研究。近期的工作 [24,25] 研究了联邦学习网络中的此类问题，但其算法不适用于去中心化网络。受 GT 方法的类似思路启发，我们引入额外的辅助变量来跟踪线性耦合项，并提出一种名为多局部更新变量和跟踪（*MUST*）算法的新算法，以求解（38）-（39）中的学习问题。

最后，我们通过大量数值实验评估了所提出的 LSGT 和 MUST 算法的性能。

概要： 第 2 节详细介绍网络模型和假设。为求解问题（1），第 3 节提出 LSGT 算法并给出其收敛性分析。第 4 节展示 LSGT 算法主要理论结果的证明细节。此外，第 5 节探讨了 LSGT 算法在混合数据学习问题上的扩展，即 MUST 算法。第 6 节给出数值结果，第 7 节总结全文。

记号： $\boldsymbol{I}_n$ 是一个 $n \times n$ 的单位矩阵，且$\mathbf{1}$ 是全 1 向量。$A_{i,j}$ 是矩阵 $\boldsymbol{A}$ 的第（i,j）个元素。 $\otimes$ 表示克罗内克积； $\boldsymbol{a}^\top$ 和 $\boldsymbol{A}^\top$ 分别表示向量 $\boldsymbol{a}$ 和矩阵 $\boldsymbol{A}$的转置运算； $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ 表示向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$的内积， $\|\boldsymbol{a}\|$ 是欧几里得范数， $\|\boldsymbol{A}\|$ 表示 $\boldsymbol{A}$的最大奇异值； $\|\boldsymbol{A}\|_F$ 表示矩阵 Frobenius 范数。

2 网络模型与假设

本节介绍问题（1）的网络模型和假设。

2.1 网络模型与假设

我们将多智能体网络建模为无向图 $\mathcal{G} = (\mathcal{E}, \mathcal{V})$，其中 $\mathcal{E}$ 是边的集合， $\mathcal{V} = [N] = \{1, \ldots, N\}$ 是智能体的集合。每个智能体拥有本地数据集 $\mathcal{D}_n$，且只能与其邻居通信和交换信息；也就是说，当且仅当（n, m） $\in \mathcal{E}$时，智能体 n 和 m才能彼此通信。此外，我们有以下标准假设。

假设 1 底层图是连通的。

为实现智能体之间的消息交换，我们定义混合矩阵 $\boldsymbol{W} \in \mathbb{R}^{N \times N}$，其中$W_{n,\ m} > 0, \forall (n,\ m) \in \mathcal{E}$ 且 $W_{n,\ m} = 0$ 否则。此外，我们有如下假设。

假设2 混合矩阵 $\boldsymbol{W}$ 是满足以下条件的双随机矩阵

$$\boldsymbol{W}\mathbf{1} = \mathbf{1}, \quad \mathbf{1}^\top \boldsymbol{W} = \mathbf{1}, \quad |\lambda_w| < 1, \tag{2}$$

其中， $\lambda_w = \|\boldsymbol{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\|$ 是 $\boldsymbol{W}$的第二大特征值。

可以通过多种方式构造假设 2 [13, 备注 2]中的混合矩阵 $\boldsymbol{W}$；例如最大度规则 [26]。请注意，我们对 $\boldsymbol{W}$ 的假设比 [15] 中的 D^2算法更弱，后者要求$\lambda_w \in \left(-\frac{1}{3}, 1\right)$。此外，与 [13,15,16]不同，我们不假设 $\boldsymbol{W}$是对称的。

2.2 问题假设

为了便于算法开发和收敛性分析，我们对问题 (1) 作出如下假设。

假设3 损失函数 $F(\boldsymbol{y}) \triangleq \frac{1}{N}\sum_{n=1}^{N} f_n(\boldsymbol{y})$ 有下界，记为 $F(\boldsymbol{y}) \geq F,\ \forall \boldsymbol{y}$。

假设4 每个 f_n 都是光滑的，且其梯度满足

$$\left\|\nabla f_n(\boldsymbol{y}) - \nabla f_n(\boldsymbol{y}')\right\| \leq L\left\|\boldsymbol{y} - \boldsymbol{y}'\right\|,\ \forall \boldsymbol{y}, \boldsymbol{y}', \tag{3}$$

其中 L 是 *Lipschitz*常数。

在实践中，计算全梯度的计算成本可能很高。作为替代方案，通常使用基于小批量数据采样的随机梯度。具体而言，对于每个智能体 n，我们定义如下随机梯度:

$$\boldsymbol{g}_n \triangleq \frac{1}{|\mathcal{I}_n|}\sum_{\xi \in \mathcal{I}_n} \nabla \ell_n(\boldsymbol{y}_n; \xi), \tag{4}$$

其中， $\mathcal{I}_n \subseteq \mathcal{D}_n$ 是智能体 n 处随机选择的小批量数据集。不失一般性地，我们假设所有智能体使用相同的小批量大小，记为 $|\mathcal{I}|$。以下假设对于随机梯度下降方法而言是标准的。

假设5 对于每个智能体 $n \in [N]$，有
- 无偏梯度: $\mathbb{E}[\boldsymbol{g}_n] = \nabla f_n(\boldsymbol{y}_n)$;
- 一致有界方差: $\mathbb{E}[\|\boldsymbol{g}_n - \nabla f_n(\boldsymbol{y}_n)\|^2] \leq \frac{\sigma^2}{|\mathcal{I}|}$.

3 所提出的 LSGT 算法

本节提出一种新的去中心化算法，称为局部随机梯度跟踪（*LSGT*），用于求解非凸问题 (1)。所提出的 LSGT 算法基于著名的随机 GT 方法 [13, 20]，并在每个通信轮次中纳入多次本地 SGD 更新，以提高收敛速度。我们首先回顾原始 GT 方法，然后介绍所提出的算法及其收敛性分析。

3.1 随机 GT 方法概述

GT 方法是一种基于共识的方法，通过引入额外的辅助变量来跟踪目标函数的全局梯度，从而模拟集中式梯度下降方法[13,20]。具体而言，除局部变量 $\boldsymbol{y}_n$ 外，每个智能体 n 还拥有另一个辅助变量 $\tilde{\boldsymbol{v}}_n \in \mathbb{R}^p$，用于估计全局梯度 $\nabla F(\frac{1}{N}\sum_{n=1}^N \nabla f_n\ \boldsymbol{y}$ $\boldsymbol{y}_n^0$ $\tilde{\boldsymbol{v}}_n^0 = \boldsymbol{g}_n^0$) 的随机近似 ( )。算法以每个智能体 n的初始值和随机梯度开始。然后，在第 r 次迭代中，所有智能体 $n \in [N]$ 并行执行以下两个步骤:

$$\boldsymbol{y}_n^{r+1} = \sum_{m=1}^N W_{n,m}\boldsymbol{y}_m^r - \gamma\tilde{\boldsymbol{v}}_n^r, \tag{5a}$$

$$\tilde{\boldsymbol{v}}_n^{r+1} = \sum_{m=1}^N W_{n,m}\tilde{\boldsymbol{v}}_m^r + \boldsymbol{g}_n^{r+1} - \boldsymbol{g}_n^r, \tag{5b}$$

其中， $\gamma > 0$ 表示步长。在（5a）中，智能体对来自其邻居的局部变量进行共识平均，然后沿着 $\tilde{\boldsymbol{v}}_n$的方向执行梯度下降；在（5b）中，智能体利用共识平均来跟踪随机梯度 $\sum_{n=1}^N \boldsymbol{g}_n^{r+1}$ 之和。梯度跟踪方法背后的思想是，当变量一致性近似达到时，即 $\boldsymbol{y}_n \approx \frac{1}{N}\sum_{m=1}^N \boldsymbol{y}_m$ 和 $\tilde{\boldsymbol{v}}_n \approx \frac{1}{N}\sum_{m=1}^N \tilde{\boldsymbol{v}}_m\ \forall n \in [N]$时， $\underline{\tilde{\boldsymbol{v}}}_n$ 近似于全局梯度∇F（$\boldsymbol{y}_n$），此时（5a）与集中式梯度下降（GD）方法相同。

理论上已经证明，与分布式随机梯度下降算法相比，随机 GT 方法的收敛速度不依赖于各智能体局部代价函数之间的差异。具体而言， [8,15] 中的收敛性分析表明，DSGD 的收敛速度取决于界$\frac{1}{N}\sum_{n=1}^N \|\nabla f_n$（$\boldsymbol{y}$）$-\nabla F$（$\underline{\boldsymbol{y}}$）$\| \le \xi$，并且当 ξ 较大时可能会变慢，而 [13,14] 中的随机 GT 方法则不会受到这种影响。这意味着，随机 GT 方法对于所谓的“异质数据”（即智能体的局部数据集具有不同统计特性）具有更强的鲁棒性。

然而，与 DSGD 方法类似，(5) 中的随机 GT 方法需要许多通信轮次才能收敛，因此通信开销仍然很大。为降低通信成本，我们提出采用局部随机梯度下降技术，即允许智能体在每个通信轮次执行多步随机梯度下降。关于本地 SGD 加速算法收敛的有效性，已有理论和实证研究 [21,22], ，但除近期工作 [16], 外，这种方法尚未在基于梯度跟踪的方法中得到充分研究。

在下一小节中，我们将随机 GT 方法与本地 SGD 技术相结合，提出一种新的通信高效 LSGT 算法，并进一步建立其收敛性质。

3.2 提出的 LSGT 算法

所提出的 LSGT 算法见算法1。与式（5）中的原始随机 GT 方法相比，在所提出的 LSGT 算法中，每个智能体 n 在每个通信轮次内执行 E 次连续的随机梯度下降步骤。具体而言，在每一步 $q \in [E]$中，智能体 n 沿着方向 $\boldsymbol{v}^{r,\,n^{q-1}}$执行梯度下降，如式（7a）所示。然后，它随机选择一个小批量数据集 $\mathcal{I}^{r,\,n^q}$（大小为 $|\mathcal{I}|$），并使用 $\boldsymbol{y}^{r,\,n^q}$计算相应的随机梯度 $\boldsymbol{g}^{r,\,n^q}$。局部辅助变量 $\boldsymbol{v}^{r,\,n^{q-1}}$按照式（7b）在本地更新。最后，经过 E 次本地更新后，每个智能体 n 向其邻居发送 $(\boldsymbol{y}_n^{r+1},\ \boldsymbol{v}_n^{r+1})$。

算法1 用于求解（1）的所提出的 LSGT 算法

1: **初始化：** 令 $\boldsymbol{y}_1^0 = \ldots = \boldsymbol{y}_N^0$ 和 $\boldsymbol{v}_n^0 = \boldsymbol{g}_n^0$, $\forall n \in [N]$。 2: **对于** 通信轮次 $r = 0$ **至** T **执行** 3: **对于** 智能体 $n = 1$ **至** N 并行 **执行** 4: 从邻居接收信息并设置

$$\begin{bmatrix} \boldsymbol{y}_n^{r,0} \\ \boldsymbol{v}_n^{r,0} \end{bmatrix} = \sum_{m=1}^{N} W_{n,m} \begin{bmatrix} \boldsymbol{y}_m^r \\ \boldsymbol{v}_m^r \end{bmatrix}, \ \boldsymbol{g}_n^{r,0} = \boldsymbol{g}_n^r. \tag{6}$$

5: **对于**本地更新 $q = 1, \ldots, E$, **执行** 6:

$$\boldsymbol{y}_n^{r,q} = \boldsymbol{y}_n^{r,q-1} - \gamma \boldsymbol{v}_n^{r,q-1}, \tag{7a}$$

$$\boldsymbol{v}_n^{r,q} = \boldsymbol{v}_n^{r,q-1} + \boldsymbol{g}_n^{r,q} - \boldsymbol{g}_n^{r,q-1}, \tag{7b}$$

其中，随机梯度 $\boldsymbol{g}^{r,n}{}^q$按照（4）计算，使用 $\boldsymbol{y}^{r,n}{}^q$和一个小批量$\mathcal{I}^{r,n}{}^q \subseteq \mathcal{D}_n$。 7: **结束循环**8: 设置 $\boldsymbol{y}_n^{r+1} = \boldsymbol{y}^{r,n}{}^E, \boldsymbol{v}_n^{r+1} = \boldsymbol{v}^{r,n}{}^E, \boldsymbol{g}_n^{r+1} = \boldsymbol{g}^{r,n}{}^E$,并将（$\boldsymbol{y}_n^{r+1}, \boldsymbol{v}_n^{r+1}$）发送给邻居。 9: **结束循环**10: **结束循环**

备注 1（与 *[16]* 的比较）值得注意的是，近期提出的本地更新梯度跟踪（*LU-GT*）算法 *[16]* 也考虑了梯度跟踪方法的多次本地更新。然而，*LU-GT* 算法与算法*1*中所提出的 *LSGT* 算法存在若干区别。首先，*LU-GT* 算法不采用随机梯度下降，而是在智能体处使用全梯度。其次，*LU-GT* 算法要求在更新局部辅助变量时使用额外的步长参数，而我们的方法在（7b）中无需任何额外参数。此外，*LU-GT* 要求混合矩阵 $\boldsymbol{W}$ 为对称矩阵，而我们的 *LSGT* 算法则没有此要求。

第三，也是最重要的一点，*[16]*中的收敛性分析未能充分刻画局部更新对算法收敛的影响。事实上，他们的结果（例如，*[16,*注释 *2]*）在某种程度上暗示，即使网络图连通性良好，局部更新也会减慢算法的收敛。

在下一小节中，我们将为 LSGT 算法提出一种新颖的收敛性分析，揭示 LSGT 算法在何种条件下确实能够受益于本地更新，并随着网络规模 N 和局部更新次数E的增加实现线性加速。

3.3 收敛速度分析

令 $\boldsymbol{Y}^r \triangleq [\boldsymbol{y}_1^r, \ldots, \boldsymbol{y}_N^r]^\top$、 $\boldsymbol{V}^r \triangleq [\boldsymbol{v}_1^r, \ldots, \boldsymbol{v}_N^r]^\top$，并将局部变量的平均值定义为 $\bar{\boldsymbol{y}}^r = \frac{1}{N}\sum_{n=1}^N \boldsymbol{y}_n^r$ 和 $\bar{\boldsymbol{v}}^r = \frac{1}{N}\sum_{n=1}^N \boldsymbol{v}_n^r$。于是，我们将共识误差和跟踪误差写成如下紧凑形式

$$\boldsymbol{\phi}^r = \begin{bmatrix} \phi_y^r \\ \phi_v^r \end{bmatrix} = \begin{bmatrix} \mathbb{E}\left[\|\boldsymbol{Y}^r - \mathbf{1}(\bar{\boldsymbol{y}}^r)^\top\|_F^2\right] \\ \mathbb{E}\left[\|\boldsymbol{V}^r - \mathbf{1}(\bar{\boldsymbol{v}}^r)^\top\|_F^2\right] \end{bmatrix} \in \mathbb{R}^2. \tag{8}$$

LSGT 方法的主要理论结果如下定理所示。

定理 1 假设假设 *1* 至 *5* 均成立。对于充分小的 $\gamma < 1$ （满足(21)、(26)、(31)、(37)以及引理 *1* 至引理 *3* 中的条件），我们有

$$\frac{1}{T}\sum_{r=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] \leq \underbrace{\frac{4(F(\bar{\boldsymbol{y}}^0)-\underline{F})}{\gamma ET}+\frac{40L\gamma\sigma^2}{N|\mathcal{I}|}}_{\text{terms same as centralized SGD}} + \underbrace{\frac{16(1+7\lambda_w^2)^2E^2L^2\gamma^2}{(1-\lambda_w^2)^4}\left(\frac{2577N\sigma^2}{|\mathcal{I}|}+\frac{111\phi_v^0}{T}\right)}_{\text{terms due to decentralized optimization}}. \tag{9}$$

证明： 证明的第一个关键步骤是建立如下共识误差矩阵和跟踪误差矩阵的动态方程

$$\boldsymbol{\phi}^{r+1} \leq \boldsymbol{A}\boldsymbol{\phi}^r + \boldsymbol{C}\boldsymbol{e}^r,$$

其中， $\boldsymbol{A}$ 和 $\boldsymbol{C}$ 为系数矩阵，另有扰动向量

$$\boldsymbol{e}^r = \begin{bmatrix} \mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] \\ \frac{\sigma^2}{|\mathcal{I}|} \end{bmatrix}. \tag{10}$$

第二个关键步骤是分析 $\boldsymbol{\phi}^r$ 如何影响式 (1) 中目标值的下降。综合这两步，即可得到式 (9) 中的界。详细内容见第 4 节。■

定理 1 表明，在采用适当的常数步长 γ时，所提出的 LSGT 算法收敛到问题 (1) 的驻点解邻域。具体而言，式 (9) 右端（RHS）的前两项与网络拓扑结构无关，且阶数与集中式 SGD 方法相同 [27]。与此同时，最后两项源于去中心化优化，并取决于网络连通性λ_w 和初始跟踪误差 ϕ_v^0。

更重要的是，以下推论给出了所提出的 LSGT 算法在本地 SGD 数量 E 和网络规模 N 下实现线性加速的条件。

推论 1 令 $\gamma = \sqrt{\frac{N}{ET}}$ 和 $E \leq (\frac{T}{N^5})^{\frac{1}{3}}$，其中 T 足够大，使得 γ 满足定理 *1* 中的条件。那么，对于所提出的 *LSGT* 算法，我们有

$$\frac{1}{T}\sum_{r=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] \leq \frac{4(F(\bar{\boldsymbol{y}}^0)-\underline{F})}{\sqrt{NET}}+\frac{40L\sigma^2}{\sqrt{NET}|\mathcal{I}|} + \frac{16(1+7\lambda_w^2)^2L^2}{(1-\lambda_w^2)^4\sqrt{NET}}\left(\frac{2577\sigma^2}{|\mathcal{I}|}+\frac{111\phi_v^0}{NT}\right). \tag{11}$$

推论 1 表明，当 T 较大时，LSGT 的驻点间隙以 $\mathcal{O}(1/\sqrt{NET})$ 的速度次线性衰减。这快于原始随机 GT 方法 [14,推论 1], 的收敛速度 $\mathcal{O}(1/\sqrt{NT})$，充分体现了采用本地 SGD 并结合 $E > 1$ 来降低通信开销的优势。

备注2 （网络连通性与随机梯度误差的影响）从 (9) 和(11)可以看出，较小的 λ_w 能够减小右端项界中的第 *3* 项和第 *4* 项。由于较小的 λ_w 意味着更高的网络连通性，这表明网络连通性越高，*LSGT* 算法就能越快收敛并实现变量一致性。另一方面，还可以看出，增大小批量大小 $|\mathcal{I}|$ 能够提升收敛性能。

4 定理 1 的证明

本节给出定理 1 的证明。对证明不感兴趣的读者可以跳过本节，直接阅读第 5 节，了解 LSGT 算法的扩展。

4.1 关键引理

定理 1 的证明依赖于以下四个关键引理。其证明放在附录 B 至 D 以及补充材料的第 I 节中。为便于阐述，附录 A 中提供了一些预备知识。

第一个引理给出了局部变量（$\boldsymbol{y}^{r,\,n^q}$，$\boldsymbol{v}^{r,\,n^q}$）与其对应平均值（$\bar{\boldsymbol{y}}^r$，$\bar{\boldsymbol{v}}^r$）之间距离的上界。

引理1 (局部更新差距**)** 假设假设 *1* 至 *5* 成立。对于足够小的$\gamma \leq \frac{\lambda_w^2}{32ENL}$，有

$$\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{y}_n^{r,q}-\bar{\boldsymbol{y}}^r\|^2] \leq 16(E-1)\lambda_w^2\phi_y^r + 8(E-1)\lambda_w^2[1+E(E-1)]\gamma^2\phi_v^r + 4E^2(E-1)N\gamma^2\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2] + 64E(E-1)^2N\gamma^2\frac{\sigma^2}{|\mathcal{I}|}, \tag{12a}$$

and

$$\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{v}_n^{r,q}-\bar{\boldsymbol{v}}^r\|^2] \leq 8(E-1)L^2(1+16\lambda_w^2)\phi_y^r + 4(E-1)\lambda_w^2\phi_v^r + 32E^2(E-1)NL^2\gamma^2\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2] + 32(E-1)N\frac{\sigma^2}{|\mathcal{I}|}. \tag{12b}$$

基于上述引理1，在接下来的两个引理中，我们分别确定一致性误差ϕ_y^r 和跟踪误差 ϕ_v^r的收缩性质。

引理2 (一致性误差界**)** 假设假设 *1* 至 *5*成立。对于满足以下条件的充分小的 γ

$$\gamma \leq \frac{(1-\lambda_w^2)^2}{32L^2(1+7\lambda_w^2)[8(1+16\lambda_w^2)(E-1)^2]}, \tag{13}$$

我们有

$$\phi_y^{r+1} = \mathbb{E}\left[\left\|\boldsymbol{Y}^{r+1}-\mathbf{1}(\bar{\boldsymbol{y}}^r)^\top\right\|^2\right]$$

$$\leq \frac{(1+3\lambda_w^2)}{4}\phi_y^r + \frac{2(1+7\lambda_w^2)}{1-\lambda_w^2}\left[1+8(E-1)^2\lambda_w^2\right]\gamma^2\phi_v^r$$
$$+\frac{256(1+7\lambda_w^2)E^2(E-1)^2NL^2\gamma^4}{1-\lambda_w^2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] + \frac{256(1+7\lambda_w^2)(E-1)^2N\gamma^2\sigma^2}{(1-\lambda_w^2)|\mathcal{I}|}. \tag{14}$$

引理 3（跟踪误差）假设假设 *1* 至 *5* 成立。对于充分小的$\gamma \leq \frac{(1-\lambda_w^2)}{3600ENL}$，我们有

$$\begin{aligned}\phi_v^{r+1} &= \mathbb{E}\left[\left\|\boldsymbol{V}^{r+1} - \mathbf{1}(\bar{\boldsymbol{v}}^r)^\top\right\|^2\right] \\ &\leq \frac{30(1+7\lambda_w^2)(1+\lambda_w^2)L^2}{1-\lambda_w^2}\phi_y^r + \frac{1+3\lambda_w^2}{4}\phi_v^r + \frac{12(1+7\lambda_w^2)N\sigma^2}{(1-\lambda_w^2)|\mathcal{I}|} \\ &\quad + \frac{60(1+7\lambda_w^2)}{1-\lambda_w^2}E^2NL^2\gamma^2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right].\end{aligned} \tag{15}$$

接下来，我们刻画全局目标函数的一个关键下降性质。

引理4 (下降引理**)** 假设 *1* 至 *5* 成立。根据引理*1*至*3*，我们有$\forall r \geq 0$,

$$\begin{aligned}F(\bar{\boldsymbol{y}}^{r+1}) \leq F(\bar{\boldsymbol{y}}^r) - \frac{\gamma E}{2}\mathbb{E}[\|\nabla F(\bar{\boldsymbol{y}}^r)\|^2] - \frac{\gamma}{2}\left[1-2E^2L\gamma-8E^2(E-1)L^2\gamma^2\right]\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] \\ + \frac{L^2\gamma}{N}[1+16(E-1)\lambda_w^2]\phi_y^r + 10EL\gamma^2\frac{\sigma^2}{N|\mathcal{I}|} + 8(E-1)\lambda_w^2[1+E(E-1)]\frac{L^2\gamma^3}{N}\phi_v^r,\end{aligned} \tag{16}$$

其中 $\nabla F(\bar{\boldsymbol{y}}^r) = \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\bar{\boldsymbol{y}}^r)]$。

4.2 定理 1 的证明

根据引理2和引理3，我们可以建立如下 $\boldsymbol{\phi}^r$ 动力系统

$$\boldsymbol{\phi}^{r+1} \leq \boldsymbol{A}\boldsymbol{\phi}^r + \boldsymbol{C}\boldsymbol{e}^r, \tag{17}$$

其中，不等式是逐元素的，矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{C}$ 定义如下

$$\boldsymbol{A} = \begin{bmatrix} \frac{(1+3\lambda_w^2)}{4} & \frac{2(1+7\lambda_w^2)\left[1+8(E-1)^2\lambda_w^2\right]\gamma^2}{1-\lambda_w^2} \\ \frac{30(1+7\lambda_w^2)(1+\lambda_w^2)L^2}{1-\lambda_w^2} & \frac{1+3\lambda_w^2}{4} \end{bmatrix}, \tag{18}$$

$$\boldsymbol{C} = \begin{bmatrix} \frac{256(1+7\lambda_w^2)E^2(E-1)^2NL^2\gamma^4}{1-\lambda_w^2} & \frac{256(1+7\lambda_w^2)(E-1)^2N\gamma^2}{1-\lambda_w^2} \\ \frac{60(1+7\lambda_w^2)E^2NL^2\gamma^2}{1-\lambda_w^2} & \frac{12(1+7\lambda_w^2)N}{1-\lambda_w^2} \end{bmatrix}. \tag{19}$$

可以验证，存在一个正向量 $\boldsymbol{s} = [s_1, s_2]^\top \in \mathbb{R}^2$ 满足

$$s_1 < \frac{(1-\lambda_w^2)^2}{40(1+7\lambda_w^2)(1+\lambda_w^2)L^2}s_2, \tag{20}$$

使得 $\boldsymbol{A}\boldsymbol{s} < \boldsymbol{s}$ 只要

$$\gamma \leq \frac{3(1-\lambda_w^2)s_1}{16(1+7\lambda_w^2)[1+8(E-1)\lambda_w]s_2}. \tag{21}$$

因此，根据 (21)， $\boldsymbol{A}$ 的谱半径满足 $\rho(\boldsymbol{A}) < 1$，由 [28, 推论 8.1.29]可得。然后，根据 [28, 推论 5.6.16], ，可得

$$\det(\mathbf{I} - \boldsymbol{A}) > 0, \quad \sum_{r=0}^{\infty} \boldsymbol{A}^r = (\mathbf{I} - \boldsymbol{A})^{-1}. \tag{22}$$

因此，(17) 可界定为

$$\boldsymbol{\phi}^r \leq \boldsymbol{A}\boldsymbol{\phi}^{r-1} + \boldsymbol{C}\boldsymbol{e}^{r-1} \leq \boldsymbol{A}^r\boldsymbol{\phi}_0 + \sum_{t=0}^{r-1} \boldsymbol{A}^t\boldsymbol{C}\boldsymbol{e}^{r-1-t}. \tag{23}$$

将 (23) 对 r 从 0 到 T求和，进一步可得

$$\begin{aligned}
\sum_{r=0}^{T} \boldsymbol{\phi}^r &\leq \sum_{r=0}^{T} \boldsymbol{A}^r\boldsymbol{\phi}^0 + \sum_{r=0}^{T}\sum_{t=1}^{r-1} \boldsymbol{A}^t\boldsymbol{C}\boldsymbol{e}^{r-1-t} \\
&\leq \left(\sum_{r=0}^{\infty} \boldsymbol{A}^r\right)\boldsymbol{\phi}^0 + \left(\sum_{r=0}^{\infty} \boldsymbol{A}^r\right)\sum_{r=0}^{T-1} \boldsymbol{C}\boldsymbol{e}^r \\
&= (\mathbf{I} - \boldsymbol{A})^{-1}\boldsymbol{\phi}^0 + (\mathbf{I} - \boldsymbol{A})^{-1}\sum_{r=0}^{T-1} \boldsymbol{C}\boldsymbol{e}^r,
\end{aligned} \tag{24}$$

其中，最后一个等号由 (22) 得出。

由于 $\boldsymbol{\phi}^r$ 始终为正，因此有 $\sum_{r=0}^{T-1} \boldsymbol{\phi}^r \leq \sum_{r=0}^{T} \boldsymbol{\phi}^r$。于是，由 (24) 可得

$$\sum_{r=0}^{T-1} \boldsymbol{\phi}^r \leq (\mathbf{I} - \boldsymbol{A})^{-1}\boldsymbol{\phi}^0 + (\mathbf{I} - \boldsymbol{A})^{-1}\sum_{r=0}^{T-1} \boldsymbol{C}\boldsymbol{e}^r. \tag{25}$$

补充材料的第 III 节证明了 $(\boldsymbol{I} - \boldsymbol{A})^{-1}$ 可界定如下。

引理 5 假设 γ 满足

$$\gamma \leq \tfrac{3(1-\lambda_w^2)^2}{320(1+7\lambda_w^2)(1+\lambda_w^2)[1+8(E-1)\lambda_w]NL}. \tag{26}$$

那么，$(\boldsymbol{I} - \boldsymbol{A})^{-1}$ 具有如下逐元素上界

$$\begin{aligned}
&(\mathbf{I} - \boldsymbol{A})^{-1} \\
&\leq \frac{8(1+7\lambda_w^2)}{(1-\lambda_w^2)^3}\begin{bmatrix} 1 & 2[1+8(E-1)^2\lambda_w^2]\gamma^2 \\ 30(1+\lambda_w^2)L^2 & 1 \end{bmatrix}.
\end{aligned} \tag{27}$$

将 (27) 代入 (25)，并利用 $\phi_y^0 = 0$，可得

$$\sum_{r=0}^{T-1} \phi_y^r \leq \frac{64(1+7\lambda_w^2)^2}{(1-\lambda_w^2)^4}\left[32(E-1)^2+15\lambda_w^2[1+8(E-1)^2]\right]E^2NL^2\gamma^4\sum_{r=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]$$

$$+\frac{64(1+7\lambda_w^2)^2}{(1-\lambda_w^2)^4}\left[32(E-1)^2+3\lambda_w^2[1+8(E-1)^2]\right]\frac{N\gamma^2T\sigma^2}{|\mathcal{I}|}$$
$$+\frac{16(1+7\lambda_w^2)^2}{(1-\lambda_w^2)^4}\lambda_w^2[1+8(E-1)^2]\gamma^2\phi_v^0, \tag{28}$$

and

$$\sum_{r=0}^{T-1}\phi_v^r \le \frac{8(1+7\lambda_w^2)^2}{(1-\lambda_w^2)^4}\phi_v^0+\frac{32(1+7\lambda_w^2)^2}{(1-\lambda_w^2)^4}[1920(E-1)^2(1+\lambda_w^2)L^2\gamma^2+3]\frac{NT\sigma^2}{|\mathcal{I}|}$$
$$+\frac{480(1+7\lambda_w^2)^2E^2NL^2\gamma^2}{(1-\lambda_w^2)^4}[128(E-1)^2(1+\lambda^2)L^2\gamma^2+1]\sum_{r=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] \tag{29}$$
$$\le \frac{8(1+7\lambda_w^2)^2}{(1-\lambda_w^2)^4}\phi_v^0+\frac{128(1+7\lambda_w^2)^2NT\sigma^2}{(1-\lambda_w^2)^4|\mathcal{I}|}+\frac{512(1+7\lambda_w^2)^2E^2NL^2\gamma^2}{(1-\lambda_w^2)^4}\sum_{r=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right], \tag{30}$$

其中，为了得到 (30)，我们假设

$$\gamma \le \frac{1}{62E(1+\lambda_w^2)NL}, \tag{31}$$

并将其应用于（29）右端的第2项和第3项。

另一方面，将引理4中的（4）从 $r=0$ 到 $r=T-1$进行望远镜求和，可得

$$\frac{\gamma E}{2}\sum_{r=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{\boldsymbol{y}}^r)\|^2]$$
$$\le F(\bar{\boldsymbol{y}}^0)-F(\bar{\boldsymbol{y}}^T)+8(E-1)\lambda_w^2[1+E(E-1)]\frac{L^2\gamma^3}{N}\sum_{r=0}^{T-1}\phi_v^r$$
$$-\frac{\gamma}{2}[1-2E^2L\gamma-8E^2(E-1)L^2\gamma^2]\sum_{r=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]$$
$$+\frac{L^2\gamma[1+16(E-1)\lambda_w^2]}{N}\sum_{r=0}^{T-1}\phi_y^r+\frac{10EL\gamma^2T\sigma^2}{N|\mathcal{I}|}. \tag{32}$$

此外，根据假设4，有

$$\sum_{r=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]$$
$$\le 2\sum_{r=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}[\nabla f_n(\boldsymbol{y}_n^r)-\nabla f_n(\bar{\boldsymbol{y}}^r)]\right\|^2+2\|\nabla F(\bar{\boldsymbol{y}}^r)\|^2\right]$$
$$\le \frac{2L^2}{N}\sum_{r=0}^{T-1}\phi_y^r+2\sum_{r=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{\boldsymbol{y}}^r)\|^2]. \tag{33}$$

在等式（33）两边乘以 $\frac{\gamma E}{4}$，再将其代入（32），并根据假设3，可得

$$
\begin{aligned}
&\frac{\gamma E}{4}\sum_{r=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] \\
&\leq F(\bar{\boldsymbol{y}}^0)-\underline{F}+\frac{10EL\gamma^2T\sigma^2}{N|\mathcal{I}|}-\frac{\gamma}{2}[1-2E^2L\gamma-8E^2(E-1)L^2\gamma^2]\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] \\
&+\left\{2[1+16(E-1)\lambda_w^2]+E\right\}\frac{L^2\gamma}{2N}\sum_{r=0}^{T-1}\phi_y^r+8(Q-1)\lambda_w^2[1+E(E-1)]\frac{L^2\gamma^3}{N}\sum_{r=0}^{T-1}\phi_v^r. \qquad (34)
\end{aligned}
$$

然后，将 (28) 和 (30) 代入上述 (34)，可得

$$
\begin{aligned}
&\frac{\gamma E}{4}\sum_{r=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] \\
&\leq (F(\bar{\boldsymbol{y}}^0)-\underline{F})+\frac{10EL\gamma^2T\sigma^2}{N|\mathcal{I}|} \\
&+\left\{\frac{4(1+7\lambda_w^2)^2}{(1-\lambda_w^2)^4}\left[32(E-1)^2+3\lambda_w^2[1+8(E-1)^2]\right]\left[2[1+16(E-1)\lambda_w^2]+E\right]NL\gamma\right. \\
&\quad\left.+\frac{128(1+7\lambda_w^2)^2}{(1-\lambda_w^2)^4}(E-1)\lambda_w^2[1+E(E-1)]NL\gamma\right\}\frac{8L\gamma^2T\sigma^2}{N|\mathcal{I}|} \\
&+\frac{8(1+7\lambda_w^2)^2}{(1-\lambda_w^2)^4}\left\{8(E-1)\lambda_w^2[1+E(E-1)]+\lambda_w^2[1+8(E-1)^2]\{2[1+16(E-1)\lambda_w^2]+E\}\right\}\frac{L^2\gamma^3}{N}\phi_v^0 \\
&-\frac{\gamma}{2}\left\{1-2E^2L\gamma-8E^2(E-1)N^2L^2\gamma^2-\frac{32(1+7\lambda_w^2)^2}{(1-\lambda_w^2)^4}\{2[1+16(E-1)\lambda_w^2]+E\}\right. \\
&\quad\{32(E-1)^2+15\lambda_w^2[1+8(E-1)^2]\}E^2L^4\gamma^4 \\
&\quad\left.-\frac{4096(1+7\lambda_w^2)^2}{(1-\lambda_w^2)^4}(E-1)\lambda_w^2[1+E(E-1)]E^2L^4\gamma^4\right\}\sum_{r=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] \qquad (35) \\
&\leq (F(\bar{\boldsymbol{y}}^0)-\underline{F})+\frac{10EL\gamma^2T\sigma^2}{N|\mathcal{I}|}+\frac{444(1+7\lambda_w^2)^2E^3L^2\gamma^3\phi_v^0}{N(1-\lambda_w^2)^4}+\frac{10308(1+7\lambda_w^2)^2E^3L^2\gamma^3T\sigma^2}{(1-\lambda_w^2)^4|\mathcal{I}|}, \qquad (36)
\end{aligned}
$$

为得到 (36)，我们假设

$$
\gamma\leq\frac{1-\lambda_w^2}{320E^2NL}, \qquad (37)
$$

使得 (35) 右端最后一项的系数为负，因此可以去掉该项。此外，我们利用 $\lambda_w^2<1$ 和 $E\geq 1$ 的性质，得到了 (36) 右端第 3 项和第 4 项系数的界。最后，将 (36) 两边除以 $\frac{T\gamma E}{4}$，即可得到定理 1 的结论。

5 混合数据上的学习扩展

本节将讨论 LSGT 算法的一种扩展，以求解以下优化问题

$$\min_{(\boldsymbol{x},\boldsymbol{\theta})\in\mathbb{R}^{J+K}} \quad \frac{1}{NS}\sum_{i=1}^{S} f\left(\boldsymbol{B}_i\boldsymbol{x},\boldsymbol{\theta}\right) \tag{38}$$

其中

$$\boldsymbol{B}_i=\sum_{n=1}^{N}\boldsymbol{B}_{n,i}\in\mathbb{R}^{M\times J},\forall i\in[S]. \tag{39}$$

问题（38）的一个实例出现在混合数据上的分布式学习问题中：智能体只能访问部分数据样本，并且只知道部分数据特征，但它们会协同训练一个深度神经网络（DNN）模型。关于混合数据学习的详细描述见 [25]。具体而言，在问题（38）中， S 表示全局数据集 $\mathcal{D}=\cup_{n=1}^{N}\mathcal{D}_n$ 中的数据样本数， $\boldsymbol{x}$ 和 $\boldsymbol{\theta}$ 分别是 DNN 模型第一层和其余层的系数参数，而 $\boldsymbol{B}_i$ 是与第 i 个数据样本相关的矩阵。根据（39），$\boldsymbol{B}_{n,\,i}$ 是智能体 i 独自掌握的第 i 个数据样本的特征信息矩阵。如果 $i\notin D_n$，即智能体 i 不拥有样本 i，那么 $\boldsymbol{B}_{n,\,i}=\boldsymbol{0}$。

因此，与问题（1）相比，问题（38）的难点在于，每个智能体 n不知道 $\boldsymbol{B}_i$ ，但只能访问 $\boldsymbol{B}_{n,\,i}$ 。因此，算法1中的 LSGT 算法无法直接用于处理问题（38）。为说明这一点，请注意，问题（38）的局部更新需要计算

$$\boldsymbol{\theta}_n^{r,q}=\boldsymbol{\theta}_n^{r,q-1}-\alpha\frac{1}{NS}\sum_{i=1}^{S}\nabla_{\boldsymbol{\theta}}f\big(\underbrace{\boldsymbol{B}_i\boldsymbol{x}_n^{r,q-1}}_{(a)},\boldsymbol{\theta}_n^{r,q-1}\big), \tag{40a}$$

$$\boldsymbol{x}_n^{r,q}=\boldsymbol{x}_n^{r,q-1}-\beta\underbrace{\frac{1}{NS}\sum_{i=1}^{S}\boldsymbol{B}_i^{\top}\nabla_{\boldsymbol{z}}f\big(\boldsymbol{B}_i\boldsymbol{x}_n^{r,q-1},\boldsymbol{\theta}_n^{r,q-1}\big)}_{(b)}, \tag{40b}$$

其中， $\alpha,\beta>0$ 是步长参数，而 $\nabla_{\boldsymbol{z}}f(\boldsymbol{z},\theta)$ 是关于 $\boldsymbol{z}$ 的梯度。不难看出，式（40）中的更新遗憾的是无法实现的，因为每个智能体都缺少关于 $\boldsymbol{B}_i$ 的全梯度信息，无法在本地计算项 (a) 和 (b)。

鉴于此，我们引入一个新的辅助变量 $\boldsymbol{z}^{r,\,n\,q}_{\,i\,,}$ 用于估计求和项 (a)，即 $\boldsymbol{B}_i\boldsymbol{x}^{r,\,n\,q}_{\,,}$，并引入另一个辅助变量 $\boldsymbol{u}^{r,\,n\,q}_{\,,}$用于估计项 (b)。给定 $\{\boldsymbol{z}_{n,i}^{r,q}\}_{i\in D}$，智能体 n 可以在本地计算小批量梯度，计算方式为

$$\boldsymbol{g}_{\theta,n}^{r,q}\triangleq\frac{1}{N|\mathcal{I}|}\sum_{\xi\in\mathcal{I}_n^{r,q}}\nabla_{\boldsymbol{\theta}}f(\boldsymbol{z}_{n,\xi}^{r,q},\boldsymbol{\theta}_n^{r,q}), \tag{41a}$$

$$\boldsymbol{g}_{x,n}^{r,q}\triangleq\frac{1}{|\mathcal{I}|}\sum_{\xi\in\mathcal{I}_n^{r,q}}\boldsymbol{B}_{n,\xi}^{\top}\nabla_{\boldsymbol{z}}f(\boldsymbol{z}_{n,\xi}^{r,q},\boldsymbol{\theta}_n^{r,q}), \tag{41b}$$

算法2 求解 (38) 的提出的 MUST 方法

1: **初始化** 令 $\boldsymbol{x}_1^0 = \ldots = \boldsymbol{x}_N^0$, $\boldsymbol{\theta}_1^0 = \ldots = \boldsymbol{\theta}_N^0$, $\boldsymbol{z}_{n,i}^0 = N\boldsymbol{B}_{n,i}\boldsymbol{x}_n^0$ 和 $\boldsymbol{u}_n^0 = \boldsymbol{g}_{x,n}^0, \forall n \in [N]$。 2: **对于** 通信轮次 $r = 0$ **至** T **执行** 3: **对于** 智能体 $n = 1$ **至** N，并行**执行** 4: 从邻居接收信息并设置

$$\begin{bmatrix} \boldsymbol{\theta}_n^{r,0} \\ \boldsymbol{x}_n^{r,0} \\ \boldsymbol{z}_n^{r,0} \\ \boldsymbol{u}_n^{r,0} \end{bmatrix} = \sum_{m=1}^{N} W_{n,m} \begin{bmatrix} \boldsymbol{\theta}_m^{r,0} \\ \boldsymbol{x}_m^{r,0} \\ \boldsymbol{z}_m^{r,0} \\ \boldsymbol{u}_m^{r,0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{g}_{\theta,n}^{r,0} \\ \boldsymbol{g}_{x,n}^{r,0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{g}_{\theta,n}^{r} \\ \boldsymbol{g}_{x,n}^{r} \end{bmatrix} \tag{44}$$

5: **针对** 本地更新 $q = 1, \ldots, E$ **do**
6:

$$\boldsymbol{\theta}_n^{r,q} = \boldsymbol{\theta}_n^{r,q-1} - \alpha \boldsymbol{g}_{\theta,n}^{r,q-1}, \tag{45a}$$

$$\boldsymbol{x}_n^{r,q} = \boldsymbol{x}_n^{r,q-1} - \beta \boldsymbol{u}_n^{r,q-1}, \tag{45b}$$

$$\boldsymbol{z}_n^{r,q} = \boldsymbol{z}_n^{r,q-1} + N\boldsymbol{B}_n(\boldsymbol{x}_n^{r,q} - \boldsymbol{x}_n^{r,q-1}), \tag{45c}$$

$$\boldsymbol{u}_n^{r,q} = \boldsymbol{u}_n^{r,q-1} + (\boldsymbol{g}_{x,n}^{r,q} - \boldsymbol{g}_{x,n}^{r,q-1}), \tag{45d}$$

其中，随机梯度 $\boldsymbol{g}^{r,q}_{\theta,n}$ 以及 $\boldsymbol{g}^{r,q}_{x,n}$ 是基于式 (41)，使用 $\boldsymbol{\theta}^{r,q}_n$ 和 $\boldsymbol{x}^{r,q}_n$ 以及一个小批量 $\mathcal{I}^{r,q}_n$ 和 $\boldsymbol{B}_n = [\boldsymbol{B}_{n,1}^\top, \ldots, \boldsymbol{B}_{n,S}^\top]^\top$ 计算得到的。

7: 结束循环 8: 设置 $\boldsymbol{\theta}_n^{r+1} = \boldsymbol{\theta}^{r,E}_n$, $\boldsymbol{x}_n^{r+1} = \boldsymbol{x}^{r,E}_n$, $\boldsymbol{z}_n^{r+1} = \boldsymbol{z}^{r,E}_n$, $\boldsymbol{u}_n^{r+1} = \boldsymbol{u}^{r,E}_n$, $\boldsymbol{g}_\theta^{r+1}$, $_n = \boldsymbol{g}^{r,E}_{\theta,n}$, $\boldsymbol{g}_x^{r+1}$, $_n = \boldsymbol{g}^{r,E}_{x,n}$, **, 并将** $(\boldsymbol{\theta}_n^{r+1}, \boldsymbol{x}_n^{r+1}, \boldsymbol{z}_n^{r+1}, \boldsymbol{u}_n^{r+1})$ **发送给邻居。9: 结束循环 10: 结束循环**

同时，通过共识平均，智能体分别通过交换跟踪变量 $\boldsymbol{z}^{r,q}_n = [(\boldsymbol{z}^{r,q}_{n,1})^\top, \ldots, (\boldsymbol{z}^{r,q}_{n,S})^\top]^\top$ 和 $\boldsymbol{g}^{r,q}_{x,n}$ 与邻居交换跟踪项 (a) 和项 (b)。因此，(40) 可替换为

$$\boldsymbol{\theta}_n^{r,q} = \boldsymbol{\theta}_n^{r,q-1} - \alpha \boldsymbol{g}_{\theta,n}^{r,q-1}, \tag{42}$$

$$\boldsymbol{x}_n^{r,q} = \boldsymbol{x}_n^{r,q-1} - \beta \boldsymbol{u}_n^{r,q-1}. \tag{43}$$

与 LSGT 算法类似，智能体仍可在每个通信轮次执行 E 本地随机梯度下降步，以降低通信成本。我们在算法2中总结了用于求解问题 (38) 的提出的方法，并将其称为多次本地更新变量求和跟踪（*MUST*）算法。

5.1 收敛速度分析

在本小节中，我们为 MUST 算法建立收敛条件。与假设5类似，我们对 (41) 中的小批量梯度作出如下假设。

假设 6 对于每个智能体 $n \in [N]$，均有 ●无偏梯度：

$\mathbb{E}[\boldsymbol{g}_{\theta,n}] = \frac{1}{NS}\sum_{i=1}^{S} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{z}_{n,i}, \boldsymbol{\theta}_n)$;

$\mathbb{E}[\boldsymbol{g}_{x,n}] = \frac{1}{S}\sum_{i=1}^{S} \boldsymbol{B}_{n,i}^\top \nabla_{\boldsymbol{z}} f(\boldsymbol{z}_{n,i}, \boldsymbol{\theta}_n)$;

- *Uniform bounded variance:*
  $\mathbb{E}[\|\boldsymbol{g}_{\theta,n} - \frac{1}{NS}\sum_{i=1}^{S}\nabla_{\boldsymbol{\theta}} f(\boldsymbol{z}_{n,i},\boldsymbol{\theta}_n)\|^2] \le \frac{\sigma^2}{|\mathcal{I}|}$,
  $\mathbb{E}[\|\boldsymbol{g}_{x,n} - \frac{1}{S}\sum_{i=1}^{S}\boldsymbol{B}_{n,i}^{\top}\nabla_{\boldsymbol{z}} f(\boldsymbol{z}_{n,i},\boldsymbol{\theta}_n)\|^2] \le \frac{\sigma^2}{|\mathcal{I}|}$.

我们还假设（38）中的成本函数 f 有下界且光滑，与假设3和4类似。MUST算法的主要理论结果由以下定理给出。

定理 2 设假设 *1* 至假设 *4*以及假设 *6*均成立。定义平均值$\bar{\boldsymbol{x}}^r = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n^r$，并以相同方式记 $\bar{\boldsymbol{\theta}}^r$、 $\bar{\boldsymbol{z}}^r$、 $\bar{\boldsymbol{u}}^r$ 。取 $\alpha = \beta = \sqrt{\frac{N}{ET}}$，对于$E \le \frac{T^{\frac{1}{3}}}{N^3}$、 $T \ge N^{\frac{5}{2}}E^3$，有

$$
\begin{aligned}
&\frac{1}{T}\sum_{r=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbb{E}\left[\left\|\frac{1}{NS}\sum_{i=1}^{S}\nabla_{\boldsymbol{\theta}} f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\right)\right\|^2\right]\right.\\
&\quad\left.+\mathbb{E}\left[\left\|\frac{1}{NS}\sum_{i=1}^{S}\sum_{t=1}^{N}\boldsymbol{B}_{t,i}^{\top}\nabla_{\boldsymbol{B}_i\boldsymbol{x}} f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\right)\right\|^2\right]\right\}\\
&\le \frac{1000[\tilde{F}(\bar{\boldsymbol{\theta}}^0,\bar{\boldsymbol{x}}^0)-\underline{F}]}{\sqrt{NET}}+\frac{14L\sigma^2}{\sqrt{NET}|\mathcal{I}|}+\frac{222L^2(1+7\lambda_w^2)^2B_{\max}^2\sigma^2}{\sqrt{NET}N(1-\lambda_w^2)^6|\mathcal{I}|}\\
&\quad+\frac{2L^2B_{\max}^2}{\sqrt{NET}N(1-\lambda_w^2)^4}\left[\frac{990}{SEN}\phi_z^0+\frac{59(1+7\lambda_w^2)}{(1-\lambda_w^2)T}\phi_u^0\right],
\end{aligned}
\tag{46}
$$

其中 $\tilde{F}(\bar{\boldsymbol{x}}^0、\ \bar{\boldsymbol{\theta}}^0) \triangleq \frac{1}{NS}\sum_{i=1}^{S} f(\boldsymbol{B}_i\bar{\boldsymbol{x}}^0、\ \bar{\boldsymbol{\theta}}^0)$，$\phi_z^0 \triangleq \mathbb{E}\left[\|\boldsymbol{Z}^r - \mathbf{1}(\bar{\boldsymbol{z}}^r)^{\top}\|_F^2\right]$，其中 $\boldsymbol{Z}^r = [\boldsymbol{z}_1^r,\ldots,\boldsymbol{z}_N^r]^{\top}$，且 ϕ_u^0的定义方式相同。

定理 2 的证明方式与定理 1 类似。证明的预备内容和详细过程见附录H至附录I。根据定理 2 可知，MUST算法在 E 和N方面也具有相同的线性加速。

6 实验结果

本节展示所提出的 LSGT 和 MUST 算法的数值性能。

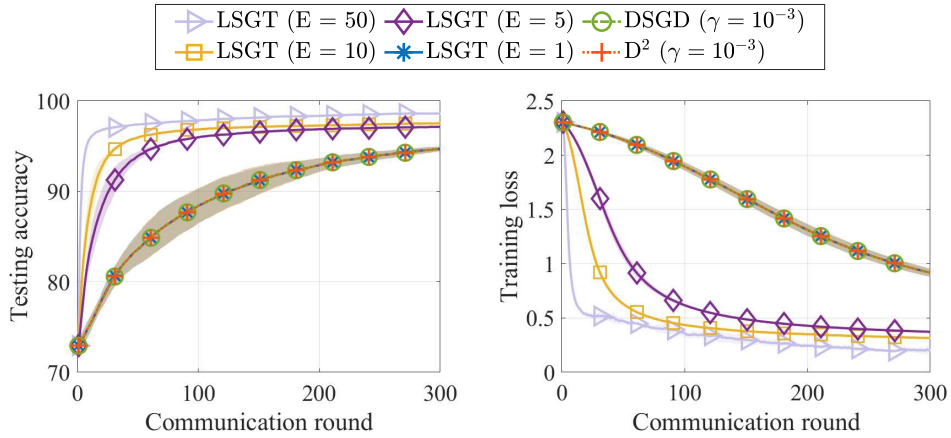
6.1 LSGT评估

6.1.1 实验设置

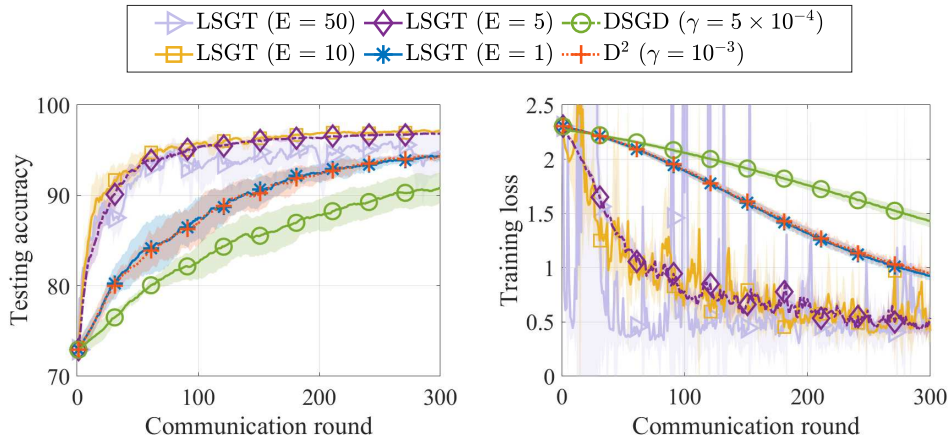
在实验中，我们设置一个包含 20 个智能体的连通网络，这些智能体通过由最大度规则 [12, 26]得到的混合矩阵 $\boldsymbol{W}$ 交换信息。假设这些智能体由一个随机网络图连接，该图按 [29]中的方法生成。

我们考虑基于 MNIST 数据集 [30]的 10 类手写数字图像分类任务。该数据集包含 60000 张训练图像和 10000 张测试图像。每张图像的维度为 28×28，并被向量化为一个 1×784 向量。这些数据样本以独立同分布或非独立同分布的方式划分到各智能体的本地数据集中。

• **IID 设置** [31]: 首先将 60000 个训练样本打乱，然后平均分配给各个智能体。因此，每个智能体持有 3000 个样本，且类别分布相同。



(a) 独立同分布设置。



(b) 非 IID 设置。

图 1： 所提出的 LSGT 算法在不同 E($\gamma = 10^{-3}$、随机图)下的收敛曲线。

• **非 IID 设置** [22]：我们考虑一种病态的非 IID 情形，其中每个智能体接收 3000 个样本，这些样本最多涉及 2 个数字。按照数字标签从 9 到 0 对训练样本排序；然后将它们划分为 40 个子集，使每个子集包含 1500 个样本；接下来，随机将每 2 个子集分配给一个智能体。

为了对 MNIST 数据集中的图像进行分类，我们考虑一个具有 30 个神经元隐藏层 [32] 的两层深度神经网络（DNN）。隐藏层和输出层的激活函数分别为整流线性单元（ReLU）和 softmax 函数。训练损失函数为交叉熵 [33]。实验结果为 5 次独立试验的平均值。

6.1.2 结果与讨论

下面分别研究局部迭代次数 E、步长 γ 和网络拓扑 λ_w 的影响。

• **影响 E：** 我们设置本地更新次数为 $E = 1$、5、10、50，以研究本地迭代对所提出的 LSGT 算法性能的影响。如图 1(a) 所示，

表 2： 达到特定测试准确率所需的通信轮次。

acc.	独立同分布设置				非 IID 设置			
	$E=1$	$E=5$	$E=10$	$E=50$	$E=1$	$E=5$	$E=10$	$E=50$
85%	62	14	8	**3**	75	16	**10**	11
90%	126	26	14	**4**	136	30	**21**	41
95%	335	68	35	**8**	378	91	**65**	191

“acc.” 表示测试准确率。

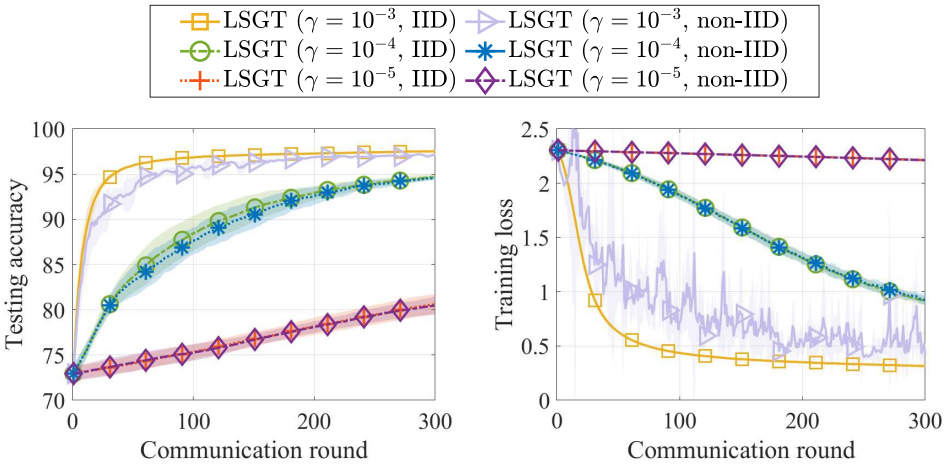


图 2： 所提出的 LSGT 算法在不同步长 ($E = 10$，随机图) 下的收敛曲线。

在 IID 情形下，随着 E 增加，训练损失和测试准确率收敛得更快。这是因为在 IID 设置中，去中心化智能体之间的数据服从相同分布，因此更多的局部更新有助于学习更好的公共模型。另一方面，从图 1(b) 可以看出，LSGT 的性能先从 $E = 1$ 提升到 $E = 10$，然后在 $E = 50$ 时变差。其原因在于，较大的 E 会扩大非独立同分布数据智能体之间的模型差异。通信轮次与局部更新 E 之间更具体的关系将在表 2 中进一步研究。该表展示了 LSGT 算法在不同 E 下达到 85%、90%、95% 测试准确率所需的通信轮次。在 IID 设置中，更多的局部更新会减少所需的通信轮次；而在非独立同分布情形下，随着 E 增加，所需的通信轮次先减少后增加。这与定理 1 的结论一致，即 $E \leq (\frac{T}{N^5})^{\frac{1}{3}}$ 不应过大。此外，在图 1 中，采用 $E = 5$ 的 LSGT 算法在 IID 和非 IID 设置下均优于 DSGD 和 D^2 方法。

- **影响 γ：** 从图 2 和图 3 可以看出，对于独立同分布和非独立同分布设置，采用较小步长（如 $\gamma = 10^{-5}$）的所提出 LSGT 算法在训练损失和测试准确率方面的收敛速度都较慢，这与定理 1 的分析一致。需要指出的是，在图 3 中，步长为 $E = 100$，$\gamma = 10^{-3}$ 的 LSGT 算法无法很好地收敛。根据推论 1 中的步长设置 $\gamma = \sqrt{N/ET}$，当 $E = 100$ 较大时，我们可以考虑减小步长。如图 3 所示，当 $E = 100$，$\gamma = 10^{-4}$ 时，LSGT 算法会收敛到更高的测试准确率。

- **Im对 λ_w的影响：** 为研究网络连通性的影响，我们考虑 r

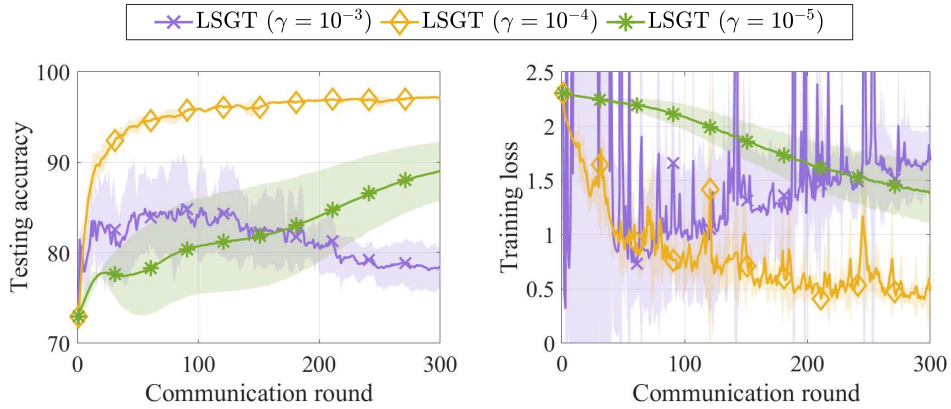


图3： 非独立同分布设置下，所提出的 LSGT 算法在不同步长下的收敛曲线（$E = 100$，随机图）。

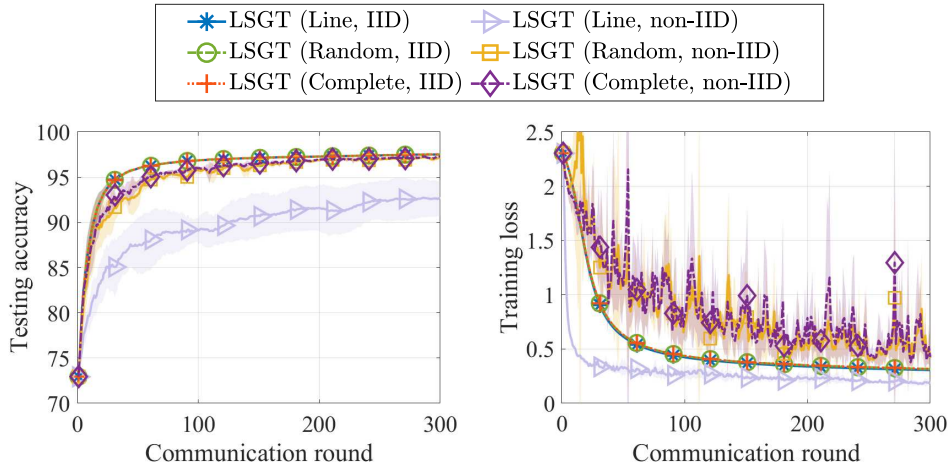


图4： 所提出的 LSGT 算法在不同拓扑结构下的收敛曲线（$E = 10$，$\gamma = 10^{-3}$）。

3 种阶数为 λ_w（线形图）、 $> \lambda_w$（随机图）、 $> \lambda_w$（完全图）的图，其连通性阶数依次相反。在图 4 中，对于 IID 设置，网络拓扑对收敛性能的影响较小。在非独立同分布情形下，当网络的连通性更强（完全图）时，提出的方法可获得更高的测试准确率。这与备注2中的分析一致。有趣的是，连通性更强的 LSGT 算法会收敛到更大的训练损失。原因在于，随着网络拓扑连通性的增强，训练模型会更接近全局模型。当训练模型应用于本地数据时，由于本地数据异质性，它可能无法很好地发挥作用 [34]。

6.2 MUST算法的评估

为考察所提出的 MUST 算法的有效性，我们考虑与第6.1.1节相同的随机 20智能体网络和深度神经网络。在混合数据设置中，首先打乱训练样本，然后将其划分为 3000 个大小为 20 的子集。接着，按照图5所示分配每个子集。具体而言，在每个包含 20 个样本的子集中，每个图像被分为 2 个图像块。每个子集中的第一个图像块按照正向

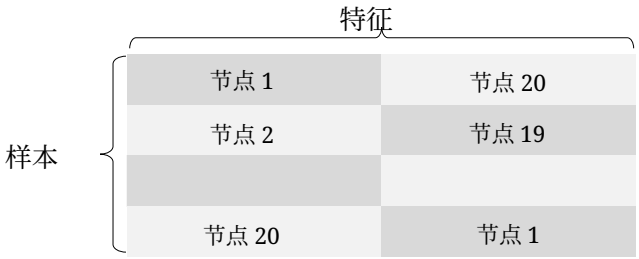


图 5： 混合数据情况下的数据划分。

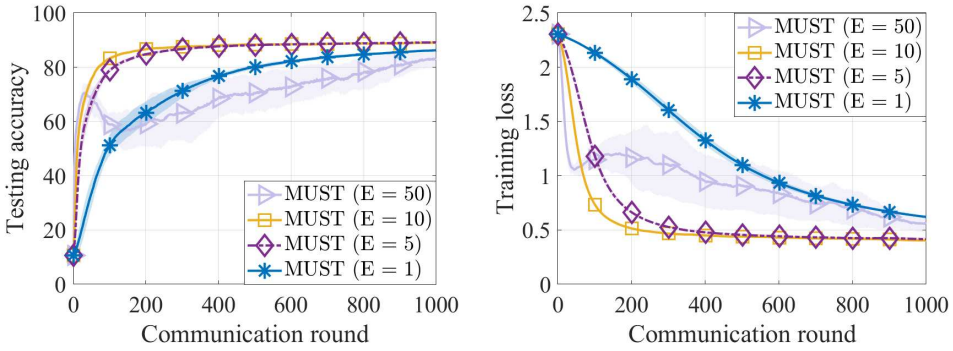


图 6： 提出的 MUST 算法在不同 E（$\gamma = 10^{-3}$，随机图）下的收敛曲线。

序列分配给 20 个智能体，第二个图像块则按相反顺序分配。因此，每个智能体都拥有一部分特征不完整的样本。

我们评估了 MUST 算法在局部迭代次数方面的性能。该次数设为 $E = 1$，5，10，50。如图 6 所示，提出的 MUST 方法在混合数据上表现稳健。特别是，随着 E 从 1 增加到 10，MUST算法获得的测试准确率和训练损失随迭代次数变化得更快。在较大的 $E = 50$ 下，正如预期，MUST 算法无法学习出有效的深度神经网络来对数字进行分类，甚至在准确率 $= 80$ % 附近波动。

7 结论

本文中，为了提高现有 GT 方法求解问题 (1) 时的通信效率，我们提出了一种新的 LSGT 算法（算法1），将局部 SGD 技术融入随机 GT 方法。理论上，我们建立了 LSGT 算法的收敛条件（定理 1），并证明其能够随着本地 SGD 更新次数的增加而获得线性加速 E（推论 1）。这与现有 GT 方法形成鲜明对比：这些方法要么未考虑多步本地 SGD，要么未能充分刻画本地 SGD 对 GT 方法的优势。作为扩展，我们还将梯度与变量跟踪思想拓展到 MUST 算法（算法2），以解决混合数据学习问题 (38)。实验结果表明，提出的 LSGT 方法比单步随机GT方法具有更好的学习性能，而 MUST 算法在异构混合数据上表现稳健。值得指出的是，据我们所知，提出的 LSGT 和 MUST 方法是首个用于去中心化学习、支持多次本地 SGD 更新的随机梯度跟踪算法，

就我们所知，首次提出了用于去中心化学习且支持多次本地 SGD 更新的随机梯度跟踪算法。未来，我们计划将这一框架拓展至时变通信网络拓扑 [35] 以及带压缩的情形 [36]。

算法1的另一种表达式

为便于分析，我们可以将第 r 轮第 q 次迭代时 $n \in [N]$ 的式（7）写为

$$\boldsymbol{y}_n^{r,q} = \sum_{m=1}^{N} W_{n,m}\boldsymbol{y}_m^r - \gamma \sum_{m=1}^{N} W_{n,m}\boldsymbol{v}_m^r - \gamma \sum_{k=1}^{q-1} \boldsymbol{v}_n^{r,k}, \tag{47a}$$

$$\boldsymbol{v}_n^{r,q} = \sum_{m=1}^{N} W_{n,m}\boldsymbol{v}_m^r + \boldsymbol{g}_n^{r,q} - \boldsymbol{g}_n^r, \tag{47b}$$

其中，$\boldsymbol{g}_n^r \triangleq \frac{1}{|\mathcal{I}|}\sum_{\xi\in\mathcal{I}_n^r} g_n(\boldsymbol{y}_n^r,\ \xi)$ 在式（4）中定义，$q \in [E]$，且 $r \in [T]$。将 $q = E$ 代入式（47），并根据算法1中的 $\boldsymbol{y}_n^{r+1} = \boldsymbol{y}^{r,\ n^E}$ 和 $\boldsymbol{v}_n^{r+1} = \boldsymbol{v}^{r,\ n^E}$，可得

$$\boldsymbol{y}_n^{r+1} = \sum_{m=1}^{N} W_{n,m}\boldsymbol{y}_m^r - \gamma \sum_{m=1}^{N} W_{n,m}\boldsymbol{v}_m^r - \gamma \sum_{q=1}^{E-1} \boldsymbol{v}_n^{r,q}, \tag{48a}$$

$$\boldsymbol{v}_n^{r+1} = \sum_{m=1}^{N} W_{n,m}\boldsymbol{v}_m^r + \boldsymbol{g}_n^{r+1} - \boldsymbol{g}_n^r, \tag{48b}$$

对任意 $n \in [N]$，均有。

对上述两个等式取平均，得到

$$\bar{\boldsymbol{y}}^{r+1} = \bar{\boldsymbol{y}}^r - \gamma\bar{\boldsymbol{v}}^r - \gamma \cdot \frac{1}{N}\sum_{n=1}^{N}\sum_{q=1}^{E-1} \boldsymbol{v}_n^{r,q}, \tag{49a}$$

$$\bar{\boldsymbol{v}}^{r+1} = \bar{\boldsymbol{v}}^r + \bar{\boldsymbol{g}}_n^{r+1} - \bar{\boldsymbol{g}}_n^r, \tag{49b}$$

其中 $\bar{\boldsymbol{g}}^r \triangleq \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{g}_n^r$。

与 [37, 公式 (27)]，类似，可以用归纳法证明：

引理 6 对于所有 $n \in [N]$，初始化 $\boldsymbol{v}_n^0 = \boldsymbol{g}_n^0$ 后，根据 (48b)，我们有

$$\bar{\boldsymbol{v}}^r \triangleq \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{v}^r, \quad \bar{\boldsymbol{v}}^r = \bar{\boldsymbol{g}}^r, \forall r \geq 0. \tag{50}$$

接下来，根据假设5，我们有

$$\mathbb{E}[\bar{\boldsymbol{v}}^r] = \mathbb{E}[\bar{\boldsymbol{g}}^r] = \frac{1}{N}\sum_{n=1}^{N} \nabla f_n(\boldsymbol{y}_n^r), \tag{51}$$

$$\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2] = \mathbb{E}[\|\bar{\boldsymbol{g}}^r\|^2]$$

$$= \mathbb{E}\left[\left\|\bar{\boldsymbol{g}}^r - \frac{1}{N}\sum_{n=1}^{N} \nabla f_n(\boldsymbol{y}_n^r) + \frac{1}{N}\sum_{n=1}^{N} \nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]$$

$$\leq 2\mathbb{E}\left[\left\|\bar{\boldsymbol{g}}^r-\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]+2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]$$
$$=\frac{2\sigma^2}{N|\mathcal{I}|}+2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]. \tag{52}$$

B 引理1的证明

下面，我们分别证明（12a）和（12b）。

• **证明**(12a)**：**我们有

$$\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{y}_n^{r,q}-\bar{\boldsymbol{y}}^r\|^2]$$
$$\overset{\text{(i)}}{=}\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}\left[\left\|\sum_{m=1}^{N}W_{n,m}(\boldsymbol{y}_m^r-\gamma\boldsymbol{v}_m^r)-\bar{\boldsymbol{y}}^r-\gamma\sum_{k=1}^{q-1}\boldsymbol{v}_n^{r,k}\right\|^2\right]$$
$$=\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}\left[\left\|\sum_{m=1}^{N}W_{n,m}\boldsymbol{y}_m^r-\bar{\boldsymbol{y}}^r-\gamma\sum_{m=1}^{N}W_{n,m}\boldsymbol{v}_m^r-\gamma\bar{\boldsymbol{v}}^r-\gamma\sum_{k=1}^{q-1}(\boldsymbol{v}_n^{r,k}-\bar{\boldsymbol{v}}^r)-q\gamma\bar{\boldsymbol{v}}^r\right\|^2\right]$$
$$\overset{\text{(ii)}}{\leq}4\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}\left[\left\|\sum_{m=1}^{N}W_{n,m}\boldsymbol{y}_m^r-\bar{\boldsymbol{y}}^r\right\|^2\right]+4\gamma^2\sum_{n=1}^{N}\sum_{q=1}^{Q-1}\mathbb{E}\left[\left\|\sum_{m=1}^{N}W_{n,m}\boldsymbol{v}_m^r-\bar{\boldsymbol{v}}^r\right\|^2\right]$$
$$+4\gamma^2\sum_{n=1}^{N}\sum_{q=1}^{Q-1}\mathbb{E}\left[\left\|\sum_{k=1}^{q-1}(\boldsymbol{v}_n^{r,k}-\bar{\boldsymbol{v}}^r)\right\|^2\right]+4\gamma^2\sum_{n=1}^{N}\sum_{q=1}^{Q-1}q^2\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2]$$
$$\leq 4(E-1)\mathbb{E}\left[\left\|\boldsymbol{W}\boldsymbol{Y}^r-\mathbf{1}(\bar{\boldsymbol{y}}^r)^\top\right\|_F^2\right]+4(E-1)\lambda_w^2\gamma^2\mathbb{E}\left[\left\|\boldsymbol{V}^r-\mathbf{1}(\bar{\boldsymbol{v}}^r)^\top\right\|_F^2\right]$$
$$+4\gamma^2\sum_{n=1}^{N}\sum_{q=1}^{E-1}(q-1)\sum_{k=1}^{q-1}\mathbb{E}[\|\boldsymbol{v}_n^{r,k}-\bar{\boldsymbol{v}}^r\|^2]+4\gamma^2N\sum_{q=1}^{E-1}q^2\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2]$$
$$\overset{\text{(iii)}}{\leq}4(E-1)\lambda_w^2\phi_y^r+4(E-1)\lambda_w^2\gamma^2\phi_v^r+4\gamma^2\frac{(E-1)E(2E-1)}{6}N\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2]$$
$$+4\gamma^2\sum_{n=1}^{N}\sum_{q=1}^{E-1}\frac{(q+(E-1))(E-q)}{2}\mathbb{E}[\|\boldsymbol{v}_n^{r,q}-\bar{\boldsymbol{v}}^r\|^2]$$
$$\overset{\text{(iv)}}{\leq}4(E-1)\lambda_w^2\phi_y^r+4(E-1)\lambda_w^2\gamma^2\phi_v^r+2E^2(E-1)N\gamma^2\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2]$$
$$+2E(E-1)\gamma^2\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{v}_n^{r,q}-\bar{\boldsymbol{v}}^r\|^2], \tag{53}$$

其中，(i) 由 (47a) 得出；(ii) 来自詹森不等式 $\sum_{i=1}^{n}f(\underbrace{\quad y_i\quad}_{n})\geq f\left(\underbrace{\sum_{i=1}^{n}y_i}_{n}\right)$，其中 $f(\cdot)=\|\cdot\|^2$；(iii) 的第一项由以下事实得出：$\|\boldsymbol{W}\boldsymbol{Y}^r-\mathbf{1}(\bar{\boldsymbol{y}}^r)^\top\|_F^2=\|(\boldsymbol{W}-\mathbf{1}\mathbf{1}^\top/N)[\boldsymbol{Y}^r-\mathbf{1}(\bar{\boldsymbol{y}}^r)^\top]\|_F^2\leq\lambda_w^2\|\boldsymbol{Y}^r-\mathbf{1}(\bar{\boldsymbol{y}}^r)^\top\|_F^2$；(iii) 的第二项来自 $\sum_{i=1}^{n}i^2=\frac{n(n+1)(2n+1)}{6}$；

(iii) 的最后一项则由以下易于证明的事实得出：

$$\sum_{q=2}^{E}(q-1)\sum_{k=1}^{q-1}y^k \leq \sum_{q=1}^{E}\frac{(q+(E-1))(E-q)}{2}y^q, \tag{54}$$

对于 $y^k > 0, k = 1, \ldots, E-1$；(iv) 中的最后一项源于以下事实：$(q+E-1)(E-q) \leq E(E-1)$。

Next, $\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{v}_n^{r,q}-\bar{\boldsymbol{v}}^r\|^2]$ 的上界可推导如下：

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{v}_n^{r,q}-\bar{\boldsymbol{v}}^r\|^2]\\
&\overset{\text{(i)}}{=}\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}\left[\left\|\sum_{m=1}^{N}W_{n,m}\boldsymbol{v}_m^r-\bar{\boldsymbol{v}}^r+(\boldsymbol{g}_n^{r,q}-\boldsymbol{g}_n^r)\right\|^2\right]\\
&\overset{\text{(ii)}}{\leq} 2\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}\left[\left\|\sum_{m=1}^{N}W_{n,m}\boldsymbol{v}_m^r-\bar{\boldsymbol{v}}^r\right\|^2\right]+2\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{g}_n^{r,q}-\nabla f_n(\boldsymbol{y}_n^{r,q})+\nabla f_n(\boldsymbol{y}_n^{r,q})-\nabla f_n(\bar{\boldsymbol{y}}^r)\\
&\quad+\nabla f_n(\bar{\boldsymbol{y}}^r)-\nabla f_n(\boldsymbol{y}_n^r)+\nabla f_n(\boldsymbol{y}_n^r)-\boldsymbol{g}_n^r\|^2]\\
&\overset{\text{(iii)}}{\leq} 2(E-1)\lambda_w^2\phi_v^r+16(E-1)N\frac{\sigma^2}{|\mathcal{I}|}+8(E-1)L^2\phi_y^r+8L^2\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{y}_n^{r,q}-\bar{\boldsymbol{y}}^r\|^2],
\end{aligned}\tag{55}$$

其中，(i) 由 (47b) 得出；(ii) 源于詹森不等式；(iii) 则依据假设4和假设5。

将 (55) 代入 (53)，可得：

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}\left[\|\boldsymbol{y}_n^{r,q}-\bar{\boldsymbol{y}}^r\|^2\right]\\
&\leq\left[4(E-1)\lambda_w^2+16E(E-1)^2L^2\gamma^2\right]\phi_y^r+4(E-1)\lambda_w^2[1+E(E-1)]\gamma^2\phi_v^r\\
&\quad+2E^2(E-1)N\gamma^2\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2]+32E(E-1)^2N\gamma^2\frac{\sigma^2}{|\mathcal{I}|}\\
&\quad+16E(E-1)L^2\gamma^2\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}\left[\|\boldsymbol{y}_n^{r,q}-\bar{\boldsymbol{y}}^r\|^2\right].
\end{aligned}\tag{56}$$

令 $\gamma \leq \frac{1}{32EL}$，即 $\frac{1}{1-16E(E-1)L^2\gamma^2} \leq 2$。然后，令 $\gamma \leq \frac{\lambda_w^2}{2EL}$ 以简化 (56) 右端（RHS）的第一项，并重新整理 (56)，可得

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}\left[\|\boldsymbol{y}_n^{r,q}-\bar{\boldsymbol{y}}^r\|^2\right]\\
&\leq 16(E-1)\lambda_w^2\phi_y^r+8(E-1)\lambda_w^2[1+E(E-1)]\gamma^2\phi_v^r\\
&\quad+4E^2(E-1)N\gamma^2\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2]+64E(E-1)^2N\gamma^2\frac{\sigma^2}{|\mathcal{I}|}.
\end{aligned}\tag{57}$$

- **(12b) 的证明：** 将 (57) 代入 (55)，可得

$$
\begin{aligned}
&\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}\left[\|\boldsymbol{v}_n^{r,q}-\bar{\boldsymbol{v}}^r\|^2\right]\\
&\le 8(E-1)L^2(1+16\lambda_w^2)\phi_y^r+\left\{2(E-1)\lambda_w^2+64(E-1)\lambda_w^2[1+E(E-1)]\ L^2N^2\gamma^2\right\}\phi_v^r\\
&\quad+32E^2(E-1)NL^2\gamma^2\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2]+16(E-1)N\left[1+32E(E-1)L^2\gamma^2\right]\frac{\sigma^2}{|\mathcal{I}|} \qquad (58)\\
&\le 8(E-1)L^2(1+16\lambda_w^2)\phi_y^r+4(E-1)\lambda_w^2\phi_v^r+32E^2(E-1)NL^2\gamma^2\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2]\\
&\quad+32(E-1)N\frac{\sigma^2}{|\mathcal{I}|}, \qquad (59)
\end{aligned}
$$

其中，(59) 的第 2 项和第 3 项分别由 (58) 的第 2 项和第 3 项令 $\gamma\le\frac{1}{8ENL}$ 和 $\gamma\le\frac{1}{6EL}$ 得到。

因此，求解上述关于 γ $\gamma\le\frac{\lambda_w^2}{32ENL}$ 的条件交集，即可得到引理1中的期望结果。■

C 引理2的证明

由 (48a) 和 (49a)，可得

$$
\begin{aligned}
&\phi_y^{r+1}=\mathbb{E}\left[\left\|\boldsymbol{Y}^{r+1}-\mathbf{1}(\bar{\boldsymbol{y}}^{r+1})^\top\right\|_F^2\right]\\
&=\sum_{n=1}^{N}\mathbb{E}\left[\left\|\sum_{m=1}^{N}W_{n,m}\boldsymbol{y}_m^r-\bar{\boldsymbol{y}}^r-\gamma\left(\sum_{m=1}^{N}W_{n,m}\boldsymbol{v}_m^r-\bar{\boldsymbol{v}}^r\right)-\gamma\frac{1}{N}\sum_{t=1}^{N}\sum_{q=1}^{E-1}(\boldsymbol{v}_n^{r,q}-\boldsymbol{v}_t^{r,q})\right\|^2\right]\\
&\overset{\text{(i)}}{\le}(1+\delta_1)\lambda_w^2\phi_y^r+2\left(1+\frac{1}{\delta_1}\right)\lambda_w^2\gamma^2\phi_v^r\\
&\quad+2\left(1+\frac{1}{\delta_1}\right)\gamma^2\sum_{n=1}^{N}\frac{1}{N}\sum_{t=1}^{N}(E-1)\sum_{q=1}^{E-1}\mathbb{E}\left[\left\|\boldsymbol{v}_n^{r,q}-\bar{\boldsymbol{v}}^r+\bar{\boldsymbol{v}}^r-\boldsymbol{v}_t^{r,q}\right\|^2\right]\\
&\overset{\text{(ii)}}{\le}(1+\delta_1)\lambda_w^2\phi_y^r+2\left(1+\frac{1}{\delta_1}\right)\lambda_w^2\gamma^2\phi_v^r+4\left(1+\frac{1}{\delta_1}\right)(E-1)\gamma^2\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{v}_n^{r,q}-\bar{\boldsymbol{v}}^r\|^2]\\
&\le\left\{(1+\delta_1)\lambda_w^2+32\left(1+\frac{1}{\delta_1}\right)L^2[(1+16\lambda_w^2)(E-1)^2]\gamma^2\right\}\phi_y^r+2\left(1+\frac{1}{\delta_1}\right)\lambda_w^2[1+8(E-1)^2]\gamma^2\phi_v^r\\
&\quad+256\left(1+\frac{1}{\delta_1}\right)E^2(E-1)^2NL^2\gamma^4\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]\\
&\quad+128\left(1+\frac{1}{\delta_1}\right)(E-1)^2\gamma^2(N+2E^2L^2\gamma^2)\frac{\sigma^2}{|\mathcal{I}|}, \qquad (60)
\end{aligned}
$$

其中，(i) 基于杨氏不等式和 $\delta_1>0$；(ii) 由詹森不等式得出；最后一个不等式则由引理1中的 (59) 和 (52) 得出。

取 $\delta_1 = \frac{1-\lambda_w^2}{8\lambda_w^2}$，并令 γ 满足

$$\gamma \leq \frac{(1-\lambda_w^2)^2}{32L^2(1+7\lambda_w^2)[8(1+16\lambda_w^2)(E-1)^2]}, \tag{61}$$

这使得（60）右端的第1项简化为 $\frac{1+3\lambda_w^2}{4}\phi_y^r$，且（60）右端的最后一项中的 $N+2E^2L^2\gamma^2 \leq 2N$。于是，（60）便化为所需结果。■

D 引理3的证明

为便于分析，回顾 $\boldsymbol{g}_n^r \triangleq \frac{1}{|\mathcal{I}|}\sum_{\xi\in\mathcal{I}_n^r} g_n(\boldsymbol{y}_n^r,\ \xi)$，并定义 $\boldsymbol{G}^r = [\boldsymbol{g}_1^r,\ldots,\boldsymbol{g}_N^r]^\top$，$\bar{\boldsymbol{g}}^r = \frac{1}{N}\sum_{n=1}^N \boldsymbol{g}_{n\circ}^r$。于是，可将（48b）和（49b）写成如下紧凑形式

$$\boldsymbol{V}^{r+1} = \boldsymbol{W}\boldsymbol{V}^r + \boldsymbol{G}^{r+1} - \boldsymbol{G}^r, \tag{62}$$

$$\bar{\boldsymbol{v}}^{r+1} = \bar{\boldsymbol{v}}^r + \bar{\boldsymbol{g}}^{r+1} - \bar{\boldsymbol{g}}^r. \tag{63}$$

由（62）和（63），可得

$$\begin{aligned}
\phi_v^{r+1} &= \mathbb{E}\left[\left\|\boldsymbol{V}^{r+1} - \mathbf{1}(\bar{\boldsymbol{v}}^{r+1})^\top\right\|_F^2\right] \\
&= \mathbb{E}\left[\left\|\boldsymbol{W}\boldsymbol{V}^r - \mathbf{1}(\bar{\boldsymbol{v}}^r)^\top + \left(\mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)(\boldsymbol{G}^{r+1} - \boldsymbol{G}^r)\right\|_F^2\right] \\
&\overset{\text{(i)}}{\leq} (1+\delta_2)\lambda_w^2\phi_v^r + \left(1+\frac{1}{\delta_2}\right)\mathbb{E}\left[\left\|\boldsymbol{G}^{r+1} - \boldsymbol{G}^r\right\|_F^2\right] \\
&= (1+\delta_2)\lambda_w^2\phi_v^r + \left(1+\frac{1}{\delta_2}\right)\sum_{n=1}^N \mathbb{E}[\|\boldsymbol{g}_n^{r+1} - \boldsymbol{g}_n^r\|^2] \\
&= (1+\delta_2)\lambda_w^2\phi_v^r + \left(1+\frac{1}{\delta_2}\right)\sum_{n=1}^N \mathbb{E}[\|\boldsymbol{g}_n^{r+1} - \nabla f_n(\boldsymbol{y}_n^{r+1}) \\
&\qquad + \nabla_{\boldsymbol{y}_n} f_n(\boldsymbol{y}_n^{r+1}) - \nabla_{\boldsymbol{y}_n} f_n(\boldsymbol{y}_n^r) + \nabla_{\boldsymbol{y}_n} f_n(\boldsymbol{y}_n^r) - \boldsymbol{g}_n^r\|^2] \\
&\overset{\text{(ii)}}{\leq} (1+\delta_2)\lambda_w^2\phi_v^r + 6\left(1+\frac{1}{\delta_2}\right)N\frac{\sigma^2}{|\mathcal{I}|} + 3\left(1+\frac{1}{\delta_2}\right)L^2\sum_{n=1}^N \mathbb{E}[\|\boldsymbol{y}_n^{r+1} - \boldsymbol{y}_n^r\|^2],
\end{aligned} \tag{64}$$

其中，（i）由杨氏不等式、$\delta_2 > 0$ 以及 $\|\mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\| = 1$ 这一事实得出；（ii）则由假设5得出。

Next，我们求解 $\sum_{n=1}^N \mathbb{E}[\|\boldsymbol{y}_n^{r+1} - \boldsymbol{y}_n^r\|^2]$ 的上界。由 (48a)，可得

$$\begin{aligned}
&\sum_{n=1}^N \mathbb{E}[\|\boldsymbol{y}_n^{r+1} - \boldsymbol{y}_n^r\|^2] \\
&\leq \sum_{n=1}^N \mathbb{E}\left[\left\|\sum_{m=1}^N W_{n,m}\boldsymbol{y}_m^r - \bar{\boldsymbol{y}}^r - (\boldsymbol{y}_n^r - \bar{\boldsymbol{y}}^r) - \gamma\left(\sum_{m=1}^N W_{n,m}\boldsymbol{v}_m^r - \bar{\boldsymbol{v}}^r\right) - \gamma\sum_{q=1}^{E-1}(\boldsymbol{v}_n^{r,q} - \bar{\boldsymbol{v}}^r) - E\gamma\bar{\boldsymbol{v}}^r\right\|^2\right] \\
&\overset{\text{(i)}}{\leq} 5(1+\lambda_w^2)\phi_y^r + 5\lambda_w^2\gamma^2\phi_v^r + 5E^2N\gamma^2\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2] + 5(E-1)\gamma^2\sum_{n=1}^N\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{v}_n^{r,q} - \bar{\boldsymbol{v}}^r\|^2]
\end{aligned}$$

$$\overset{\text{(ii)}}{\leq} [5(1+\lambda_w^2)+40(E-1)^2(1+16\lambda_w^2)L^2\gamma^2]\phi_y^r + 5\lambda_w^2\gamma^2[1+4(E-1)^2]\phi_v^r$$
$$+ 5E^2N\gamma^2[1+32(E-1)^2L^2\gamma^2]\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2] + 160(E-1)^2N\gamma^2\frac{\sigma^2}{|\mathcal{I}|} \tag{65}$$
$$\overset{\text{(iii)}}{\leq} 10(1+\lambda_w^2)\phi_y^r + 5\lambda_w^2[1+4(E-1)^2]\gamma^2\phi_v^r + 10E^2N\gamma^2\mathbb{E}[\|\bar{\boldsymbol{v}}^r\|^2] + 160(E-1)^2N\gamma^2\frac{\sigma^2}{|\mathcal{I}|}, \tag{66}$$

其中，(i) 根据詹森不等式得出；(ii) 是通过在引理1中应用 (12b) 得到的；(iii) 则是令 $\gamma \leq \frac{1}{136EL}$，因此 (66) 的第1项和第3项可分别化简为 (65) 的第1项和第3项。

将 (66) 和 (52) 代入 (64)，得到

$$\begin{aligned}\phi_v^{r+1} \leq & 30\left(1+\frac{1}{\delta_2}\right)(1+\lambda_w^2)L^2\phi_y^r + \left[(1+\delta_2)\lambda_w^2 + 15\left(1+\frac{1}{\delta_2}\right)\lambda_w^2[1+4(E-1)^2]L^2\gamma^2\right]\phi_v^r \\ & + 60\left(1+\frac{1}{\delta_2}\right)E^2NL^2\gamma^2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^N\nabla f(\boldsymbol{y}_n^r)\right\|^2\right] \\ & + 6\left(1+\frac{1}{\delta_2}\right)N\left\{1+80(E-1)^2L^2\gamma^2+\frac{5E^2L^2\gamma^2}{N}\right\}\frac{\sigma^2}{|\mathcal{I}|}.\end{aligned} \tag{67}$$

通过选择 $\delta_2 = \frac{1-\lambda_w^2}{8\lambda_w^2}$，并令 $\gamma \leq \frac{(1-\lambda_w^2)}{3600ENL}$，使得 (67) 的第2项和第4项化为 (15) 中对应的项。■

引理4的证明

将 $\bar{\boldsymbol{y}}^r$ 上的梯度记为 $\nabla F(\bar{\boldsymbol{y}}^r) \triangleq \frac{1}{N}\sum_{n=1}^N \nabla f_n(\bar{\boldsymbol{y}}^r)$。根据假设4中的 Lipschitz 光滑性，我们有

$$\mathbb{E}[F(\bar{\boldsymbol{y}}^{r+1})] \leq \mathbb{E}[F(\bar{\boldsymbol{y}}^r)] + \mathbb{E}[\langle\nabla F(\bar{\boldsymbol{y}}^r), \bar{\boldsymbol{y}}^{r+1}-\bar{\boldsymbol{y}}^r\rangle] + \frac{L}{2}\mathbb{E}[\|\bar{\boldsymbol{y}}^{r+1}-\bar{\boldsymbol{y}}^r\|^2]. \tag{68}$$

这里，$\mathbb{E}\left[\langle\nabla F(\bar{\boldsymbol{y}}^r)\ \bar{\boldsymbol{y}}^{r+1}-\bar{\boldsymbol{y}}^r\rangle\right]$ 可以简化为

$$\begin{aligned}&\mathbb{E}\left[\langle\nabla F(\bar{\boldsymbol{y}}^r), \bar{\boldsymbol{y}}^{r+1}-\bar{\boldsymbol{y}}^r\rangle\right] \\ &\overset{\text{(i)}}{=} \mathbb{E}\left[\left\langle\nabla F(\bar{\boldsymbol{y}}^r), -\gamma\bar{\boldsymbol{v}}^r - \gamma\frac{1}{N}\sum_{n=1}^N\sum_{q=1}^{E-1}\boldsymbol{v}_n^{r,q}\right\rangle\right] \\ &\overset{\text{(ii)}}{=} -\gamma\mathbb{E}\left[\left\langle\nabla F(\bar{\boldsymbol{y}}^r), \bar{\boldsymbol{v}}^r + \frac{1}{N}\sum_{n=1}^N\sum_{q=1}^{E-1}\sum_{m=1}^N W_{n,m}\boldsymbol{v}_m^r\right\rangle\right] - \gamma\mathbb{E}\left[\left\langle\nabla F(\bar{\boldsymbol{y}}^r), \frac{1}{N}\sum_{n=1}^N\sum_{q=1}^{E-1}\boldsymbol{g}_n^{r,q}\right\rangle\right] \\ &\quad + \gamma(E-1)\mathbb{E}\left[\left\langle\nabla F(\bar{\boldsymbol{y}}^r), \frac{1}{N}\sum_{n=1}^N\boldsymbol{g}_n^r\right\rangle\right] \\ &\overset{\text{(iii)}}{=} -\gamma E\mathbb{E}\left[\left\langle\nabla F(\bar{\boldsymbol{y}}^r), \bar{\boldsymbol{v}}^r\right\rangle\right] - \gamma\mathbb{E}\left[\left\langle\nabla F(\bar{\boldsymbol{y}}^r), \frac{1}{N}\sum_{n=1}^N\sum_{q=1}^{E-1}\nabla f_n(\boldsymbol{y}_n^{r,q})\right\rangle\right]\end{aligned}$$

$$
\begin{aligned}
&+\gamma(E-1)\mathbb{E}\left[\left\langle \nabla F(\bar{\boldsymbol{y}}^r), \frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\rangle\right] \\
&\overset{\text{(iv)}}{=} -\gamma\mathbb{E}\left[\left\langle \nabla F(\bar{\boldsymbol{y}}^r), \frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\rangle\right] - \gamma\sum_{q=1}^{E-1}\mathbb{E}\left[\left\langle \nabla F(\bar{\boldsymbol{y}}^r), \frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^{r,q})\right\rangle\right] \\
&\overset{\text{(v)}}{=} -\gamma\left\{\frac{1}{2}\mathbb{E}[\|\nabla F(\bar{\boldsymbol{y}}^r)\|^2] + \frac{1}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] - \frac{1}{2}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{y}}^r) - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]\right\} \\
&\quad -\gamma\left\{\frac{E-1}{2}\mathbb{E}[\|\nabla F(\bar{\boldsymbol{y}}^r)\|^2] + \frac{1}{2}\sum_{q=1}^{E-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^{r,q})\right\|^2\right]\right. \\
&\qquad \left. -\frac{1}{2}\sum_{q=1}^{E-1}\mathbb{E}\left[\left\|\nabla F(\bar{\boldsymbol{y}}^r) - \frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^{r,q})\right\|^2\right]\right\} \\
&\le -\frac{\gamma E}{2}\|\nabla F(\bar{\boldsymbol{y}}^r)\|^2 - \frac{\gamma}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right] + \frac{\gamma L^2}{2N}\phi_y^r + \frac{\gamma L^2}{2N}\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{y}_n^{r,q} - \bar{\boldsymbol{y}}^r\|^2],
\end{aligned}
\tag{69}
$$

其中，(i) 根据 (49a) 得出；(ii) 源于 (47b)；(iii) 的第一项由 $\boldsymbol{W}$ 的双随机性质得出；(iii) 的第二项和第三项由假设5中的无偏梯度得到；(iv) 源于 (51)；(v) 根据如下事实得出：对于某些向量 $\boldsymbol{a}$，$\boldsymbol{b}$，$\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \frac{1}{2}(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\boldsymbol{a}-\boldsymbol{b}\|^2)$；最后一个不等式则是通过略去负项 $-\frac{\gamma}{2}\sum_{q=1}^{E-1}\left\|\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}^r, {}_n{}^q)\right\|^2$ 得到的。

Next，我们求解（68）中 $\frac{L}{2}\mathbb{E}\left[\|\bar{\boldsymbol{y}}^{r+1} - \bar{\boldsymbol{y}}^r\|^2\right]$ 的上界。由（49a）可得

$$
\begin{aligned}
&\frac{L}{2}\mathbb{E}\left[\|\bar{\boldsymbol{y}}^{r+1} - \bar{\boldsymbol{y}}^r\|^2\right] \\
&= \frac{L\gamma^2}{2}\mathbb{E}\left[\left\|\bar{\boldsymbol{v}}^r + \frac{1}{N}\sum_{n=1}^{N}\sum_{q=1}^{E-1}\boldsymbol{v}_n^{r,q}\right\|^2\right] \\
&\overset{\text{(i)}}{=} \frac{L\gamma^2}{2}\mathbb{E}\left[\left\|E\bar{\boldsymbol{v}}^r + \frac{1}{N}\sum_{n=1}^{N}\sum_{q=1}^{E-1}(\boldsymbol{g}_n^{r,q} - \boldsymbol{g}_n^r)\right\|^2\right] \\
&\overset{\text{(ii)}}{\le} L\gamma^2\mathbb{E}\left[\left\|E\left(\bar{\boldsymbol{g}}^r - \frac{1}{N}\sum_{n=1}^{N}\nabla f(\boldsymbol{y}_n^r)\right) + \frac{E}{N}\sum_{n=1}^{N}\nabla f(\boldsymbol{y}_n^r)\right\|^2\right] \\
&\quad + 3L\gamma^2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\sum_{q=1}^{E-1}[\boldsymbol{g}_n^{r,q} - \nabla f_n(\boldsymbol{y}_n^{r,q})]\right\|^2\right] + 3L\gamma^2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\sum_{q=1}^{E-1}[\boldsymbol{g}_n^r - \nabla f_n(\boldsymbol{y}_n^r)]\right\|^2\right] \\
&\quad + 3L\gamma^2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\sum_{q=1}^{E-1}[\nabla f_n(\boldsymbol{y}_n^{r,q}) - \nabla f_n(\boldsymbol{y}_n^r)]\right\|^2\right] \\
&\overset{\text{(iii)}}{\le} 2EL\gamma^2\mathbb{E}\left[\left\|\bar{\boldsymbol{g}}^r - \frac{1}{N}\sum_{n=1}^{N}\nabla f(\boldsymbol{y}_n^r)\right\|^2\right] + 2E^2L\gamma^2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f(\boldsymbol{y}_n^r)\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+3L\gamma^2\frac{1}{N^2}\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{g}_n^{r,q}-\nabla f_n(\boldsymbol{y}_n^{r,q})\|^2]+3L\gamma^2\frac{1}{N^2}\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{g}_n^{r}-\nabla f_n(\boldsymbol{y}_n^{r})\|^2]\\
&+3(E-1)NL^3\gamma^2\frac{1}{N}\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{y}_n^{r,q}-\boldsymbol{y}_n^{r}\|^2]\\
&\overset{\text{(iv)}}{\leq} 2E^2L\gamma^2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f(\boldsymbol{y}_n^r)\right\|^2\right]+2(4E-3)L\gamma^2\frac{\sigma^2}{N|\mathcal{I}|}\\
&+\frac{6(E-1)^2L^3\gamma^2}{N}\left[\phi_y^r+\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{y}_n^{r,q}-\bar{\boldsymbol{y}}^r\|^2]\right],
\end{aligned}
\tag{70}
$$

其中，（i）由（47b）得出；（ii）由詹森不等式得出；（iii）的第二项和第三项由事实 $\|\sum_{n=1}^{N}(\boldsymbol{x}_n-\mathbb{E}[\boldsymbol{x}_n]\|^2=\sum_{n=1}^{N}\|\boldsymbol{x}_n-\mathbb{E}[\boldsymbol{x}_n]\|^2$，$\forall \boldsymbol{x}$，并依据假设5得出；（iii）的最后一项依据假设4得出；（iv）由假设5中的有界方差得出。

将 (69) 和 (70) 相加，我们得到

$$
\begin{aligned}
&\mathbb{E}\left[\langle\bar{\nabla}_{\boldsymbol{y}_n}^r,\bar{\boldsymbol{y}}^{r+1}-\bar{\boldsymbol{y}}^r\rangle\right]+\frac{L}{2}\mathbb{E}\left[\|\bar{\boldsymbol{y}}^{r+1}-\bar{\boldsymbol{y}}^r\|^2\right]\\
&\leq -\frac{\gamma E}{2}\|\bar{\nabla}_{\boldsymbol{y}}^r\|^2-\frac{\gamma}{2}[1-4E^2L\gamma]\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]+\frac{\gamma L^2}{2N}[1+12(E-1)^2L\gamma]\phi_y^r\\
&\quad+\frac{\gamma L^2}{2N}[1+12(E-1)^2L\gamma]\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{y}_n^{r,q}-\bar{\boldsymbol{y}}^r\|^2]+2(4E-3)L\gamma^2\frac{\sigma^2}{N|\mathcal{I}|}
\end{aligned}
\tag{71}
$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} -\frac{\gamma E}{2}\|\bar{\nabla}_{\boldsymbol{y}}^r\|^2-\frac{\gamma}{2}[1-4E^2L\gamma]\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]\\
&\quad+\frac{L^2\gamma}{N}\phi_y^r+\frac{L^2\gamma}{N}\sum_{n=1}^{N}\sum_{q=1}^{E-1}\mathbb{E}[\|\boldsymbol{y}_n^{r,q}-\bar{\boldsymbol{y}}^r\|^2]+2(4E-3)L\gamma^2\frac{\sigma^2}{N|\mathcal{I}|}
\end{aligned}
\tag{72}
$$

$$
\begin{aligned}
&\overset{\text{(ii)}}{\leq} -\frac{\gamma E}{2}\|\bar{\nabla}_{\boldsymbol{y}}^r\|^2-\frac{\gamma}{2}\left[1-4E^2L\gamma-16E^2(E-1)L^2\gamma^2\right]\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]\\
&\quad+\frac{L^2\gamma}{N}[1+16(E-1)\lambda_w^2]\phi_y^r+8(E-1)\lambda_w^2[1+E(E-1)]\frac{L^2\gamma^3}{N}\phi_v^r\\
&\quad+\frac{2L\gamma^2}{N}\left[4E-3+32E(E-1)^2NL\gamma\right]\frac{\sigma^2}{|\mathcal{I}|}
\end{aligned}
\tag{73}
$$

$$
\begin{aligned}
&\overset{\text{(iii)}}{\leq} -\frac{\gamma E}{2}\|\bar{\nabla}_{\boldsymbol{y}}^r\|^2-\frac{\gamma}{2}\left[1-2E^2L\gamma-8E^2(E-1)L^2\gamma^2\right]\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla f_n(\boldsymbol{y}_n^r)\right\|^2\right]\\
&\quad+\frac{L^2\gamma}{N}[1+16(E-1)\lambda_w^2]\phi_y^r+10EL\gamma^2\frac{\sigma^2}{N|\mathcal{I}|}+8(E-1)\lambda_w^2[1+E(E-1)]\frac{L^2\gamma^3}{N}\phi_v^r,
\end{aligned}
\tag{74}
$$

其中，(i) 和 (iii) 是通过令

$$
\gamma\leq\min\left\{\frac{1}{12E^2L},\frac{1}{32E^2NL}\right\}=\frac{1}{32E^2NL},
\tag{75}
$$

从而使 (71) 的第 3 项化简为 (72) 的第 3 项，并使 (73) 的最后一项化简为 (74) 的最后一项；(ii) 则是由引理1中的 (12b) 得到的。

因此，结合 (68) 和 (74)，即可得到所需结果。 ■

F 事实 $\rho(\boldsymbol{A}) < 1$ 的证明

假设 $\boldsymbol{s} = [s_1, s_2]^\top \in \mathbb{R}^2$。通过等价地求解，可以验证 $\boldsymbol{A}\boldsymbol{s} < \boldsymbol{s}$

$$
\begin{aligned}
&\begin{cases} \boldsymbol{A}_{1,1}s_1 + \boldsymbol{A}_{1,2}s_2 < s_1, \\ \boldsymbol{A}_{2,1}s_1 + \boldsymbol{A}_{2,2}s_2 < s_2, \end{cases} \\
\Leftarrow &\begin{cases} \boldsymbol{A}_{1,2}s_2 < (1 - \boldsymbol{A}_{1,1})s_1, \\ \boldsymbol{A}_{2,1}s_1 < (1 - \boldsymbol{A}_{2,2})s_2, \end{cases} \\
\Leftarrow &\begin{cases} 2\frac{(1+7\lambda_w^2)}{1-\lambda_w^2}\left[1 + 8(E-1)^2\lambda_w^2\right] s_2\gamma^2 < \frac{3(1-\lambda_w^2)}{4}s_1, \\ 30\frac{(1+7\lambda_w^2)(1+\lambda_w^2)}{1-\lambda_w^2}N^2L^2s_1 < \frac{3(1-\lambda_w^2)}{4}s_2, \end{cases} \\
\Leftarrow &\begin{cases} \gamma^2 \le \frac{3(1-\lambda_w^2)^2 s_1}{16(1+7\lambda_w^2)[1+4(E-1)^2\lambda_w^2]s_2} < \frac{3(1-\lambda_w^2)^2 s_1}{8(1+7\lambda_w^2)[1+8(E-1)^2\lambda_w^2]S_2}, \\ s_1 < \frac{3(1-\lambda_w^2)}{120\frac{(1+7\lambda_w^2)(1+\lambda_w^2)}{1-\lambda_w^2}N^2L^2}s_2. \end{cases}
\end{aligned} \tag{76}
$$

因此，对于满足以下条件的 $\boldsymbol{s}$ 和 γ

$$
\begin{aligned}
s_1 &< \frac{(1-\lambda_w^2)^2}{40(1+7\lambda_w^2)(1+\lambda_w^2)L^2}s_2, \\
\gamma &\le \frac{3(1-\lambda_w^2)s_1}{16(1+7\lambda_w^2)[1+8(E-1)\lambda_w]s_2},
\end{aligned}
$$

我们有 $\boldsymbol{A}\boldsymbol{s} < \boldsymbol{s}$。然后可以得到 $\rho(\boldsymbol{A}) < 1$。 ■

G 引理 5 的证明

根据 $(\boldsymbol{I} - \boldsymbol{A})^{-1} = (\boldsymbol{I} - \boldsymbol{A})^* / \det(\boldsymbol{I} - \boldsymbol{A})$ 这一规则，我们将推导分为两步，分别求解行列式和伴随矩阵 $(\boldsymbol{I} - \boldsymbol{A})^*$。

- **求解行列式** $\det(\boldsymbol{I} - \boldsymbol{A})$**：**

回顾 $\boldsymbol{A}$ 式（18），我们有

$$
\begin{aligned}
&\det(\mathbf{I} - \boldsymbol{A}) \\
&= (1 - A_{1,1})(1 - A_{2,2}) - A_{1,2}A_{2,1} \\
&= \frac{9(1-\lambda_w^2)^2}{16} - 60\frac{(1+7\lambda_w^2)^2(1+\lambda_w^2)}{(1-\lambda_w^2)^2}[1 + 8(Q-1)^2\lambda_w^2]L^2\gamma^2 \\
&\ge \frac{(1-\lambda_w^2)^2}{8},
\end{aligned} \tag{77}
$$

其中，最后一个不等式是通过令

$$
\gamma \le \frac{3(1-\lambda_w^2)^2}{320(1+7\lambda_w^2)(1+\lambda_w^2)[1+8(Q-1)\lambda_w]L}.
$$

• **求行列式 $(\boldsymbol{I}-\boldsymbol{A})^*$:**

$$(\mathbf{I}-\boldsymbol{A})^* = \begin{bmatrix} 1-G_{2,2} & A_{1,2} \\ A_{2,1} & 1-A_{1,1} \end{bmatrix} \leq \begin{bmatrix} 1 & \frac{2(1+7\lambda_w^2)[1+8(Q-1)^2\lambda_w^2]\gamma^2}{1-\lambda_w^2} \\ \frac{30(1+7\lambda_w^2)(1+\lambda_w^2)L^2}{1-\lambda_w^2} & 1 \end{bmatrix}, \tag{78}$$

其中，最后一个不等式是逐元素成立的。

结合式（77）和式（78），即可得到所需结果。 ■

H 算法 2 的另一种表达形式

为便于分析，我们可以将第 r 轮第 q 次迭代时 $n \in [N]$ 的式（45）写为

$$\boldsymbol{\theta}_n^{r,q} = \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{\theta}_m^r - \alpha \boldsymbol{g}_{\theta,n}^r - \alpha \sum_{k=1}^{q-1} \boldsymbol{g}_{\theta,n}^{r,k}, \tag{79a}$$

$$\boldsymbol{x}_n^{r,q} = \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{x}_m^r - \beta \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{u}_m^r - \beta \sum_{k=1}^{q-1} \boldsymbol{u}_n^{r,k}, \tag{79b}$$

$$\boldsymbol{z}_{n,i}^{r,q} = \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{z}_{m,i}^r + N\boldsymbol{B}_{n,i}\left(\boldsymbol{x}_n^{r,q} - \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{x}_m^r\right), \forall i \in [S], \tag{79c}$$

$$\boldsymbol{u}_n^{r,q} = \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{u}_m^r + \boldsymbol{g}_{x,n}^{r,q} - \boldsymbol{g}_{x,n}^r. \tag{79d}$$

将 $q = E$ 代入式（79），可得

$$\boldsymbol{\theta}_n^{r+1} = \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{\theta}_m^r - \alpha \boldsymbol{g}_{\theta,n}^r - \alpha \sum_{q=1}^{E-1} \boldsymbol{g}_{\theta,n}^{r,q}, \tag{80a}$$

$$\boldsymbol{x}_n^{r+1} = \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{x}_m^r - \beta \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{u}_m^r - \beta \sum_{q=1}^{E-1} \boldsymbol{u}_n^{r,q}, \tag{80b}$$

$$\boldsymbol{z}_{n,i}^{r+1} = \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{z}_{m,i}^r + N\boldsymbol{B}_{n,i}\left(\boldsymbol{x}_n^{r,q} - \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{x}_m^r\right), \forall i \in [S], \tag{80c}$$

$$\boldsymbol{u}_n^{r+1} = \sum_{m=1}^{N} \boldsymbol{W}_{n,m}\boldsymbol{u}_m^r + \boldsymbol{g}_{x,n}^{r+1} - \boldsymbol{g}_{x,n}^r. \tag{80d}$$

对上述子公式取平均，得到

$$\bar{\boldsymbol{\theta}}^{r+1} = \bar{\boldsymbol{\theta}}^r - \alpha \bar{\boldsymbol{g}}_\theta^r - \alpha \frac{1}{N} \sum_{n=1}^{N} \sum_{q=1}^{E-1} \boldsymbol{g}_{\theta,n}^{r,q}, \tag{81a}$$

$$\bar{\boldsymbol{x}}^{r+1} = \bar{\boldsymbol{x}}^r - \beta \bar{\boldsymbol{u}}^r - \beta \frac{1}{N} \sum_{n=1}^{N} \sum_{q=1}^{E-1} \boldsymbol{u}_n^{r,q}, \tag{81b}$$

$$\bar{\boldsymbol{z}}_i^{r+1} = \bar{\boldsymbol{z}}_i^r + \sum_{n=1}^N \boldsymbol{B}_{n,i}\left(\boldsymbol{x}_n^{r+1} - \sum_{m=1}^N \boldsymbol{W}_{n,m}\boldsymbol{x}_m^r\right), \forall i \in [S], \tag{81c}$$

$$\bar{\boldsymbol{u}}^{r+1} = \bar{\boldsymbol{u}}^r + \bar{\boldsymbol{g}}_x^{r+1} - \bar{\boldsymbol{g}}_x^r, \tag{81d}$$

其中 $\bar{\theta}^r \triangleq \frac{1}{N}\sum_{n=1}^N \boldsymbol{\theta}_n^r$，而 $\bar{\boldsymbol{x}}^r$, $\bar{\boldsymbol{z}}^r$, $\bar{\boldsymbol{u}}^r$, $\bar{\boldsymbol{g}}_\theta^r$, $\bar{\boldsymbol{g}}_x^r$ 的定义方式相同。

与引理 6 类似，我们有

引理 7 对于所有 $n \in [N]$，初始化 $\boldsymbol{z}_{n,\,i}^0 = N\boldsymbol{B}_{n,\,i}\boldsymbol{x}_n^0, \forall i \in [S]$, $\boldsymbol{u}_n^0 = \boldsymbol{g}_{x,\,n}^0$，由 (80c)和(80d)，我们有

$$\bar{\boldsymbol{z}}_i^r = \sum_{n=1}^N \boldsymbol{B}_{n,i}\boldsymbol{x}_n^r, \forall i \in [S], \tag{82a}$$

$$\bar{\boldsymbol{u}}^r = \bar{\boldsymbol{g}}_x^r, \forall r \geq 0. \tag{82b}$$

然后，与 (51) 和 (52) 类似，我们有

$$\mathbb{E}[\bar{\boldsymbol{u}}^r] = \mathbb{E}[\bar{\boldsymbol{g}}_x^r] = \frac{1}{N}\sum_{n=1}^N \frac{1}{S}\sum_{i=1}^S \boldsymbol{B}_{n,i}^\top \nabla_{\boldsymbol{z}} f(\boldsymbol{z}_{n,i}^r, \boldsymbol{\theta}_n^r), \tag{83}$$

$$\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2] = \mathbb{E}[\|\bar{\boldsymbol{g}}_x^r\|^2] \leq \frac{2\sigma}{N|\mathcal{I}|} + 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{n=1}^N \frac{1}{S}\sum_{i=1}^S \boldsymbol{B}_{n,i}^\top \nabla_{\boldsymbol{z}} f(\boldsymbol{z}_{n,i}^r, \boldsymbol{\theta}_n^r)\right\|^2\right]. \tag{84}$$

为证明收敛，对于第 r 轮，我们记 $\boldsymbol{\Theta}^r = [\boldsymbol{\theta}_1^r, \cdots, \boldsymbol{\theta}_N^r]^\top$，并以相同方式定义 $\boldsymbol{X}^r$、$\boldsymbol{Z}^r$ 和 $\boldsymbol{U}^r$。基于这些记号，我们定义

$$\hat{\boldsymbol{\phi}}^r = \begin{bmatrix}\phi_\theta^r \\ \phi_x^r \\ \phi_z^r \\ \phi_w^r\end{bmatrix} = \begin{bmatrix}\mathbb{E}\left[\|\boldsymbol{\Theta}^r - \mathbf{1}(\bar{\boldsymbol{\theta}}^r)^\top\|_F^2\right] \\ \mathbb{E}\left[\|\boldsymbol{X}^r - \mathbf{1}(\bar{\boldsymbol{x}}^r)^\top\|_F^2\right] \\ \mathbb{E}\left[\|\boldsymbol{Z}^r - \mathbf{1}(\bar{\boldsymbol{z}}^r)^\top\|_F^2\right] \\ \mathbb{E}\left[\|\boldsymbol{U}^r - \mathbf{1}(\bar{\boldsymbol{u}}^r)^\top\|_F^2\right]\end{bmatrix} \in \mathbb{R}^4 \tag{85}$$

作为共识误差矩阵，其中 $\|\cdot\|_F$ 表示矩阵 Frobenius 范数。此外，平均 $\boldsymbol{\theta}^r$ 和 $\bar{\boldsymbol{x}}^r$ 上的 $\bar{}$ 梯度定义为

$$\bar{\nabla}_{\boldsymbol{\theta}}^r \triangleq \nabla_{\boldsymbol{\theta}}\tilde{F}(\bar{\boldsymbol{\theta}}^r, \bar{\boldsymbol{x}}^r) = \frac{1}{NS}\sum_{i=1}^S \nabla_{\boldsymbol{\theta}} f\left(\boldsymbol{B}_i\bar{\boldsymbol{x}}^r, \bar{\boldsymbol{\theta}}^r\right), \tag{86}$$

$$\bar{\nabla}_{\boldsymbol{\theta}}^r \triangleq \nabla_{\boldsymbol{x}}\tilde{F}(\bar{\boldsymbol{\theta}}^r, \bar{\boldsymbol{x}}^r) = \frac{1}{NS}\sum_{i=1}^S \boldsymbol{B}_i^\top \nabla_{\boldsymbol{B}_i\bar{\boldsymbol{x}}^r} f\left(\boldsymbol{B}_i\bar{\boldsymbol{x}}^r, \bar{\boldsymbol{\theta}}^r\right). \tag{87}$$

定理 2 的证明

I.1 关键引理

引理 8 （局部更新）假设 *1* 至假设 *4* 以及假设 *6* 均成立。对于足够小的 α 且满足条件的 β

$$\alpha \leq \min\left\{\frac{\lambda_w}{20EB_{\max}L}, \frac{\lambda_w}{4EL}\right\}, \tag{88}$$

$$\beta \leq \min\left\{\frac{\lambda_w}{28ENB_{\max}L}, \frac{\lambda_w}{56ENB_{\max}^2L}, \frac{\lambda_w}{272ENB_{\max}^3L}\right\}, \tag{89}$$

我们有由所提出的 *MUST*算法生成的局部步变量 $\boldsymbol{\theta}^{r,q}_n, \boldsymbol{x}^{r,q}_n, \boldsymbol{z}^{r,q}_n, \boldsymbol{u}^{r,q}_n$与$^-$对应平均值 $\boldsymbol{\theta}^r, \bar{\boldsymbol{x}}^r, \bar{\boldsymbol{z}}^r, \bar{\boldsymbol{u}}^r$ 之间距离的如下动态关系

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{q=1}^{E-1}\frac{1}{S}\sum_{i=1}^{S}\left\|\boldsymbol{z}^{r,q}_{n,i}-\bar{\boldsymbol{z}}^r_i\right\|^2+\sum_{n=1}^{N}\sum_{q=1}^{E-1}\left\|\boldsymbol{\theta}^{r,q}_n-\bar{\boldsymbol{\theta}}^r\right\|^2\\
&\leq 7(E-1)\lambda_w^2\phi_\theta^r+200(E-1)N^2B_{\max}^2\lambda_w^2\phi_x^r+32(E-1)\lambda_w^2\frac{1}{S}\phi_z^r\\
&\quad+128(E-1)[1+E(E-1)]N^2B_{\max}^2\lambda_w^2\beta^2\phi_u^r+2E(E-1)N[5\alpha^2+768NB_{\max}^4L^2\beta^2]\frac{\sigma^2}{|\mathcal{I}|}\\
&\quad+32E^2(E-1)N\alpha^2\mathbb{E}[\|\bar{\nabla}^r_{\boldsymbol{\theta}}\|^2]+48E^2(E-1)N^2B_{\max}^2\beta^2\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2],
\end{aligned}\tag{90a}$$

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{q=1}^{E-1}\|\boldsymbol{u}^{r,q}_n-\bar{\boldsymbol{u}}^r\|^2\\
&\leq 12(E-1)(1+7\lambda_w^2)B_{\max}^2L^2\phi_\theta^r+2400(E-1)N^2B_{\max}^4L^2\lambda_w^2\phi_x^r\\
&\quad+12(E-1)(1+32\lambda_w^2)B_{\max}^2L^2\frac{1}{S}\phi_z^r+3(E-1)\lambda_w^2\phi_u^r\\
&\quad+24E(E-1)NB_{\max}^2L^2[5\alpha^2+768NB_{\max}^4L^2\beta^2]\frac{\sigma^2}{|\mathcal{I}|}+384E^2(E-1)NB_{\max}^2L^2\alpha^2\mathbb{E}[\|\bar{\nabla}^r_{\boldsymbol{\theta}}\|^2]\\
&\quad+576E^2(E-1)N^2B_{\max}^4L^2\beta^2\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2],
\end{aligned}\tag{90b}$$

and

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{q=1}^{E-1}\|\boldsymbol{x}^{r,q}_n-\bar{\boldsymbol{x}}^r\|^2\\
&\leq 24E(E-1)^2(1+7\lambda_w^2)B_{\max}^2L^2\beta^2\phi_\theta^r+5(E-1)\lambda_w^2\phi_x^r+24E(E-1)^2(1+32\lambda_w^2)B_{\max}^2L^2\frac{1}{S}\beta^2\phi_z^r\\
&\quad+5(E-1)\beta^2[1+E(E-1)]\lambda_w^2\phi_u^r+25E(E-1)^2N\beta^2\frac{\sigma^2}{|\mathcal{I}|}\\
&\quad+3E^2(E-1)\beta^2\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]+768E^3(E-1)^2NB_{\max}^2L^2\alpha^2\beta^2\mathbb{E}[\|\bar{\nabla}^r_{\boldsymbol{\theta}}\|^2].
\end{aligned}\tag{90c}$$

引理 9 （一致性误差）设假设 *1* 至假设*4*以及假设 *6* 均成立。对于足够小的 α 和满足以下条件的 β

$$\alpha \leq \frac{1}{144EL}, \quad \beta \leq \frac{1}{109ENB_{\max}^2L}, \tag{91}$$

*MUST*算法生成的迭代序列具有以下收缩性质

$$\begin{aligned}
\phi_\theta^{r+1}&=\mathbb{E}\left[\left\|\boldsymbol{\Theta}^{r+1}-\mathbf{1}(\bar{\boldsymbol{\theta}}^{r+1})^\top\right\|^2\right]\\
&\leq\frac{1+3\lambda_w^2}{4}\phi_\theta^r+80\frac{1+7\lambda_w^2}{1-\lambda_w^2}NB_{\max}^2L^2[E^2+100(E-1)\lambda_w^2N]\alpha^2\phi_x^r\\
&\quad+40\frac{1+7\lambda_w^2}{1-\lambda_w^2}[1+32(E-1)\lambda_w^2]\frac{1}{S}L^2\alpha^2\phi_z^r+5120\frac{1+7\lambda_w^2}{1-\lambda_w^2}(E-1)[1+E(E-1)]
\end{aligned}$$

$$
\begin{aligned}
&N^2B_{\max}^2L^2\alpha^2\lambda_w^2\beta^2\phi_u^r+48\frac{1+7\lambda_w^2}{1-\lambda_w^2}EN\alpha^2\frac{\sigma^2}{|\mathcal{I}|}\\
&+88\frac{1+7\lambda_w^2}{1-\lambda_w^2}E^2\alpha^2\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]\\
&+1920\frac{1+7\lambda_w^2}{1-\lambda_w^2}E^2(E-1)N^2B_{\max}^2L^2\alpha^2\beta^2\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2],
\end{aligned}
\tag{92a}
$$

and

$$
\begin{aligned}
\phi_x^r&=\mathbb{E}\left[\left\|\boldsymbol{X}^r-\mathbf{1}(\bar{\boldsymbol{x}}^r)^\top\right\|^2\right]\\
&\le 96\frac{1+7\lambda_w^2}{1-\lambda_w^2}(E-1)^2(1+7\lambda_w^2)B_{\max}^2L^2\beta^2\phi_\theta^r+\frac{1+3\lambda_w^2}{4}\phi_x^r\\
&\quad+96\frac{1+7\lambda_w^2}{1-\lambda_w^2}(E-1)^2(1+32\lambda_w^2)B_{\max}^2L^2\frac{1}{S}\beta^2\phi_z^r+2\frac{1+7\lambda_w^2}{1-\lambda_w^2}\beta^2[1+12(E-1)^2\lambda_w^2]\phi_u^r\\
&\quad+192\frac{1+7\lambda_w^2}{1-\lambda_w^2}E(E-1)^2NB_{\max}^2L^2[5\alpha^2+768NB_{\max}^4L^2\beta^2]\beta^2\frac{\sigma^2}{|\mathcal{I}|}\\
&\quad+3072\frac{1+7\lambda_w^2}{1-\lambda_w^2}E^2(E-1)^2NB_{\max}^2L^2\alpha^2\beta^2\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]\\
&\quad+4608\frac{1+7\lambda_w^2}{1-\lambda_w^2}E^2(E-1)^2N^2B_{\max}^4L^2\beta^4\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2].
\end{aligned}
\tag{92b}
$$

引理 10 （跟踪误差）设假设 *1* 至假设*4*以及假设 *6* 均成立。对于足够小的 α 和满足以下条件的 β

$$
\alpha\le\frac{1}{78EL},\ \beta\le\min\left\{\frac{1}{712ENB_{\max}^2L},\frac{1}{EN\lambda_w}\right\},
\tag{93}
$$

我们得到了 *MUST*算法生成的迭代点所具有的以下收缩性质

$$
\begin{aligned}
\phi_z^{r+1}&=\mathbb{E}\left[\left\|\boldsymbol{Z}^{r+1}-\mathbf{1}(\bar{\boldsymbol{z}}^{r+1})^\top\right\|^2\right]\\
&\le 144\frac{1+7\lambda_w^2}{1-\lambda_w^2}(E-1)^2(1+7\lambda_w^2)N^2SB_{\max}^4L^2\beta^2\phi_\theta^r\\
&\quad+28800\frac{1+7\lambda_w^2}{1-\lambda_w^2}(E-1)^2N^4SB_{\max}^6L^2\lambda_w^2\beta^2\phi_x^r\\
&\quad+\frac{1+3\lambda_w^2}{4}\phi_z^r+12\frac{1+7\lambda_w^2}{1-\lambda_w^2}[1+3(E-1)^2\lambda_w^2]N^2SB_{\max}^2\beta^2\phi_u^r\\
&\quad+288\frac{1+7\lambda_w^2}{1-\lambda_w^2}E(E-1)^2N^3SB_{\max}^4L^2[5\alpha^2+768NB_{\max}^4L^2\beta^2]\beta^2\frac{\sigma^2}{|\mathcal{I}|}\\
&\quad+4608\frac{1+7\lambda_w^2}{1-\lambda_w^2}E^2(E-1)^2N^3SB_{\max}^4L^2\alpha^2\beta^2\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]+24\frac{1+7\lambda_w^2}{1-\lambda_w^2}E^2N^3SB_{\max}^2\beta^2\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2].
\end{aligned}
\tag{94a}
$$

and

$$
\phi_u^{r+1}=\mathbb{E}\left[\left\|\boldsymbol{U}^{r+1}-\mathbf{1}(\bar{\boldsymbol{u}}^{r+1})^\top\right\|^2\right]
$$

$$
\begin{aligned}
&\leq 49\frac{1+7\lambda_w^2}{1-\lambda_w^2}(1+\lambda_w^2)B_{\max}^2L^2\phi_\theta^r \\
&\quad+12\frac{1+7\lambda_w^2}{1-\lambda_w^2}\left\{21600(E-1)^2N^4B_{\max}^6L^2\lambda_w^2\beta^2+30NB_{\max}^2L^2[E^2+160(E-1)\lambda_w^2N]\alpha^2\right\}\phi_x^r \\
&\quad+49\frac{1+7\lambda_w^2}{1-\lambda_w^2}(1+\lambda_w^2)B_{\max}^2L^2\frac{1}{S}\phi_z^r+\frac{1+3\lambda_w^2}{4}\phi_u^r+25\frac{1+7\lambda_w^2}{1-\lambda_w^2}N\frac{\sigma^2}{|\mathcal{I}|} \\
&\quad+397\frac{1+7\lambda_w^2}{1-\lambda_w^2}E^2B_{\max}^2L^2\alpha^2\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]+217\frac{1+7\lambda_w^2}{1-\lambda_w^2}E^2N^3B_{\max}^4L^2\beta^2\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2],
\end{aligned}
\tag{94b}
$$

我们建立了一个关键的下降不等式，用以刻画全局目标函数 F 从 $r=1$ 到某个 $T\geq 1$的迭代过程中的下降性质。

定义

$$
\tilde{F}(\bar{\boldsymbol{x}}^r,\bar{\boldsymbol{\theta}}^r)\triangleq\mathbb{E}\bigg[\frac{1}{NS}\sum_{i=1}^{S}f\Big(\sum_{n=1}^{N}\boldsymbol{B}_{n,i}\bar{\boldsymbol{x}}^r,\bar{\boldsymbol{\theta}}^r\Big)\bigg].
\tag{95}
$$

引理 11 （下降引理）令假设 *1* 至假设 *4* 以及假设 *6* 成立。对于 $\forall r\geq 0$，有

$$
\begin{aligned}
&\tilde{F}(\bar{\boldsymbol{x}}^{r+1},\bar{\boldsymbol{\theta}}^{r+1})-\tilde{F}(\bar{\boldsymbol{x}}^r,\bar{\boldsymbol{\theta}}^r) \\
&\leq-\frac{\alpha E}{2}\left\{1-10EL\alpha-\left[\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right]64E(E-1)N\alpha\right\}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]-\frac{\beta E}{2}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{x}}^r\|^2] \\
&\quad-\frac{\beta}{2}\left\{1-10E^2L\beta-\left[\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right]96E(E-1)N^2B_{\max}^2\beta\right\}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2] \\
&\quad+\left\{\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right\}[1+7(E-1)\lambda_w^2]\phi_\theta^r \\
&\quad+\left\{\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right\}[E+200(E-1)\lambda_w^2]N^2B_{\max}^2\phi_x^r \\
&\quad+\left\{\frac{2\alpha L^2}{SN^3}+\frac{2\beta B_{\max}^2L^2}{SN}\right\}[1+32(E-1)\lambda_w^2]\phi_z^r \\
&\quad+\left\{\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right\}128(E-1)[1+E(E-1)]\lambda_w^2N^2B_{\max}^2\beta^2\phi_u^r \\
&\quad+\frac{3EL\alpha^2+9(E-1)L\beta^2}{N}\frac{\sigma^2}{|\mathcal{I}|},
\end{aligned}
\tag{96}
$$

I.2 线性时不变系统

基于引理 9 和引理 10，我们现在将迭代收缩总结为一个线性时不变系统，该系统刻画了混合数据设置下共识过程和梯度跟踪过程的收敛性。

ˆ ˆ 此处，我们将向量 $\hat{}\boldsymbol{e}^r$ 的系统矩阵 $\boldsymbol{A}$、系数矩阵 $\boldsymbol{C}$ 定义为

$$
\hat{\boldsymbol{A}}\triangleq\begin{bmatrix}a_1 & a_2\alpha^2 & a_3\alpha^2 & a_4\alpha^2\beta^2\\ b_1\beta^2 & b_2 & b_3\beta^2 & b_4\beta^2\\ c_1\beta^2 & c_2\beta^2 & c_3 & c_4\beta^2\\ d_1 & d_{21}\alpha^2+d_{22}\beta^2 & d_3 & d_4\end{bmatrix},
\tag{97a}
$$

$$\hat{\boldsymbol{C}} \triangleq \begin{bmatrix} a_5\alpha^2 & a_6\alpha^2 & a_7\alpha^2\beta^2 \\ b_5\beta^2 & b_6\alpha^2\beta^2 & b_7\beta^4 \\ c_5\beta^2 & c_6\alpha^2\beta^2 & c_7\beta^2 \\ d_5 & d_6\alpha^2 & d_7\beta^2 \end{bmatrix}, \tag{97b}$$

$$\hat{\boldsymbol{e}}^r \triangleq \begin{bmatrix} \frac{\sigma^2}{|\mathcal{I}|} \\ \mathbb{E}[\|\nabla_{\boldsymbol{\theta}}^r\|^2] \\ \mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2] \end{bmatrix}, \tag{97c}$$

其中，参数 a_1-a_7、 b_1-b_7、 c_1-c_7、 d_1-d_7 可分别根据引理 9 和引理10 确定。

因此，可以将线性时不变系统简洁地写为

$$\hat{\boldsymbol{\phi}}^{r+1} \leq \hat{\boldsymbol{A}}\boldsymbol{\phi}^r + \hat{\boldsymbol{C}}\hat{\boldsymbol{e}}^r, \tag{98}$$

其中，不等式是逐元素成立的。

容　易确定，存在一个正向量 $\hat{\boldsymbol{s}} = [\hat{s}_1, \ldots\ldots, \hat{s}_4]^\top \in \mathbb{R}^4$ 满足

$$\hat{s}_1 \leq \frac{1-d_4}{5d_1}\hat{s}_4,\ \hat{s}_3 \leq \frac{1-d_4}{5d_3}\hat{s}_4, \tag{99}$$

使得 $\boldsymbol{A}\hat{\boldsymbol{s}} < \hat{\boldsymbol{s}}$ ，只要

$$\alpha \leq \min\left\{\frac{\sqrt{(1-d_4)\hat{s}_4}}{\sqrt{5d_{21}\hat{s}_2}}, \frac{\sqrt{(1-a_1)\hat{s}_1}}{\sqrt{a_2\hat{s}_2 + a_3\hat{s}_3 + a_4\hat{s}_4 + 1}}\right\} \tag{100}$$

$$\beta \leq \min\left\{\frac{\sqrt{(1-d_4)\hat{s}_4}}{\sqrt{5d_{22}\hat{s}_2}}, \frac{\sqrt{(1-b_2)\hat{s}_2}}{\sqrt{b_1\hat{s}_1 + b_3\hat{s}_3 + b_4\hat{s}_4 + 1}}, \frac{\sqrt{(1-c_3)\hat{s}_3}}{\sqrt{c_1\hat{s}_1 + c_2\hat{s}_2 + c_4\hat{s}_4 + 1}}\right\}. \tag{101}$$

因此，根据 [28, 推论 8.1.29]，在条件 (100) 下， $\boldsymbol{A}$ 的谱半径满足 $\rho(\boldsymbol{A}) < 1$ 。接下来，我们在以下引理中求系统矩阵 $\boldsymbol{A}$ 的逆。

引理 12 对于足够小的 α 和满足条件的 β

$$\alpha \leq \min\left\{\frac{1}{\sqrt{d_{21}}}, \frac{1}{\sqrt{d_{22}}}, \frac{\sqrt{c_2d_3}}{\sqrt{a_3c_2d_1}}, \frac{\sqrt{c_2d_3}}{\sqrt{a_2c_1d_3}}, \frac{b_4}{4\sqrt{a_4}}, \frac{b_4}{4\sqrt{a_3}}\right\}, \tag{102}$$

$$\beta \leq \min\left\{\frac{\sqrt{c_1}}{2\sqrt{b_1c_2}}, \frac{\sqrt{c_1}}{2\sqrt{b_1c_4}}, \frac{\sqrt{a_3}}{2\sqrt{a_2b_3}}, \frac{\sqrt{a_3}}{2\sqrt{a_4d_3}}, \frac{\sqrt{a_3}}{2\sqrt{a_4b_3}}, \frac{\sqrt{a_3}}{2\sqrt{a_2b_4d_3}}, \frac{1}{\sqrt{d_{22}}}, \frac{\sqrt{b_4}}{2\sqrt{b_3c_4}}, \frac{\sqrt{b_4}}{2\sqrt{b_3c_1}}, \frac{\sqrt{b_4}}{2\sqrt{b_1c_4}}\right\}, \tag{103}$$

行列式$(\boldsymbol{I} - \hat{\boldsymbol{A}})^{-1}$ 具有逐元素上界

$$(\mathbf{I} - \boldsymbol{A})^{-1} \leq \frac{1}{(1-\lambda_w^2)^4}\hat{\boldsymbol{A}}^*, \tag{104}$$

其中

$$\hat{\boldsymbol{A}}^* \triangleq \begin{bmatrix} 1 & a_2c_4d_3\alpha^2\beta^2 & 2a_3\alpha^2 & a_4b_3c_2\alpha^2\beta^6 \\ b_1c_4d_3\beta^4 & 1 & a_4b_3d_1\alpha^2\beta^4 & 2b_4\beta^2 \\ (c_4d_1+2c_1)\beta^2 & a_4c_2d_1\alpha^2\beta^4 & 1 & a_2b_1c_4\alpha^2\beta^4 \\ b_3c_2d_1\beta^4 & (2c_2d_3+d_{22})\beta^2+d_{21}\alpha^2 & a_2b_1d_3\alpha^2\beta^2 & 1 \end{bmatrix}. \tag{105}$$

与式 (25) 类似，我们有

$$\sum_{r=0}^{T-1}\hat{\boldsymbol{\phi}}^r \leq (\mathbf{I}-\hat{\boldsymbol{A}})^{-1}\hat{\boldsymbol{\phi}}^0 + (\mathbf{I}-\hat{\boldsymbol{A}})^{-1}\sum_{r=0}^{T-1}\hat{\boldsymbol{C}}\hat{\boldsymbol{e}}^r. \tag{106}$$

令 $\hat{}\,\phi_\theta^0 = \hat{}\,\phi_x^0 = 0$。然后，将 (104) 代入 (106)，可得以下引理 .

引理 13 对于足够小的 α 和满足条件的 β

$$\begin{aligned}\alpha \leq \min\Bigg\{ & \frac{\sqrt{a_6}}{2\sqrt{a_2d_3}}, \frac{\sqrt{a_6}}{4\sqrt{a_3}}, \frac{\sqrt{a_6}}{2\sqrt{a_4d_6}}, \frac{\sqrt{a_7}}{2\sqrt{a_2d_3}}, \frac{\sqrt{a_7}}{4\sqrt{a_4}}, \frac{\sqrt{b_5}}{2\sqrt{a_5d_3}}, \frac{\sqrt{b_5}}{2\sqrt{a_4d_1}}, \frac{\sqrt{b_5}}{2\sqrt{a_6d_3}}, \frac{\sqrt{b_5}}{2\sqrt{a_7d_3}} \\ & \frac{\sqrt{c_5}}{3\sqrt{a_5(c_4d_1+2c_1)}}, \frac{\sqrt{c_5}}{3\sqrt{a_4d_1}}, \frac{\sqrt{c_5}}{3\sqrt{a_2d_5}}, \frac{\sqrt{c_6}}{2\sqrt{a_4d_1}}, \frac{\sqrt{c_6}}{2\sqrt{a_1d_6}}, \frac{\sqrt{c_7}}{3\sqrt{a_7}}, \frac{\sqrt{c_7}}{3\sqrt{a_4d_1}}, \frac{\sqrt{c_7}}{3\sqrt{a_2d_7}} \\ & \frac{\sqrt{d_5}}{2\sqrt{a_5d_1}}, \frac{\sqrt{d_5}}{2\sqrt{d_{21}}}, \frac{\sqrt{d_5}}{2\sqrt{a_2d_3}}, \frac{\sqrt{d_6}}{2\sqrt{d_{21}}}, \frac{\sqrt{d_6}}{2\sqrt{a_2d_3}}, \frac{\sqrt{d_7}}{2\sqrt{a_7d_1}}, \left(\frac{d_7}{b_7(2c_2d_3+d_{22})}\right)^{\frac{1}{4}}, \frac{\sqrt{d_7}}{2\sqrt{d_{21}}}, \\ & \frac{\sqrt{d_7}}{2\sqrt{a_2d_3}}\Bigg\},\end{aligned} \tag{107}$$

$$\begin{aligned}\beta \leq \min\Bigg\{ & \left(\frac{a_5}{a_2b_5c_4d_3}\right)^{\frac{1}{4}}, \frac{\sqrt{a_5}}{4\sqrt{a_3c_5}}, \left(\frac{a_5}{a_4b_3c_2}\right)^{\frac{1}{6}}, \frac{\sqrt{a_6}}{2\sqrt{b_6c_4}}, \frac{\sqrt{a_6}}{4\sqrt{c_6}}, \frac{\sqrt{a_6}}{2\sqrt{b_3c_2}}, \frac{\sqrt{a_7}}{2\sqrt{b_7c_4}}, \frac{\sqrt{a_7}}{2\sqrt{b_3c_2}}, \\ & \frac{\sqrt{b_5}}{2\sqrt{b_1c_4}}, \frac{\sqrt{b_5}}{2\sqrt{b_3c_5}}, \frac{\sqrt{b_5}}{2\sqrt{b_3c_6}}, \frac{\sqrt{b_5}}{2\sqrt{b_3c_7}}\frac{\sqrt{c_5}}{2\sqrt{b_5c_2}}, \frac{\sqrt{c_5}}{3\sqrt{b_1c_4}}, \frac{\sqrt{c_6}}{2\sqrt{b_6c_2}}, \frac{\sqrt{c_6}}{2\sqrt{b_1c_4}}, \frac{\sqrt{c_7}}{2\sqrt{b_1c_4}}, \\ & \frac{\sqrt{c_7}}{2\sqrt{c_4d_1+2c_1}}, \frac{\sqrt{c_7}}{3\sqrt{b_7c_2}}\frac{\sqrt{d_5}}{2\sqrt{b_3c_2}}, \frac{\sqrt{d_5}}{2\sqrt{b_5}}, \left(\frac{d_5}{2b_5(2c_2d_3+d_{22})}\right)^{\frac{1}{4}}, \frac{\sqrt{d_5}}{2\sqrt{b_1c_5}}, \left(\frac{d_6}{2a_6b_3c_2d_1}\right)^{\frac{1}{4}}, \\ & \left(\frac{d_6}{2b_6(2c_2d_3+d_{22})}\right)^{\frac{1}{4}}, \frac{\sqrt{d_6}}{2\sqrt{b_6}}, \frac{\sqrt{d_6}}{2\sqrt{b_1c_6}}, \frac{\sqrt{d_7}}{2\sqrt{b_3c_2}}, \frac{\sqrt{d_7}}{2\sqrt{b_7}}, \frac{\sqrt{d_7}}{2\sqrt{b_1c_7}}\Bigg\},\end{aligned} \tag{108}$$

我们有

$$\begin{aligned}\sum_{r=0}^{T-1}\phi_\theta^r \leq & \frac{1}{(1-\lambda_w^2)^4}\Bigg\{2a_5\alpha^2T\frac{\sigma^2}{|\mathcal{I}|} + 2a_6\alpha^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2] + 2(a_7+a_3c_7)\alpha^2\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2] \\ & + 2a_3\alpha^2\phi_z^0 + a_4b_3c_2\alpha^2\beta^6\phi_u^0\Bigg\},\end{aligned} \tag{109a}$$

$$\begin{aligned}\sum_{r=0}^{T-1}\phi_x^r \leq & \frac{1}{(1-\lambda_w^2)^4}\Bigg\{2(b_5+b_4d_5)\beta^2T\frac{\sigma^2}{|\mathcal{I}|} + 2(b_6+b_4d_6)\alpha^2\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2] \\ & + 2(b_7+b_4d_7)\beta^4\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2] + a_4b_3d_1\alpha^2\beta^4\phi_z^0 + 2b_4\beta^2\phi_u^0\Bigg\},\end{aligned} \tag{109b}$$

$$\begin{aligned}\sum_{r=0}^{T-1}\phi_z^r \leq & \frac{1}{(1-\lambda_w^2)^4}\Bigg\{2c_5\beta^2T\frac{\sigma^2}{|\mathcal{I}|} + 2[a_6(c_4d_1+2c_1)+c_6]\alpha^2\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2] \\ & + 2c_7\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2] + \phi_z^0 + a_2b_1c_4\alpha^2\beta^4\phi_u^0\Bigg\},\end{aligned} \tag{109c}$$

and

$$\sum_{r=0}^{T-1}\phi_u^r \le \frac{1}{(1-\lambda_w^2)^4}\Bigg\{2d_5T\frac{\sigma^2}{|\mathcal{I}|}+2d_6\alpha^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]+2d_7\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2] + a_2b_1d_3\alpha^2\beta^2\phi_z^0+\phi_u^0\Bigg\}. \tag{109d}$$

I.3 定理 2 的证明

将 (96) 从 $r=0$ 到 $r=T-1$ 求和，然后将 (109) 代入所得结果，可得

$$\begin{aligned}
&\tilde{F}(\bar{\boldsymbol{x}}^T,\bar{\boldsymbol{\theta}}^T)-\tilde{F}(\bar{\boldsymbol{x}}^0,\bar{\boldsymbol{\theta}}^0)\\
&\le -\frac{\alpha E}{2}\Bigg\{1-10EL\alpha-\left[\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right]64E(E-1)N\alpha\Bigg\}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]\\
&\quad-\frac{\beta E}{2}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{x}}^r\|^2]\\
&\quad-\frac{\beta}{2}\Bigg\{1-10E^2L\beta-\left[\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right]96E(E-1)N^2B_{\max}^2\beta\Bigg\}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]\\
&\quad+\left\{\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right\}[1+7(E-1)\lambda_w^2]\Bigg\{\frac{1}{(1-\lambda_w^2)^4}\{2a_5\alpha^2T\frac{\sigma^2}{|\mathcal{I}|}+2a_6\alpha^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]\\
&\qquad+2(a_7+a_3c_7)\alpha^2\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]+2a_3\alpha^2\phi_z^0+a_4b_3c_2\alpha^2\beta^6\phi_u^0\}\Bigg\}\\
&\quad+\left\{\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right\}[E+200(E-1)\lambda_w^2]N^2B_{\max}^2\Bigg\{\frac{1}{(1-\lambda_w^2)^4}\{2(b_5+b_4d_5)\beta^2T\frac{\sigma^2}{|\mathcal{I}|}\\
&\qquad+2(b_6+b_4d_6)\alpha^2\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]+2(b_7+b_4d_7)\beta^4\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]\\
&\qquad+a_4b_3d_1\alpha^2\beta^2\phi_z^0+2b_4\beta^2\phi_u^0\}\Bigg\}\\
&\quad+\left\{\frac{2\alpha L^2}{SN^3}+\frac{2\beta B_{\max}^2L^2}{SN}\right\}[1+32(E-1)\lambda_w^2]\Bigg\{\frac{1}{(1-\lambda_w^2)^4}\{2c_5\beta^2T\frac{\sigma^2}{|\mathcal{I}|}+2[a_6(c_4d_1+2c_1)+c_6]\\
&\qquad\alpha^2\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]+2c_7\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]+\phi_z^0+a_2b_1c_4\alpha^2\beta^4\phi_u^0\}\Bigg\}\\
&\quad+\left\{\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right\}128(E-1)[1+E(E-1)]\lambda_w^2N^2B_{\max}^2\beta^2\Bigg\{\frac{1}{(1-\lambda_w^2)^4}\{2d_5T\frac{\sigma^2}{|\mathcal{I}|}\\
&\qquad+2d_6\alpha^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]+2d_7\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]+a_2b_1d_3\alpha^2\beta^2\phi_z^0+\phi_u^0\}\Bigg\}\\
&\quad+[3EL\alpha^2+9(E-1)L\beta^2]T\frac{\sigma^2}{|\mathcal{I}|}
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{\alpha E}{2}\Bigg\{1-10EL\alpha-\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)64E(E-1)N\alpha \\
&\quad-\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)[1+7(E-1)\lambda_w^2]\frac{4a_6\alpha}{E(1-\lambda_w^2)^4} \\
&\quad-\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)[E+200(E-1)\lambda_w^2]N^2B_{\max}^2\frac{1}{(1-\lambda_w^2)^4}4(b_6+b_4d_6)\alpha\beta^2 \\
&\quad-\left(\frac{2\alpha L^2}{SN^3}+\frac{2\beta B_{\max}^2 L^2}{SN}\right)[1+32(E-1)\lambda_w^2]\frac{1}{(1-\lambda_w^2)^4}4[a_6(c_4d_1+2c_1)+c_6]\alpha\beta^2 \\
&\quad-\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)128(E-1)[1+E(E-1)]\lambda_w^2\frac{4d_6N^2B_{\max}^2\alpha\beta^2}{(1-\lambda_w^2)^4}\Bigg\}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2] \\
&\quad-\frac{\beta}{2}\Bigg\{1-10E^2L\beta-\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)96E(E-1)N^2B_{\max}^2\beta \\
&\quad-\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)[1+7(E-1)\lambda_w^2]\frac{4(a_7+a_3c_7)\alpha^2\beta}{(1-\lambda_w^2)^4} \\
&\quad-\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)[E+200(E-1)\lambda_w^2]N^2B_{\max}^2\frac{4(b_7+b_4d_7)\beta^3}{(1-\lambda_w^2)^4} \\
&\quad-\left(\frac{2\alpha L^2}{SN^3}+\frac{2\beta B_{\max}^2 L^2}{SN}\right)[1+32(E-1)\lambda_w^2]\frac{4c_7\beta}{(1-\lambda_w^2)^4} \\
&\quad-\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)128(E-1)[1+E(E-1)]\lambda_w^2N^2B_{\max}^2\beta^2\frac{4d_7\beta^2}{(1-\lambda_w^2)^4}\Bigg\}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2] \\
&\quad-\frac{\beta E}{2}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{x}}^r\|^2] \\
&\quad+\Bigg\{[3EL\alpha^2+9(E-1)L\beta^2]+\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)[1+7(E-1)\lambda_w^2]\frac{2a_5\alpha^2}{(1-\lambda_w^2)^4} \\
&\quad+\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)[E+200(E-1)\lambda_w^2]\frac{2(b_5+b_4d_5)N^2B_{\max}^2\beta^2}{(1-\lambda_w^2)^4} \\
&\quad+\left(\frac{2\alpha L^2}{SN^3}+\frac{2\beta B_{\max}^2 L^2}{SN}\right)[1+32(E-1)\lambda_w^2]\frac{2c_5\beta^2}{(1-\lambda_w^2)^4} \\
&\quad+\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)128(E-1)[1+E(E-1)]\lambda_w^2\frac{2d_5N^2B_{\max}^2\beta^2}{(1-\lambda_w^2)^4}\Bigg\}T\frac{\sigma^2}{|\mathcal{I}|} \\
&\quad+\Bigg\{\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)[1+7(E-1)\lambda_w^2]\frac{2a_3\alpha^2}{(1-\lambda_w^2)^4} \\
&\quad+\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)[E+200(E-1)\lambda_w^2]\frac{a_4b_3d_1N^2B_{\max}^2\alpha^2\beta^2}{(1-\lambda_w^2)^4} \\
&\quad+\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)128(E-1)[1+E(E-1)]\lambda_w^2\frac{a_2b_1d_3N^2B_{\max}^2\alpha^2\beta^4}{(1-\lambda_w^2)^4} \\
&\quad+\left(\frac{2\alpha L^2}{SN^3}+\frac{2\beta B_{\max}^2 L^2}{SN}\right)\frac{[1+32(E-1)\lambda_w^2]}{(1-\lambda_w^2)^4}\Bigg\}\phi_z^0 \\
&\quad+\Bigg\{\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)[1+7(E-1)\lambda_w^2]\frac{a_4b_3c_2\alpha^2\beta^6}{(1-\lambda_w^2)^4}
\end{aligned}
$$

$$
\begin{aligned}
&+\left(\frac{2\alpha L^2}{SN^3}+\frac{2\beta B_{\max}^2 L^2}{SN}\right)[1+32(E-1)\lambda_w^2]\frac{a_2 b_1 c_4 \alpha^2\beta^4}{(1-\lambda_w^2)^4}\\
&+\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)128(E-1)[1+E(E-1)]\lambda_w^2\frac{N^2B_{\max}^2\beta^2}{(1-\lambda_w^2)^4}\\
&+\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)[E+200(E-1)\lambda_w^2]\frac{2b_4N^2B_{\max}^2\beta^2}{(1-\lambda_w^2)^4}\Bigg\}\phi_u^0 \qquad (110)\\
\le&-\frac{\alpha E}{4}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar\nabla_{\boldsymbol\theta}^r\|^2]-\frac{\beta E}{4}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar\nabla_{\boldsymbol x}^r\|^2]-\frac{\beta}{4}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol u}^r\|^2]\\
&+[4EL\alpha^2+10(E-1)L\beta^2]T\frac{\sigma^2}{|\mathcal I|}\\
&+\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)\frac{1}{(1-\lambda_w^2)^4}\Bigg\{[1+7(E-1)\lambda_w^2]2a_5\alpha^2+\frac{1}{S}[1+32(E-1)\lambda_w^2]2c_5\beta^2\\
&+[E+200(E-1)\lambda_w^2]4b_4d_5N^2B_{\max}^2\beta^2\\
&+128(E-1)[1+E(E-1)]\lambda_w^2 2d_5N^2B_{\max}^2\beta^2\Bigg\}T\frac{\sigma^2}{|\mathcal I|}\\
&+\left(\frac{2\alpha L^2}{SN^3}+\frac{2\beta B_{\max}^2 L^2}{SN}\right)[1+32(E-1)\lambda_w^2]\frac{2}{(1-\lambda_w^2)^3}\phi_z^0\\
&+\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2 L^2}{N}\right)\frac{N^2B_{\max}^2\beta^2}{(1-\lambda_w^2)^4}\Bigg\{2b_4[E+200(E-1)\lambda_w^2]+1\\
&+128(E-1)[1+E(E-1)]\lambda_w^2\Bigg\}\phi_u^0, \qquad (111)
\end{aligned}
$$

其中，若 α 和 β 满足

$$
\begin{aligned}
\alpha\le\min\Bigg\{&\frac{N^3}{4L^2},\frac{1}{120EL},\frac{1}{768E^2N},\frac{E(1-\lambda_w^2)^4}{48a_6[1+7(E-1)\lambda_w^2]},\frac{(1-\lambda_w^2)^4}{12N^2B_{\max}^2[E+200(E-1)\lambda_w^2]},\\
&\frac{(1-\lambda_w^2)^4}{48[1+32(E-1)\lambda_w^2]},\frac{(1-\lambda_w^2)^4}{6144d_6N^2B_{\max}^2(E-1)[1+E(E-1)\lambda_w^2]},\frac{(1-\lambda_w^2)^2}{48\sqrt{a_7+a_3c_7}[1+7(E-1)\lambda_w^2]},\\
&\frac{\sqrt{[1+32(E-1)\lambda_w^2]}}{8a_3[1+7(E-1)\lambda_w^2]},\sqrt{\frac{1}{a_4d_1N^2B_{\max}^2}},\sqrt{\frac{1}{a_2d_3N^2B_{\max}^2}},\frac{1}{\sqrt{a_2c_4}}\Bigg\}, \qquad (112)\\
\beta\le\min\Bigg\{&\frac{N}{4B_{\max}^2L^2},\frac{1}{2\sqrt{b_6+b_4d_6}},\frac{1}{\sqrt{a_6(c_4d_1+2c_1)+c_6}},\frac{1}{120E^2L},\frac{1}{1152E(E-1)N^2B_{\max}^2},\\
&\frac{(1-\lambda_w^2)}{\{48(b_7+b_4d_7)N^2B_{\max}^2[E+200(E-1)\lambda_w^2]\}^{\frac13}},\frac{(1-\lambda_w^2)^4}{48Sc_7[1+32(E-1)\lambda_w^2]},\\
&\frac{(1-\lambda_w^2)^2}{78NB_{\max}\sqrt{d_7(E-1)[1+E(E-1)\lambda_w^2]}},\sqrt{\frac{[1+32(E-1)\lambda_w^2]}{4b_3[E+200(E-1)\lambda_w^2]}},\\
&\left(\frac{1+32(E-1)\lambda_w^2}{512b_1(E-1)[1+E(E-1)]\lambda_w^2}\right)^{\frac14},\left(\frac{1}{2a_4b_3c_2[1+7E(E-1)\lambda_w^2]}\right)^{\frac14},\\
&\sqrt{\frac{1}{2b_1S[1+32(E-1)\lambda_w^2]}}\Bigg\}, \qquad (113)
\end{aligned}
$$

从而在 (110) 中， $\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]$ 前括号内最后 6 个负项小于$1\,12$， $\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]$ 前括号内最后 6 个负项小于$1\,12$，前 3 项小于 ϕ_z^0前的最后一项，前 2 项小于 ϕ_u^0前的$1\,2$。

通过重新整理上述不等式，并进一步代入假设3中的 $\tilde{F}(\bar{\boldsymbol{x}}^T,\ \bar{\boldsymbol{\theta}}^T) \geq \underline{F}$，我们得到

$$
\begin{aligned}
&\frac{\alpha E}{4}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]+\frac{\beta E}{4}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{x}}^r\|^2]\\
&\leq \tilde{F}(\bar{\boldsymbol{\theta}}^0,\bar{\boldsymbol{x}}^0)-\underline{F}-\frac{\beta}{4}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]+[4ELa\alpha^2+10(E-1)L\beta^2]T\frac{\sigma^2}{|\mathcal{I}|}\\
&\quad+\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right)\frac{1}{(1-\lambda_w^2)^4}\Big\{[1+7(E-1)\lambda_w^2]2a_5\alpha^2+\frac{1}{S}[1+32(E-1)\lambda_w^2]2c_5\beta^2\\
&\qquad+[E+200(E-1)\lambda_w^2]4b_4d_5N^2B_{\max}^2\beta^2\\
&\qquad+128(E-1)[1+E(E-1)]\lambda_w^2 2d_5N^2B_{\max}^2\beta^2\Big\}T\frac{\sigma^2}{|\mathcal{I}|}\\
&\quad+\left(\frac{2\alpha L^2}{SN^3}+\frac{2\beta B_{\max}^2L^2}{SN}\right)[1+32(E-1)\lambda_w^2]\frac{2}{(1-\lambda_w^2)^3}\phi_z^0\\
&\quad+\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right)\frac{N^2B_{\max}^2\beta^2}{(1-\lambda_w^2)^4}\Big\{2b_4[E+200(E-1)\lambda_w^2]+1\\
&\qquad+128(E-1)[1+E(E-1)]\lambda_w^2\Big\}\phi_u^0.
\end{aligned}
\tag{114}
$$

此外，我们对 $\boldsymbol{\theta}$ 和 $\boldsymbol{x}$ 的平均梯度进行透视分析如下。

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left\|\frac{1}{NS}\sum_{i=1}^{S}\nabla_{\boldsymbol{\theta}}f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\right)\right\|^2\right]\\
&\leq 2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left\|\frac{1}{NS}\sum_{i=1}^{S}\left[\nabla_{\boldsymbol{\theta}}f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\right)-\nabla_{\boldsymbol{\theta}}f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\bar{\boldsymbol{x}}^r,\bar{\boldsymbol{\theta}}^r\right)\right\|^2\right]+2\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]\\
&\leq 2\frac{1}{N}\sum_{n=1}^{N}\frac{1}{N^2S}\sum_{i=1}^{S}L^2\mathbb{E}\left[\left\|\sum_{t=1}^{N}\boldsymbol{B}_{t,i}(\boldsymbol{x}_n^r-\bar{\boldsymbol{x}}^r)\right\|^2\right]+2\frac{1}{N}\sum_{n=1}^{N}\frac{1}{N^2S}\sum_{i=1}^{S}L^2\mathbb{E}[\|\boldsymbol{\theta}_n^r-\bar{\boldsymbol{\theta}}^r\|^2]\\
&\quad+2\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]\\
&\leq \frac{2L^2B_{\max}^2}{N}\phi_x^r+\frac{2L^2}{N^3}\phi_\theta^r+2\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2].
\end{aligned}
\tag{115}
$$

同样，我们有

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left\|\frac{1}{NS}\sum_{i=1}^{S}\sum_{t=1}^{N}\boldsymbol{B}_{t,i}^\top\nabla_{\boldsymbol{B}_i\boldsymbol{x}}f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\right)\right\|^2\right]\\
&\leq 2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left\|\frac{1}{NS}\sum_{i=1}^{S}\sum_{t=1}^{N}\boldsymbol{B}_{t,i}^\top\left[\nabla_{\boldsymbol{B}_i\boldsymbol{x}}f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\right)-\nabla_{\boldsymbol{B}_i\boldsymbol{x}}f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\bar{\boldsymbol{x}}^r,\bar{\boldsymbol{\theta}}^r\right)\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{x}}^r\|^2] \\
&\le 2\frac{1}{N}\sum_{n=1}^{N}\frac{1}{N^2S}\sum_{i=1}^{S}N\sum_{t=1}^{N}B_{\max}^2L^2\mathbb{E}\bigg[\bigg\|\sum_{t=1}^{N}\boldsymbol{B}_{t,i}(\boldsymbol{x}_n^r-\bar{\boldsymbol{x}}^r)\bigg\|^2\bigg] \\
&\quad+\frac{1}{N}\sum_{n=1}^{N}\frac{1}{N^2S}\sum_{i=1}^{S}N\sum_{t=1}^{N}B_{\max}^2L^2\mathbb{E}[\|\boldsymbol{\theta}_n^r-\bar{\boldsymbol{\theta}}^r\|^2]+2\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{x}}^r\|^2] \\
&\le 2NB_{\max}^4L^2\phi_x^r+\frac{2B_{\max}^2L^2}{N}\phi_\theta^r+2\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{x}}^r\|^2] \qquad (116)
\end{aligned}
$$

分别将 (115) 和 (116) 两边乘以 $\frac{\alpha E}{10}$ 和 $\frac{\beta E}{10}$，然后将结果从 $r=0$ 到 $r=T-1$ 求和，可得

$$
\begin{aligned}
&\sum_{r=0}^{T-1}\frac{\alpha E}{10}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\bigg[\bigg\|\frac{1}{NS}\sum_{i=1}^{S}\nabla_{\boldsymbol{\theta}}f\Big(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\Big)\bigg\|^2\bigg] \\
&\quad+\sum_{r=0}^{T-1}\frac{\beta E}{10}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\bigg[\bigg\|\frac{1}{NS}\sum_{i=1}^{S}\sum_{t=1}^{N}\boldsymbol{B}_{t,i}^\top\nabla_{\boldsymbol{B}_i\boldsymbol{x}}f\Big(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\Big)\bigg\|^2\bigg] \\
&\le\frac{L^2B_{\max}^2}{5N}(\alpha+N^2B_{\max}^2\beta)\sum_{r=0}^{T-1}\phi_x^r+\frac{L^2}{5N^3}(\alpha+N^2B_{\max}^2\beta)\sum_{r=0}^{T-1}\phi_\theta^r+\frac{\alpha E}{5}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2] \\
&\quad+\frac{\beta E}{5}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{x}}^r\|^2] \qquad (117)
\end{aligned}
$$

将 (109a) 和 (109b) 代入 (117)，可得

$$
\begin{aligned}
&\sum_{r=0}^{T-1}\frac{\alpha E}{10}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\bigg[\bigg\|\frac{1}{NS}\sum_{i=1}^{S}\nabla_{\boldsymbol{\theta}}f\Big(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\Big)\bigg\|^2\bigg] \\
&\quad+\sum_{r=0}^{T-1}\frac{\beta E}{10}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\bigg[\bigg\|\frac{1}{NS}\sum_{i=1}^{S}\sum_{t=1}^{N}\boldsymbol{B}_{t,i}^\top\nabla_{\boldsymbol{B}_i\boldsymbol{x}}f\Big(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\Big)\bigg\|^2\bigg] \\
&\le\frac{L^2B_{\max}^2}{5N}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}\bigg\{2(b_5+b_4d_5)\beta^2T\frac{\sigma^2}{|\mathcal{I}|}+2(b_6+b_4d_6)\alpha^2\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2] \\
&\qquad+2(b_7+b_4d_7)\beta^4\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]+a_4b_3d_1\alpha^2\beta^2\phi_z^0+2b_4\beta^2\phi_u^0\bigg\} \\
&\quad+\frac{L^2}{5N^3}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}\bigg\{2a_5\alpha^2T\frac{\sigma^2}{|\mathcal{I}|}+2a_6\alpha^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2] \\
&\qquad+2(a_7+a_3c_7)\alpha^2\beta^2\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]+2a_3\alpha^2\phi_z^0+a_4b_3c_2\alpha^2\beta^6\phi_u^0\bigg\} \\
&\quad+\frac{\alpha E}{5}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]+\frac{\beta E}{5}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{x}}^r\|^2] \\
&\le\frac{\alpha E}{4}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]+\frac{\beta E}{4}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{x}}^r\|^2]+\bigg\{\frac{L^2B_{\max}^2}{5N}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2(b_7+b_4d_7)\beta^4
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{L^2}{5N^3}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2(a_7+a_3c_7)\alpha^2\beta^2\Big\}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]\\
&+\Big\{\frac{L^2B_{\max}^2}{5N}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2(b_5+b_4d_5)\beta^2\\
&+\frac{L^2}{5N^3}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2a_5\alpha^2\Big\}T\frac{\sigma^2}{|\mathcal{I}|}\\
&+\Big\{\frac{L^2B_{\max}^2}{5N}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}a_4b_3d_1\alpha^2\beta^2\\
&+\frac{L^2}{5N^3}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2a_3\alpha^2\Big\}\phi_z^0\\
&+\Big\{\frac{L^2B_{\max}^2}{5N}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2b_4\beta^2\\
&+\frac{L^2}{5N^3}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}a_4b_3c_2\alpha^2\beta^6\Big\}\phi_u^0 \qquad (118)\\
\le\ &\frac{\alpha E}{4}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{\theta}}^r\|^2]+\frac{\beta E}{4}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\nabla}_{\boldsymbol{x}}^r\|^2]+\Big\{\frac{L^2B_{\max}^2}{5N}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2(b_7+b_4d_7)\beta^4\\
&+\frac{L^2}{5N^3}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2(a_7+a_3c_7)\alpha^2\beta^2\Big\}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]\\
&+\Big\{\frac{L^2B_{\max}^2}{5N}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2(b_5+b_4d_5)\beta^2\\
&+\frac{L^2}{5N^3}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2a_5\alpha^2\Big\}T\frac{\sigma^2}{|\mathcal{I}|}\\
&+\frac{L^2}{5N^3}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2a_3\alpha^2\phi_z^0+\frac{L^2B_{\max}^2}{5N}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2b_4\beta^2\phi_u^0, \qquad (119)
\end{aligned}
$$

其中，最后一个不等式是通过令

$$
\alpha\le\frac{\sqrt{b_4}}{\sqrt{a_4}},\ \beta\le\min\Big\{\frac{\sqrt{a_3}}{\sqrt{a_4b_3d_1B_{\max}^2N^2}},\frac{\sqrt{b_4}}{\sqrt{b_3c_2N^2}}\Big\}, \qquad (120)
$$

因此，在 (118) 中，第一个项小于 ϕ_z^0 之前的第二个项，并且第一个项小于 ϕ_u^0 之前的第二个项。

将 (114) 代入上述不等式，可得

$$
\begin{aligned}
&\sum_{r=0}^{T-1}\frac{\alpha E}{10}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\frac{1}{NS}\sum_{i=1}^{S}\nabla_{\boldsymbol{\theta}}f\Big(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\Big)\Big\|^2\Big]\\
&\quad+\sum_{r=0}^{T-1}\frac{\beta E}{10}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\frac{1}{NS}\sum_{i=1}^{S}\sum_{t=1}^{N}\boldsymbol{B}_{t,i}^{\top}\nabla_{\boldsymbol{B}_i\boldsymbol{x}}f\Big(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r,\boldsymbol{\theta}_n^r\Big)\Big\|^2\Big]\\
&\le\tilde{F}(\bar{\boldsymbol{\theta}}^0,\bar{\boldsymbol{x}}^0)-\underline{F}+\Big\{\frac{L^2B_{\max}^2}{5N}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2(b_7+b_4d_7)\beta^4
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{L^2}{5N^3}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2(a_7+a_3c_7)\alpha^2\beta^2-\frac{\beta}{4}\Bigg\}\sum_{r=0}^{T-1}\mathbb{E}[\|\bar{\boldsymbol{u}}^r\|^2]\\
&+[4ELa^2+10(E-1)L\beta^2]T\frac{\sigma^2}{|\mathcal{I}|}\\
&+\Bigg\{\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right)\frac{1}{(1-\lambda_w^2)^4}\Bigg[[1+7(E-1)\lambda_w^2]2a_5\alpha^2+\frac{1}{S}[1+32(E-1)\lambda_w^2]2c_5\beta^2\\
&\quad+[E+200(E-1)\lambda_w^2]4b_4d_5N^2B_{\max}^2\beta^2+128(E-1)[1+E(E-1)]\lambda_w^2 2d_5N^2B_{\max}^2\beta^2\Bigg]\\
&\quad+\frac{(\alpha+N^2B_{\max}^2\beta)L^2}{5N^3(1-\lambda_w^2)^4}\Bigg[N^2B_{\max}^2 2(b_5+b_4d_5)\beta^2+2a_5\alpha^2\Bigg]\Bigg\}T\frac{\sigma^2}{|\mathcal{I}|}\\
&+\Bigg\{\left(\frac{2\alpha L^2}{SN^3}+\frac{2\beta B_{\max}^2L^2}{SN}\right)[1+32(E-1)\lambda_w^2]\frac{2}{(1-\lambda_w^2)^4}\\
&\quad+\frac{L^2}{5N^3}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2a_3\alpha^2\Bigg\}\phi_z^0\\
&+\Bigg\{\Bigg[\left(\frac{2\alpha L^2}{N^3}+\frac{2\beta B_{\max}^2L^2}{N}\right)\frac{N^2B_{\max}^2\beta^2}{(1-\lambda_w^2)^4}\Bigg\{2b_4[E+200(E-1)\lambda_w^2]+1\\
&\quad+128(E-1)[1+E(E-1)]\lambda_w^2\Bigg]+\frac{L^2B_{\max}^2}{5N}(\alpha+N^2B_{\max}^2\beta)\frac{1}{(1-\lambda_w^2)^4}2b_4\beta^2\Bigg\}\phi_u^0 \qquad (121)\\
&\le \tilde{F}(\bar{\boldsymbol{\theta}}^0,\bar{\boldsymbol{x}}^0)-\underline{F}\\
&+[4ELa^2+10(E-1)L\beta^2]T\frac{\sigma^2}{|\mathcal{I}|}\\
&+\frac{(\alpha+N^2B_{\max}^2\beta)L^2}{5N^3(1-\lambda_w^2)^4}\Bigg\{[1+7(E-1)\lambda_w^2]4a_5\alpha^2+\frac{1}{S}[1+32(E-1)\lambda_w^2]4c_5\beta^2\\
&\quad+[E+200(E-1)\lambda_w^2]8b_4d_5N^2B_{\max}^2\beta^2+128(E-1)[1+E(E-1)]\lambda_w^2 4d_5N^2B_{\max}^2\beta^2\\
&\quad+N^2B_{\max}^2 2(b_5+b_4d_5)\beta^2+2a_5\alpha^2\Bigg\}T\frac{\sigma^2}{|\mathcal{I}|}\\
&+\frac{6(\alpha+N^2B_{\max}^2\beta)L^2}{SN^3(1-\lambda_w^2)^4}[1+32(E-1)\lambda_w^2]\phi_z^0\\
&+\frac{(\alpha+N^2B_{\max}^2\beta)L^2}{5N^3(1-\lambda_w^2)^4}\Bigg\{5N^2B_{\max}^2\beta^2\Bigg[10b_4[E+200(E-1)\lambda_w^2]+2\\
&\quad+640(E-1)[1+E(E-1)]\lambda_w^2\Bigg]+2b_4B_{\max}^2\beta^2\Bigg\}\phi_u^0\\
&\le \tilde{F}(\bar{\boldsymbol{\theta}}^0,\ \bar{\boldsymbol{x}}^0)-\underline{F}+2EL[2\alpha^2+5\beta^2]T\frac{\sigma^2}{|\mathcal{I}|}\\
&+(\alpha+N^2B_{\max}^2\beta)(327\alpha^2+221326N^2B_{\max}^2\beta^2)\frac{E^3L^2(1+7\lambda_w^2)^2T\sigma^2}{N^2(1-\lambda_w^2)^6|\mathcal{I}|}\\
&+\frac{(\alpha+N^2B_{\max}^2\beta)198EL^2}{SN^2(1-\lambda_w^2)^4}\phi_z^0+\frac{(\alpha+N^2B_{\max}^2\beta)E^3L^2(1+7\lambda_w^2)58681B_{\max}^2\beta^2}{N^2(1-\lambda_w^2)^5}\phi_u^0, \qquad (122)
\end{aligned}
$$

其中，最后两个不等式是通过令

$$\alpha \leq \frac{\sqrt{2[1+32(E-1)\lambda_w^2]}}{\sqrt{5Sa_3}}, \tag{123}$$

使得在（121）中，第二个系数小于 ϕ_z^0 之前的第一个系数；最后一个不等式是通过代入引理 9 和引理 10 中提到的 a_5，b_4，b_5，c_5，d_5，并利用 $\lambda_w^2 < 1$ 和 $E \geq 1$ 的性质得到的。

令 $\eta \triangleq \min\{\alpha, \beta\}$，则在不等式（122）两边同时除以 $\frac{\eta ET}{10}$，可得

$$\begin{aligned}
&\frac{1}{T}\sum_{r=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbb{E}\left[\left\|\frac{1}{NS}\sum_{i=1}^{S}\nabla_{\boldsymbol{\theta}} f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r, \boldsymbol{\theta}_n^r\right)\right\|^2\right]\right. \\
&\quad \left.+\mathbb{E}\left[\left\|\frac{1}{NS}\sum_{i=1}^{S}\sum_{t=1}^{N}\boldsymbol{B}_{t,i}^{\top}\nabla_{\boldsymbol{B}_i\boldsymbol{x}} f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r, \boldsymbol{\theta}_n^r\right)\right\|^2\right]\right\} \\
&\leq \frac{10(\bar{F}^0 - \underline{F})}{\eta ET} + \frac{200L(2\alpha^2+5\beta^2)\sigma^2}{\eta EN|\mathcal{I}|} \\
&\quad + 10(\alpha + N^2B_{\max}^2\beta)(327\alpha^2 + 221326N^2B_{\max}^2\beta^2)\frac{E^2L^2(1+7\lambda_w^2)^2\sigma^2}{\eta N^2(1-\lambda_w^2)^6|\mathcal{I}|} \\
&\quad + \frac{(\alpha+N^2B_{\max}^2\beta)1980L^2}{\eta TSN^2(1-\lambda_w^2)^4}\phi_z^0 + \frac{(\alpha+N^2B_{\max}^2\beta)E^2L^2(1+7\lambda_w^2)586810B_{\max}^2\beta^2}{\eta TN^2(1-\lambda_w^2)^5}\phi_u^0,.
\end{aligned} \tag{124}$$

不失一般性，假设 $N^2B_{\max}^2 \geq 1$；这可以通过取足够大的 N 并对数据进行归一化来增大 $B_{\max}$ 实现。然后，将 $\alpha = \beta = \eta = \frac{\sqrt{N}}{100\sqrt{ET}}$ 和 $E \leq \frac{T^3}{N^3}$ 代入（124），可得

$$\begin{aligned}
&\frac{1}{T}\sum_{r=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbb{E}\left[\left\|\frac{1}{NS}\sum_{i=1}^{S}\nabla_{\boldsymbol{\theta}} f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r, \boldsymbol{\theta}_n^r\right)\right\|^2\right]\right. \\
&\quad \left.+\mathbb{E}\left[\left\|\frac{1}{NS}\sum_{i=1}^{S}\sum_{t=1}^{N}\boldsymbol{B}_{t,i}^{\top}\nabla_{\boldsymbol{B}_i\boldsymbol{x}} f\left(\sum_{t=1}^{N}\boldsymbol{B}_{t,i}\boldsymbol{x}_n^r, \boldsymbol{\theta}_n^r\right)\right\|^2\right]\right\} \\
&\leq \frac{1000[\tilde{F}(\bar{\boldsymbol{\theta}}^0, \bar{\boldsymbol{x}}^0) - \underline{F}]}{\sqrt{NET}} + \frac{14L\sigma^2}{\sqrt{NET}|\mathcal{I}|} + \frac{222L^2(1+7\lambda_w^2)^2B_{\max}^2\sigma^2}{\sqrt{NET}N(1-\lambda_w^2)^6|\mathcal{I}|} \\
&\quad + \frac{2L^2B_{\max}^2}{\sqrt{NET}N(1-\lambda_w^2)^4}\left[\frac{990}{SEN}\phi_z^0 + \frac{59(1+7\lambda_w^2)}{(1-\lambda_w^2)T}\phi_u^0\right].
\end{aligned} \tag{125}$$

因此，我们完成了定理 2 的证明。. ■

参考文献

[1] G. B. Giannakis、Q. Ling、G. Mateos、I. D. Schizas 和 H. Zhu，“无线通信与网络的去中心化学习”，载于 *Splitting Methods in Communication, Imaging, Science, and Engineering*，第 461–497 页，Springer，2016年。[2] R. Bekkerman、M. Bilenko 和 J. Langford, *Scaling up Machine Learning- Parallel and Distributed Approaches* 2012 。剑桥大学出版社，。

[3] G. Scutari 和 Y. Sun，“大数据优化的并行分布式逐次凸近似方法”，载于 *Multi-agent Optimization*，第 141–308 页，Springer，2018年。[4] T.-H. Chang、M. Hong、H.-T. Wai、X. Zhang 和 S. Lu，“非凸世界中的分布式学习：从批量数据到流式处理及更多”，*IEEE Signal Processing Magazine* 37, 3, 26 38, ，第 2020，第 卷，第 号，第 页 – 页。[5] J. Zhang 、 S. Ge 、 T.-H. Chang 和 Z.-Q. Luo，“ 具有线性耦合约束的去中心化非凸学习”， *arXiv preprint arXiv:2103.05378*，2021年。[6] T. Yang、X. Yi、J. Wu、Y. Yuan、D. Wu、Z. Meng、Y. Hong、H. Wang、Z. Lin 和 K. H. Johansson，“分布式优化综述”， *Annual Reviews in Control*，第 47 卷，第 278–305 页，2019年。[7] A. Nedi ´ *Annual Review of Control Robotics and* c 和 *Autonomous Systems* 1, 77 103, 2018 J. Liu，“控制优化的分布式优化”，，，第 卷，第 页–页。[8] X. Lian、C. Zhang、H. Zhang、C.-J. Hsieh、W. Zhang 和 J. 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”,*Advances in Neural Information Processing Systems* 30, 2017 ，第 卷。[9] M. Hong、D. Hajinezhad 和 M.-M. Zhao，“Prox-PDA：用于快速分布式非凸优化和网络学习的近端原始-对偶算法”，载于 *Proc. Int. Conf. on Mach. Learn.*，第 1529–1538 页，2017年。[10] H. Sun 和 M. Hong，“分布式非凸一阶优化与信息处理：复杂度下界和速率最优算法”， *IEEE Transactions on Signal processing*，第 67 卷，第 22 期，第 5912–5928 页，2019年。[11] G. Qu 和 N. Li，“利用光滑性加速分布式优化”，*IEEE Transactions on Control of Network Systems* 5, 3, 1245 1260, 2017 ，第 卷，第 号，第 页–页。[12] S. Pu 和 A. Nedi´c，“分布式随机梯度跟踪方法”， *Mathematical Programming*，第 187 卷，第 1 期，第 409–457 页，2021年。[13] S. Lu、X. Zhang、H. Sun 和 M. Hong，“GNSD：一种用于去中心化优化的基于梯度跟踪的非凸随机算法”，载于 *Proc. IEEE Data Sci. Workshop*，第 315–321 页，2019年。[14] R. Xin、U. A. Khan 和 S. Kar，“去中心化在线随机非凸优化的改进收敛性分析”，*IEEE Transactions on Signal Processing*，第 69 卷，第 1842–1858 页，2021年。[15] H. Tang、X. Lian、M. Yan、C. Zhang 和 J. Liu，“D^2：在去中心化数据上进行去中心化训练”，载于 *Proc. Int. Conf. on Mach. Learn.*，第 4848–4856 页，2018年。[16] E. D. H. Nguyen、S. A. Alghunaim、K. Yuan 和 C. A. Uribe，“局部更新下梯度跟踪的性能”， *arXiv preprint arXiv:2210.04757*，2022年。[17] A. Nedic、A. Olshevsky 和 W. Shi，“在时变图上实现分布式优化的几何收敛”， *SIAM Journal on Optimization*，第 27 卷，第 4 期，第 2597–2633 页，2017年。[18] Z. Wang、J. Zhang、T.-H. Chang、J. Li 和 Z.-Q. Luo，“用于非凸非光滑问题的带动量分布式随机一致性优化”， *IEEE Transactions on Signal Processing*，第 69 卷，第 4486–4501 页，2021年。[19] P. Di Lorenzo 和 G. Scutari，“Next：网络内非凸优化 ”，*IEEE Transactions on Signal and Information Processing over Networks* 2, 2, 120 136, 2016 ，第 卷，第 号，第 页–页。[20] R. Xin、A. K. Sahu、U. A. Khan 和 S. Kar，“强连通网络上的梯度跟踪分布式随机优化”，载于 *Proc. IEEE Conf. Decis. Control*，第 8353–8358 页，2019年。[21] S. U. Stich，“本地 SGD 收敛迅速且通信量少”， *arXiv preprint arXiv:1805.09767*，2018年。[22] B. McMahan、E. Moore、D. Ramage、S. Hampson 和 B. A. y Arcas，“利用去中心化数据实现深度网络的通信高效学习”，载于 *Artificial intelligence and statistics*，第 1273–1282 页，2017年。[23] X. Li、K. Huang、W. Yang、S. Wang 和 Z. Zhang，“ 关于 FedAvg 在非独立同分布数据上的收敛性 ”，*arXiv preprint arXiv:1907.02189* 2019 ，。

[24] X. Zhang、W. Yin、M. Hong 和 T. Chen，“混合联邦学习：算法与实现”，*arXiv preprint arXiv:2012.12420*，2020年。[25] H. Gao、S. Ge 和 T.-H. Chang，“FedHD：基于混合数据的通信高效联邦学习”，*submitted to Journal of The Franklin Institute and under review*，2022年。

[26] A. H. Sayed, 《自适应网络》, *Proceedings of the IEEE*, 第102卷, 第4期, 第460–497页, 2014年。[27] L. Bottou、F. E. Curtis 和 J. Nocedal, 《大规模机器学习的优化方法》, *Siam Review* 60, 2, 223 311, 2018, 第 卷, 第 期, 第 – 页。[28] R. A. Horn 和 C. R. Johnson, *Matrix analysis*。剑桥大学出版社, 2012。[29] M. E. Yildiz 和 A. Scaglione, 《用于速率约束共识的利用边信息编码》, *IEEE Transactions on Signal Processing* 56, 8, 3753 3764, 2008, 第 卷, 第 期, 第 – 页。[30] Y. LeCun 和 C. Cortes, 《MNIST 手写数字数据库》, *Available: http://yann.lecun.com/exdb/mnist/* 2010。[31] C. Briggs、Z. Fan 和 P. Andras, 《通过对局部更新进行层次聚类, 改善非独立同分布数据上的联邦学习训练》, 载于 *Proc. IEEE Int. Joint Conf. on Neural Networks (IJCNN)*, 第1–9页, 2020年。[32] Y. Wang、Y. Xu、Q. Shi 和 T.-H. Chang, 《传输时延和中断约束下的量化联邦学习》, *IEEE Journal on Selected Areas in Communications*, 第40卷, 第1期, 第323–341页, 2021年。[33] I. Goodfellow、Y. Bengio 和 A. Courville, *Deep Learning*。麻省理工学院出版社, 2016。http://www. deeplearningbook.org。[34] V. Kulkarni、M. Kulkarni 和 A. Pant, 《联邦学习个性化技术综述》, 载于 *Proc. IEEE 4th World Conf. on Smart Trends in Systems, Security and Sustainability (WorldS4)*, 第794–797页, 2020年。[35] G. Scutari 和 Y. Sun, 《时变有向图上的分布式非凸约束优化》, *Mathematical Programming*, 第176卷, 第1期, 第497–544页, 2019年。[36] Y. Liao、Z. Li、K. Huang 和 S. Pu, 《具有线性收敛性的去中心化优化压缩梯度跟踪方法》, *IEEE Transactions on Automatic Control*, 第67卷, 第10期, 第5622–5629页, 2022年。[37] B. Ying、K. Yuan 和 A. H. Sayed, 《分布式特征下的监督学习》, *IEEE Transactions on Signal Processing* 67, 4, 977 992, 2019, 第 卷, 第 期, 第 – 页。