广义光滑性下有向图上的可证明收敛去中心化优化

Yanan Bo 和 Yongqiang Wang，IEEE 高级会员

arXiv:2601.03566v1 [math.OC] 7 Jan 2026

摘要——去中心化优化已成为大规模学习系统的一项基础工具；然而，现有的大多数方法都依赖经典的 Lipschitz 光滑性假设，而在梯度快速变化的问题中，这一假设往往不成立。针对这一局限，我们研究了广义 (L_0, L_1)-光滑性框架下的去中心化优化。在该框架中，海森范数允许随梯度范数线性增长，因此能够适用于超出经典 Lipschitz 光滑性范围的梯度快速变化情形。我们将梯度跟踪技术与梯度裁剪相结合，并仔细设计裁剪阈值，以确保在广义光滑性条件下，算法在有向通信图上准确收敛。与现有要求梯度差异性有界的广义光滑性分布式优化研究不同，即使梯度差异性无界，我们的结果仍然成立，因此所提出的框架更适用于真实的异构数据环境。我们在标准基准数据集上进行了数值实验，包括 LIBSVM 和 CIFAR-10，并采用正则化逻辑回归和卷积神经网络进行验证。实验结果表明，与现有方法相比，我们的方法具有更好的稳定性和更快的收敛速度。

索引词——去中心化优化、广义光滑性、有向图、梯度差异性、收敛保证

I. 引言

去中心化优化已成为大规模系统中的一种基础范式，在这类系统中，数据与计算分布在多个智能体之间。这类场景自然出现于广泛的应用中，包括分布式机器学习 [1], 多智能体协调与控制 [2],传感器与通信网络 [3]，智能电网和电力系统 [4], 以及现代云计算和内容分发基础设施 [5], [6]。在这些场景中，各智能体协同最小化全局目标函数，同时保持数据局部性，并在没有中央协调器的情况下运行。

尽管取得了重大进展，现有去中心化优化的理论基础大多建立在经典的 Lipschitz 光滑性假设之上，该假设要求每个局部目标函数的梯度在全局范围内均具有 Lipschitz 连续性。然而，在现代大规模和非凸学习问题中，这一假设可能过于严格，尤其是在涉及深度神经网络的问题中 [7], [8], [9], [10], [11], [12]。事实上，神经网络训练中的实验证据表明，海森范数通常会随损失函数的梯度范数而变化，这说明梯度可能沿着优化轨迹快速变化，从而不满足标准的光滑性条件。

一种更一般且更贴近现实的光滑性框架，称为 (L_0, L_1)-光滑性，由 [13]提出。在这一条件下，可微函数 g 满足

$$\|\nabla^2 g(\boldsymbol{\theta})\| \leqslant L_0 + L_1 \|\nabla g(\boldsymbol{\theta})\|, \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^d,$$

该条件允许海森范数随梯度范数增大，并严格推广了经典的 L_0-光滑性概念；当 $L_1 = 0$ 时，便可恢复经典概念。该框架涵盖了更广泛的函数类别，包括多项式函数和指数函数 [14], ，并能更准确地刻画现代机器学习模型（如长短期记忆网络 [13] 和 Transformer [15]）中遇到的损失景观。

本文考虑广义光滑性条件下的以下去中心化优化问题

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} F(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{\theta}), \tag{1}$$

其中，每个局部目标函数 f_i 由智能体 i 私有持有，且满足 (L_0, L_1)-光滑性。由于智能体只能获取局部信息，求解 (1) 需要通过通信网络进行局部计算和信息交换，以实现协同。

自提出以来，(L_0, L_1)-光滑性框架在优化与学习研究中日益受到关注，但现有结果大多聚焦于集中式场景。Zhang 等人 [13], [16] 采用 (L_0, L_1)-光滑性条件，从理论上解释了裁剪随机梯度下降相较于标准随机梯度下降的加速效应。后续研究将这些结果拓展到不同的算法变体和优化情境，包括加速梯度方法 [17], [18], 、带动量的裁剪随机梯度下降 [16], 、带动量的归一化梯度下降 [19], [20], 、差分隐私随机梯度下降 [21], 、广义 SignSGD [15], 、AdaGrad-Norm/AdaGrad [22], [23], 、Adam [24], 、(L_0, L_1)-Spider [25]以及分布鲁棒优化 [26]。

相比之下，在 (L_0, L_1)-光滑性条件下去中心化优化的最新成果仍十分有限。最近，Jiang 等人 [27] 提出了一种在宽松光滑性

本研究部分得到了美国国家科学基金会资助，项目编号为 CCF-2106293、CCF-2215088、CNS-2219487、CCF-2334449 和 CNS-2422312（通讯作者：Yongqiang Wang）。作者现任职于美国南卡罗来纳州克莱姆森市克莱姆森大学电气与计算机工程系，邮编 29634（电子邮箱：ybo@clemson.edu；yongqiw@clemson.edu）。

假设下用于去中心化优化的一阶方法，但该方法要求在每次迭代时进行全局平均，因此并非完全去中心化。Luo 等人 [28] 和 Sun [29] 提出了基于去中心化梯度下降并结合梯度归一化或裁剪的算法，以应对广义光滑性。然而，这些方法依赖有界梯度差异条件，极大地限制了它们在异质性场景中的适用性。具体而言，它们假设

$$\|\nabla f_i(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\| \leq \hat{b}, \tag{2}$$

ˆ 对某些常数 $b > 0$ [27], [30], [31], [32], [33] 成立。事实上，其他关于去中心化 (L_0, L_1)-光滑优化的现有研究也采用了这一条件 [27], [28], [29]。尽管这一条件在分析上较为方便，但在异质或大规模网络中可能过于严格：智能体的数据分布或模型架构可能存在显著差异，导致梯度差异随梯度范数增大，从而使这一假设失效。

本文放宽了这一条件，允许梯度差异性无界。具体而言，我们作如下假设：

$$\|\nabla f_i(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\| \leqslant (\ell - 1)\|\nabla F(\boldsymbol{\theta})\| + b,^{1} \tag{3}$$

其中，ℓ 和 b 是满足 $\ell \geqslant 1$ 和 $b > 0$ 的常数。这一放宽自然允许梯度差异随梯度范数增大，从而拓宽了适用范围。

此外，不同于现有依赖对称通信网络的去中心化优化 (L_0, L_1)- 光滑性研究，我们考虑的通信网络可以是非对称的，这可能是由非对称信息流、单向链路或异质的传输能力造成的 [34], [35]。值得注意的是，与无向图上的优化相比，有向图上的优化要复杂得多（例如，参见 [7], [8], [9], [36], [37], [38]）：非对称通信会导致有偏信息混合、无法采用双随机权重，以及非平凡的误差传播 [39]。事实上，所有关于有向图的现有分布式优化研究，包括 SGP [40]、SONATA [41]、Push-DIGing/ADDOPT [42] 和 Push–Pull/AB [43]，都依赖于标准的 Lipschitz 光滑性假设。如何将这些方法扩展到能够适应快速变化且具有异质性的梯度的更一般光滑性情形，仍鲜有研究。

在本研究中，我们探讨了广义光滑性条件下有向图上的分布式优化，且不要求有界梯度差异。不同于 [29]，对原始局部梯度进行裁剪、因而需要有界差异条件，我们的算法对全局梯度的局部估计进行裁剪。这种方法使我们能够有效处理局部目标函数之间的显著差异，以及有向通信带来的额外不平衡。由此，我们弥补了上述研究空白，并在广义且符合实际的条件下建立了可证明的收敛保证。

我们的主要贡献总结如下：

- 我们提出了一种去中心化优化算法，在无需有界梯度差异性的情况下，可在广义（L_0，L_1）-光滑性下实现有证明的收敛——据我们所知，这是此前尚未实现的性质。广义光滑性下的去中心化优化与经典 Lipschitz 光滑性下的现有工作（例如，[7]，[8]，[42]，[43]，[44]，[45]，[46]）存在根本差异，因为广义光滑性允许不同智能体之间的梯度范数发生快速且无界的变化。这种变化会显著放大去中心化优化中的一致性误差和梯度异质性，从而给收敛分析带来重大挑战。不同于依赖有界梯度差异性假设来处理广义光滑性的现有方法 [27]，[28]，[29]，我们的方法允许梯度差异性无界，从而反映出更加现实的异质性场景。

- 本研究的一项关键贡献是提出了一种新颖的算法设计，使算法能够在广义光滑条件下准确收敛。不同于 [29]，直接裁剪局部梯度，我们的方法对全局梯度的局部估计进行裁剪，从而无需有界梯度差异假设。这一设计并非易事，因为简单地裁剪局部梯度可能会放大智能体之间的差异，并严重阻碍收敛。因此，所提出的算法能够有效处理局部目标函数之间的显著异质性，以及有向通信造成的额外不平衡。

- 本工作的另一项重要贡献在于开发了新的证明技术。裁剪与局部梯度估计相结合会引入非线性、状态依赖的扰动，因此无法采用基于 Lipschitz 梯度的传统收敛分析方法。为此，我们通过精心设计算法参数，并推导出适用于我们更新结构的精细不等式，建立了一个新的理论框架。借助这一方法，我们首次为去中心化优化提供了收敛保证，同时考虑了梯度裁剪、有向通信网络和广义光滑性。事实上，我们证明该算法在 $\mathcal{O}(1/\epsilon^2)$ 次迭代内收敛至一个 ϵ 驻点，其复杂度界与相同光滑性条件 [13]下的集中式算法相匹配。

- 我们在基准数据集上通过数值实验验证了所提方法，包括 LIBSVM 和 CIFAR-10，并采用正则化逻辑回归和卷积神经网络进行测试。结果表明，与现有方法相比，我们的算法稳定性显著提升，收敛速度也更快。

本文结构安排如下。第二节介绍问题表述和假设。第三节介绍所提出的算法。第四节给出主要收敛结果，详细证明见附录。第五节提供数值实验，第六节

[1]此处采用参数化（$\ell - 1$），以便后续证明中的记号使用。

总结全文。

符号说明：令 $\boldsymbol{x}_i^k \in \mathbb{R}^d$ 表示智能体 i 在第 k次迭代时的局部优化变量，并将所有局部变量的集合定义为 $\boldsymbol{x}^k = [(\boldsymbol{x}_1^k)^\top; \cdots; (\boldsymbol{x}_N^k)^\top] \in \mathbb{R}^{N\times d}$。类似地，令 $\boldsymbol{y}_i^k \in \mathbb{R}^d$ 表示全局梯度的局部估计，$\boldsymbol{y}^k = [(\boldsymbol{y}_1^k)^\top; \cdots; (\boldsymbol{y}_N^k)^\top] \in \mathbb{R}^{N\times d}$ 表示所有智能体的估计值堆叠而成的向量。记在局部变量处计算得到的局部梯度集合为 $\nabla f(\boldsymbol{x}^k) = [\nabla f_1^\top(\boldsymbol{x}_1^k); \cdots; \nabla f_N^\top(\boldsymbol{x}_N^k)] \in \mathbb{R}^{N\times d}$。为便于分析，定义智能体 i在第 k次迭代中裁剪后的有效步长为 $\alpha_i^k = \alpha\min\{1, c_0/\|\boldsymbol{y}_i^k\|\}$，其中 α 和 c_0 为常数。注意，α_i^k 会随智能体和迭代次数而变化。缩放梯度记为 $\boldsymbol{\alpha}^k\boldsymbol{y}^k = [(\alpha_1^k\boldsymbol{y}_1^k)^\top; \cdots; (\alpha_N^k\boldsymbol{y}_N^k)^\top] \in \mathbb{R}^{N\times d}$。在平均变量 $\bar{\boldsymbol{x}}^k = \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i^k$处计算得到的全局梯度记为 $\nabla F(\bar{\boldsymbol{x}}^k) \in \mathbb{R}^d$。

II. 问题表述与预备知识

我们考虑一个由 N 个智能体组成的去中心化网络，这些智能体通过有向图进行通信。每个智能体 $i \in [N] \coloneqq \{1, 2, \cdots, N\}$ 都维护一个局部目标函数 $f_i\colon \mathbb{R}^d \to \mathbb{R}$，全局目标是最小化所有局部目标函数的平均值，即

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathbb{R}^{N\times d}} f(\boldsymbol{x}) = \tfrac{1}{N}\textstyle\sum_{i=1}^N f_i(\boldsymbol{x}_i), \\ &\text{s.t. } \boldsymbol{x}_1 = \boldsymbol{x}_2 = \cdots = \boldsymbol{x}_N, \end{aligned} \tag{4}$$

其中 $\boldsymbol{x} = [\boldsymbol{x}_1^\top;\ \boldsymbol{x}_2^\top;\ .....;\ \boldsymbol{x}_N^\top] \in \mathbb{R}^{N\times d}$。在本文中，局部目标函数和全局目标函数可以是非凸的。

我们首先介绍目标函数所需的假设和性质。

假设 1（下有界目标函数）。全局函数 f 下有界，即

$$\underline{f} := \inf_{\boldsymbol{x}\in\mathbb{R}^{N\times d}} f(\boldsymbol{x}) > -\infty.$$

假设 2（(L_0, L_1)-光滑性）。每个局部函数$f_i(\cdot)$二阶连续可微且满足(L_0^i, L_1^i)-光滑性，即存在常数 $L_0^i, L_1^i > 0$ 使得

$$\|\nabla^2 f_i(\boldsymbol{\theta})\| \leqslant L_0^i + L_1^i\|\nabla f_i(\boldsymbol{\theta})\|, \quad \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{5}$$

在假设 2 下，局部目标函数的海森范数可随梯度范数线性增长。这推广了标准的梯度利普希茨假设，该假设对应于 $L_1^i = 0$ 的情形。

假设 2 适用于广泛的实际目标函数。逻辑回归、用于图像分类的深度神经网络以及语言建模的实验证据表明，训练过程中局部光滑性会随梯度范数近似线性增长 [13], [27]。这一行为从根本上违背了传统的一致利普希茨光滑性假设，但可由（L_0, L_1）-光滑性条件自然刻画。

以下引理总结了 (L_0, L_1)-光滑性的两个关键性质，它们将在我们的收敛分析中发挥重要作用。

引理 1（[16], 引理 A.3）。设 g 是 (L_0, L_1)-光滑的，并设 $c > 0$ 为常数。对于任意满足 $\|\boldsymbol{\theta}-\boldsymbol{\vartheta}\| \leqslant c/L_1$ 的 $\boldsymbol{\theta}, \boldsymbol{\vartheta} \in \mathbb{R}^d$，有

$$\begin{aligned} g(\boldsymbol{\theta}) \leqslant\ & g(\boldsymbol{\vartheta}) + \left\langle \nabla g(\boldsymbol{\vartheta}),\, \boldsymbol{\theta} - \boldsymbol{\vartheta} \right\rangle \\ & + \frac{AL_0 + BL_1\|\nabla g(\boldsymbol{\vartheta})\|}{2}\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|^2, \end{aligned} \tag{6}$$

其中

$$A = 1 + e^c - \frac{e^c - 1}{c}, \qquad B = \frac{e^c - 1}{c}.$$

引理 2（[16], 推论 A.4）。设 g 是（L_0, L_1）-光滑的，且 $c > 0$ 为常数。对于任意满足$\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\| \leqslant c/L_1$ 的 $\boldsymbol{\theta}, \boldsymbol{\vartheta} \in \mathbb{R}^d$，有

$$\begin{aligned} &\|\nabla g(\boldsymbol{\theta}) - \nabla g(\boldsymbol{\vartheta})\| \\ &\leqslant \Big(AL_0 + BL_1\|\nabla g(\boldsymbol{\vartheta})\|\Big)\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|, \end{aligned} \tag{7}$$

其中

$$A = 1 + e^c - \frac{e^c - 1}{c}, \qquad B = \frac{e^c - 1}{c}.$$

接下来，我们讨论如何量化局部目标函数之间的异质性。

在去中心化优化的收敛分析中，一种常见做法是假设局部梯度之间的差异一致有界，即对所有 $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\frac{1}{N}\sum_{i=1}^N \big\|\nabla f_i(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\big\|^2 \leqslant \hat{b}^2 \tag{8}$$

ˆ 对某个常数 $b > 0$ 成立。这一条件在分布式优化中被广泛采用 [30], [31], [32], [33]。

当局部目标函数之间的异质性较弱时，式 (8) 中的一致有界条件可能成立；但在更一般的情形下，即使满足传统光滑性条件，该条件也会过于严格。事实上，即使是简单的二次函数，只要局部目标函数具有不同的曲率，这一假设也可能被违反 [47]。在异质或大规模网络中常见的广义光滑条件下，情况甚至更加棘手 [48]。例如，Bridge 回归可能采用 L_q-范数正则化项 $r(\boldsymbol{\theta}) = \sum_{j=1}^d |\theta_j|^q$，其中 $q > 2$ [49]。此类正则化项使目标函数具有 (L_0, L_1)-光滑性，但不具备 L-光滑性，因为当 $\|\boldsymbol{\theta}\| \to \infty$ 时，$\|\nabla^2 r(\boldsymbol{\theta})\| = \mathcal{O}(\|\boldsymbol{\theta}\|^{q-2})$ 会无界增长。当不同智能体为了应对非独立同分布数据 [48], 、学习个性化模型 [50], 或执行多任务学习 [51], ，而采用权重不同的正则化项时，正则化项之间的差异 $\|\nabla r_i(\boldsymbol{\theta}) - \frac{1}{N}\sum_{j=1}^N \nabla r_j(\boldsymbol{\theta})\|$ 会随 $\|\boldsymbol{\theta}\|$ 呈多项式增长。因此，目标函数本身会违反式 (8)。

受这一局限的启发，我们放宽了有界梯度差异性条件，允许局部梯度与全局梯度之间的差异随全局梯度的大小而变化，从而更好地刻画异质优化景观。

假设 3. 存在常数 $\ell \geqslant 1$ 和 $b > 0$，使得对任意 $\boldsymbol{\theta} \in \mathbb{R}^d$, 以下不等式成立：

$$\|\nabla f_i(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\| \leqslant (\ell - 1)\|\nabla F(\boldsymbol{\theta})\| + b. \tag{9}$$

假设 3 将有界梯度差异性条件推广为允许局部梯度与全局梯度之间的偏差随 $\|\nabla F(\boldsymbol{\theta})\|$ 线性变化。容易验证，式 (8) 是假设 3 在 $\ell = 1$ 条件下的特例。此外，可以验证，即使不同智能体在其正则化项中采用不同的 q，上述桥接回归问题仍满足假设 3。

此外，在假设 2 下，我们可以证明，全局函数 $F(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{\theta})$ 也是 (L_0, L_1)-光滑的，具体见下方的引理3，其证明见附录A。

引理3. 当每个 f_i 根据假设 2 均为 (L_0, L_1)-光滑时，我们有全局目标函数 F $\boldsymbol{\theta}$ — $\boldsymbol{\theta}$ L_0 L_1 $L_0 = ( )$ $\frac{1}{N}\sum_{i=1}^{N}\left(L_0^i + L_1^i b\right)$ $L_1 = \frac{\ell}{N}\sum_{i=1}^{N} L_1^i$ $= \frac{1}{N}\sum_{i=1}^{N} f_i( )$ 也是 (,)-光滑的，其中 和 。

最后，我们描述关于底层有向通信图的假设，该图由两个混合矩阵 $\boldsymbol{R}$ 和 $\boldsymbol{C}$ 描述。

假设 4（混合矩阵）。矩阵 $\boldsymbol{R} \in \mathbb{R}^{N\times N}$ 非负且为行随机矩阵（$\boldsymbol{R}\mathbf{1} = \mathbf{1}$），矩阵 $\boldsymbol{C} \in \mathbb{R}^{N\times N}$ 非负且为列随机矩阵（$\mathbf{1}^\top \boldsymbol{C} = \mathbf{1}^\top$）。二者的对角元素均为正。由 $\boldsymbol{R}$ 诱导的有向图 $\mathcal{G}_{\boldsymbol{R}}$ 至少包含一棵生成树，而由 $\boldsymbol{C}$ 诱导的有向图 $\mathcal{G}_{\boldsymbol{C}^\top}$ 是强连通的。

在假设 4 下，我们回顾 [43] 中关于混合矩阵谱性质的一些结果。

引理 4（[43], 引理 1）。在假设 4 下，矩阵$\boldsymbol{R}$ 具有一个非负左特征向量 $\boldsymbol{u}^\top$ （对应于特征值 1），满足 $\boldsymbol{u}^\top\mathbf{1} = N$。类似地，矩阵$\boldsymbol{C}$ 具有一个严格为正的右特征向量 $\boldsymbol{v}$ （对应于特征值 1），满足 $\mathbf{1}^\top\boldsymbol{v} = N$。此外，我们有$\boldsymbol{u}^\top\boldsymbol{v} > 0$。

引理 5（[43], 引理3）。假设假设 4 成立。令 ρ_R 和 ρ_C 分别为$\left(\boldsymbol{R} - \frac{1}{N}\mathbf{1}\boldsymbol{u}^\top\right)$ 和 $\left(\boldsymbol{C} - \frac{1}{N}\boldsymbol{v}\mathbf{1}^\top\right)$的谱半径。则有 $\rho_R < 1$ 和 $\rho_C < 1$。

引理 6（[43], 引理 4）。存在矩阵范数 $\|\cdot\|_R$和 $\|\cdot\|_C$，使得

$$\sigma_R := \left\|\boldsymbol{R} - \frac{1}{N}\mathbf{1}\boldsymbol{u}^\top\right\|_R < 1, \quad \sigma_C := \left\|\boldsymbol{C} - \frac{1}{N}\boldsymbol{v}\mathbf{1}^\top\right\|_C < 1,$$

且 σ_R 和 σ_C 分别可以任意接近 ρ_R 和 ρ_C。

此外，给定任意对角矩阵 $\boldsymbol{W} \in \mathbb{R}^{N\times N}$，有

$$\|\boldsymbol{W}\|_R = \|\boldsymbol{W}\|_C = \|\boldsymbol{W}\|_2.$$

我们还回顾以下范数等价结果：

引理 7（[43], 引理 6）。存在常数$\delta_{C,\,R}$、$\delta_{C,\,2}$、$\delta_{R,\,C}$、$\delta_{R,\,2} > 0$，使得对所有 $\boldsymbol{\theta} \in \mathbb{R}^d$，均有

$$\|\boldsymbol{\theta}\|_C \leqslant \delta_{C,R}\|\boldsymbol{\theta}\|_R, \quad \|\boldsymbol{\theta}\|_C \leqslant \delta_{C,2}\|\boldsymbol{\theta}\|_2,$$
$$\|\boldsymbol{\theta}\|_R \leqslant \delta_{R,C}\|\boldsymbol{\theta}\|_C, \quad \|\boldsymbol{\theta}\|_R \leqslant \delta_{R,2}\|\boldsymbol{\theta}\|_2.$$

此外，对范数 $\|\cdot\|_R$和$\|\cdot\|_C$进行适当的缩放后，对所有 $\boldsymbol{\theta} \in \mathbb{R}^d$，均有 $\|\boldsymbol{\theta}\|_2 \leqslant \|\boldsymbol{\theta}\|_R$ 和 $\|\boldsymbol{\theta}\|_2 \leqslant\|\boldsymbol{\theta}\|_C$。

本节中的假设和引理是证明所提出的分散式优化算法在有向通信图上的收敛性所必需的。

III. 所提出的算法

本节提出一种新的去中心化优化算法，即使智能体之间的梯度差异无界，该算法仍能在有向图上满足广义光滑条件时确保精确收敛。其基本思路是利用梯度跟踪框架，对全局梯度的局部估计进行梯度裁剪。据我们所知，这是首项将梯度裁剪融入梯度跟踪的工作，旨在抑制广义光滑性导致各智能体优化变量之间差异迅速增大的问题。

值得注意的是，这种融合给收敛分析带来了重大挑战。具体而言，引入裁剪会产生非线性、依赖状态的扰动，因此无法直接采用依赖梯度满足利普希茨条件的传统梯度跟踪收敛分析方法。为克服这些困难，我们通过精心设计算法参数，并推导适用于该更新结构的精细不等式，建立了新的理论框架。所提出的算法总结于算法 1 中，收敛分析将在下一节给出。

算法 1 裁剪梯度跟踪（CGT）

选择步长 $\alpha > 0$、裁剪阈值 $c_0 > 0$，
对所有 $j \in \mathcal{N}^{\text{in}}_{\boldsymbol{R},i}$，选取界内混合权重 $R_{ij} \geqslant 0$
并对所有 $l \in \mathcal{N}_{\text{out}\boldsymbol{C},i}$，选取界外权重 $C_{li} \geqslant 0$；
每个智能体 i 以任意的 $\boldsymbol{x}_i^0 \in \mathbb{R}^d$ 和 $\boldsymbol{y}_i^0 =$ 初始化 $\nabla f_i(\boldsymbol{x}_i^0)$;
对于 $k = 0, 1, \cdots$, **do**
对于每个 $i \in [N]$,
智能体 i 从每个 $j \in \mathcal{N}^{\text{in}}_{\boldsymbol{R}}$接收 $\boldsymbol{x}_j^k,\ i$;
智能体 i向每个 $l \in \mathcal{N}$发送 $C_{li}\boldsymbol{y}_i^k{}_{\text{out}\boldsymbol{C},\ i}$;
对于每个 $i \in [N]$,

$$\boldsymbol{x}_i^{k+1} = \sum_{j=1}^{N} R_{ij}\boldsymbol{x}_j^k - \alpha \min\left\{1, \frac{c_0}{\|\boldsymbol{y}_i^k\|}\right\}\boldsymbol{y}_i^k \tag{10}$$

$$\boldsymbol{y}_i^{k+1} = \sum_{j=1}^{N} C_{ij}\boldsymbol{y}_j^k + \nabla f_i(\boldsymbol{x}_i^{k+1}) - \nabla f_i(\boldsymbol{x}_i^k) \tag{11}$$

结束循环

该算法遵循标准的梯度跟踪框架：除了维护局部优化变量 $\boldsymbol{x}_i^k$ 外，每个智能体 i 还维护一个辅助变量 $\boldsymbol{y}_i^k$，用于跟踪全局梯度的变化。如 [52], 中所述，引入这一额外变量对于确保分散式优化中的下降方向准确至关重要，尤其是在各智能体的数据存在异质性的情况下。

在每次迭代中，每个智能体都会通过行随机矩阵 $\boldsymbol{R}$，将当前优化变量 $\boldsymbol{x}_i^k$ 与从其入邻居处接收的变量进行混合；同时，梯度跟踪变量 $\boldsymbol{y}_i^k$ 则通过列随机矩阵 $\boldsymbol{C}$ 进行混合。

与传统梯度跟踪框架的一个根本区别在于，我们对局部跟踪变量 $\boldsymbol{y}_i^k$ 采用裁剪操作，这对于抑制因 (L_0, L_1)-光滑性所允许的快速梯度变化而导致的智能体间差异迅速增大至关重要。具体而言，当 $\boldsymbol{y}_i^k$ 的范数超过 c_0 时，将其限制为 c_0；当其范数低于 c_0 时，则保持不变。值得一提的是，在我们的算法中，这种裁剪操作施加于 $\boldsymbol{y}_i^k$，而不是像 [29] 那样直接作用于局部梯度 $\nabla f_i(\boldsymbol{x}_i^k)$。我们认为，这对于我们获得优于 [29] 的结果至关重要，因为在异质环境中，不同智能体的局部梯度可能存在显著差异，直接对其进行裁剪会加剧智能体间的状态差异。相比之下，跟踪变量是全局梯度的估计量，因此更加稳定，可以对其进行裁剪而不影响收敛性。

IV. 收敛分析

本节将严格证明，即使智能体间的梯度差异可能无界，算法 1 在广义光滑性和有向通信图条件下仍能确保精确收敛。据我们所知，这是首次建立此类结果。为此，我们首先引入简洁的符号表示，并刻画关键误差量。随后，我们针对裁剪步长、梯度有界性和误差动态给出辅助结果。最后，我们结合这些结果证明主要的收敛定理。

A. 矩阵形式与误差定义

为分析所提出算法的收敛性，我们首先将算法 1 中的更新规则表示为紧凑的矩阵形式。(10) 和 (11) 中的迭代可写为

$$\boldsymbol{x}^{k+1} = \boldsymbol{R}\boldsymbol{x}^k - \boldsymbol{\alpha}_k \boldsymbol{y}^k, \tag{12}$$

$$\boldsymbol{y}^{k+1} = \boldsymbol{C}\boldsymbol{y}^k + \nabla f(\boldsymbol{x}^{k+1}) - \nabla f(\boldsymbol{x}^k), \tag{13}$$

其中 $\boldsymbol{x}^k = [(\boldsymbol{x}_1^k)^\top; \cdots; (\boldsymbol{x}_N^k)^\top] \in \mathbb{R}^{N\times d}$，$\boldsymbol{y}^k = [(\boldsymbol{y}_1^k)^\top; \cdots; (\boldsymbol{y}_N^k)^\top] \in \mathbb{R}^{N\times d}$，以及 $\boldsymbol{\alpha}^k\boldsymbol{y}^k = [(\alpha_1^k\boldsymbol{y}_1^k)^\top; \cdots; (\alpha_N^k\boldsymbol{y}_N^k)^\top] \in \mathbb{R}^{N\times d}$。

我们将网络范围平均变量定义为

$$\bar{\boldsymbol{x}}^k := \tfrac{1}{N}\boldsymbol{u}^\top\boldsymbol{x}^k \in \mathbb{R}^d, \qquad \bar{\boldsymbol{y}}^k := \tfrac{1}{N}\mathbf{1}^\top\boldsymbol{y}^k \in \mathbb{R}^d,$$

其中，$\boldsymbol{u}$ 是矩阵 $\boldsymbol{R}$ 对应于特征值 1 的左特征向量（见引理 4）。利用上述式 (12) 和 (13) 中的更新规则，可以得到 $\bar{\boldsymbol{x}}^k$ 和 $\bar{\boldsymbol{y}}^k$ 的动力学如下：

$$\bar{\boldsymbol{x}}^{k+1} = \bar{\boldsymbol{x}}^k - \tfrac{1}{N}\boldsymbol{u}^\top\boldsymbol{\alpha}_k\boldsymbol{y}^k, \tag{14}$$

$$\bar{\boldsymbol{y}}^{k+1} = \bar{\boldsymbol{y}}^k + \tfrac{1}{N}\mathbf{1}^\top\big(\nabla f(\boldsymbol{x}^{k+1}) - \nabla f(\boldsymbol{x}^k)\big). \tag{15}$$

为刻画智能体之间的分歧，我们将一致性误差定义如下：

$$\boldsymbol{e}_{x,k} := \boldsymbol{x}^k - \mathbf{1}(\bar{\boldsymbol{x}}^k)^\top \in \mathbb{R}^{N\times d}, \tag{16}$$

该变量衡量每个智能体的局部变量偏离全局平均值的程度。

类似地，梯度跟踪误差定义为

$$\boldsymbol{e}_{y,k} := \boldsymbol{y}^k - \boldsymbol{v}(\bar{\boldsymbol{y}}^k)^\top \in \mathbb{R}^{N\times d}, \tag{17}$$

其中，$\boldsymbol{v}$ 是矩阵 $\boldsymbol{C}$ 对应于特征值 1 的右特征向量。

误差矩阵$\boldsymbol{e}_{x,\,k}$ 和$\boldsymbol{e}_{y,\,k}$ 的第 i 行满足

$$\boldsymbol{e}_{x,k,i} = (\boldsymbol{x}_i^k)^\top - (\bar{\boldsymbol{x}}^k)^\top, \qquad \boldsymbol{e}_{y,k,i} = (\boldsymbol{y}_i^k)^\top - (v_i\bar{\boldsymbol{y}}^k)^\top.$$

利用恒等式 $\boldsymbol{R}\mathbf{1} = \mathbf{1}$ 和 $\mathbf{1}^\top\boldsymbol{C} = \mathbf{1}^\top$，以及式（10）和（11）中的更新规则，可以验证一致性误差的演化形式为

$$\boldsymbol{e}_{x,k+1} = \left(\boldsymbol{R} - \tfrac{\mathbf{1}\boldsymbol{u}^\top}{N}\right)\boldsymbol{e}_{x,k} - \left(\boldsymbol{I} - \tfrac{\mathbf{1}\boldsymbol{u}^\top}{N}\right)\boldsymbol{\alpha}_k\boldsymbol{y}^k, \tag{18}$$

梯度跟踪误差的演化形式为

$$\boldsymbol{e}_{y,k+1} = \left(\boldsymbol{C} - \tfrac{\boldsymbol{v}\mathbf{1}^\top}{N}\right)\boldsymbol{e}_{y,k} + \left(\boldsymbol{I} - \tfrac{\boldsymbol{v}\mathbf{1}^\top}{N}\right)\big(\nabla f(\boldsymbol{x}^{k+1}) - \nabla f(\boldsymbol{x}^k)\big). \tag{19}$$

这些关系对于我们的收敛分析至关重要。

B. 辅助结果

我们首先量化局部裁剪步长之间的差异。

引理 8. 对于算法 1 中的任意智能体 $i \in [N]$，将裁剪后的局部步长记为

$$\alpha_i^k = \alpha \min\left\{1,\ \frac{c_0}{\|\boldsymbol{y}_i^k\|}\right\},$$

将基于网络平均梯度的步长记为

$$\bar{\alpha}_i^k = \alpha \min\left\{1,\ \frac{c_0}{v_i\|\nabla F(\bar{\boldsymbol{x}}^k)\|}\right\}.$$

那么，在假设 1、假设 2 和假设 4 下，以下不等式成立：

$$\left|\alpha_i^k - \bar{\alpha}_i^k\right| \|\boldsymbol{y}_i^k\| \ \leqslant\ \bar{\alpha}_i^k \|\boldsymbol{y}_i^k - v_i\nabla F(\bar{\boldsymbol{x}}^k)\|. \tag{20}$$

此外，将全局步长记为

$$\bar{\alpha}_k = \alpha \min\left\{1,\ \frac{c_0}{\|\boldsymbol{v}\|\|\nabla F(\bar{\boldsymbol{x}}^k)\|}\right\},$$

则有

$$\bar{\alpha}_k \ \leqslant\ \bar{\alpha}_i^k \ \leqslant\ \frac{\|\boldsymbol{v}\|}{v_i}\bar{\alpha}_k,$$

and

$$\left|\alpha_i^k - \bar{\alpha}_i^k\right| \|\boldsymbol{y}_i^k\| \ \leqslant\ \frac{\|\boldsymbol{v}\|}{v_i}\bar{\alpha}_k \|\boldsymbol{y}_i^k - v_i\nabla F(\bar{\boldsymbol{x}}^k)\|. \tag{21}$$

证明。详见附录 C。 □

为建立梯度的有界性，我们需要在引理 9 和引理 10 中推导一致性误差和梯度跟踪误差的界。

引理 9。假设 1、2 和 4 成立。对于由算法 1 生成的迭代序列，一致性误差$\|\boldsymbol{e}_{x,\ k}\|_R^2$ 是一致有界的。具体而言，我们有

$$\|\boldsymbol{e}_{x,k}\|_R^2 \leqslant \mathcal{C}_x\alpha^2c_0^2, \qquad \forall k\geqslant 0, \tag{22}$$

where $\mathcal{C}_x = \frac{2N\sigma_R^2(1+\sigma_R^2)\delta_{R,2}^2\|\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{u}^\top}{N}\|_R^2}{(1-\sigma_R^2)^2}$.

证明。见附录 D。 □

接下来，我们证明梯度跟踪误差也一致有界。

引理 10。假设假设 1、2、3 和 4 成立。在算法 1 下，若梯度对所有 $k\geqslant 0$ 满足 $\|\nabla F(\bar{\boldsymbol{x}}^k)\|\leqslant G$，则梯度跟踪误差满足

$$\|\boldsymbol{e}_{y,k}\|_C^2 \leqslant \mathcal{C}_y\alpha^2c_0^2, \qquad \forall k\geqslant 0,$$

where $\mathcal{C}_y = \frac{2(1+\sigma_C^2)\delta_{C,2}^2\|\boldsymbol{I}-\frac{\boldsymbol{v}\boldsymbol{1}^\top}{N}\|_C^2}{(1-\sigma_C^2)^2}\mathcal{C}_1$, *with* $\mathcal{C}_1 = 2(AL_0+BL_1b+BL_1\ell G)^2\big(2N+(1+2\sigma_R^2)\mathcal{C}_x\big)$.

证明。见附录 E。 □

利用这些误差界，我们现在证明 $\|\nabla F(\bar{\boldsymbol{x}}^k)\|$ 一致有界。

引理 11。假设假设 1、2、3 和 4 成立。若 α 满足 $0<\alpha\leqslant\frac{\boldsymbol{u}^\top\boldsymbol{v}}{9LN\|\boldsymbol{v}\|^2}$，且 $c_0=1/\sqrt{K}$，则算法 1 生成的迭代序列满足

$$\|\nabla F(\bar{\boldsymbol{x}}^k)\|\leqslant G,\quad \forall\, k\leqslant K,$$

其中

$$G=\sup\Big\{t>0\ \Big|\ t^2\leqslant 2\big(L_0+2L_1t\big)\big(f(\bar{\boldsymbol{x}}^0)-\underline{f}+\alpha^3\mathcal{C}_f\big)\Big\},$$

and

$$\mathcal{C}_f=\left(\frac{3L\kappa_{uv}^2\|\boldsymbol{v}\|^2\alpha}{N}+\frac{2\kappa_{uv}^2\|\boldsymbol{v}\|^2}{N\boldsymbol{u}^\top\boldsymbol{v}}\right)\big(2\mathcal{C}_y+2L^2\|\boldsymbol{v}\|^2\mathcal{C}_x\big).$$

证明。我们利用数学归纳法证明该结论，并以引理 9、10 和 15（附录 B）为基础。

显然，对于情形 $k=0$，由于$\|\nabla F(\bar{\boldsymbol{x}}^0)\|\leqslant G$，结论显然成立。

接下来，我们证明：若 $\|\nabla F(\bar{\boldsymbol{x}}^k)\|\leqslant G$ 对 $k\geqslant 0$ 成立，则该不等式对 $k+1$ 也成立。

根据 (14) 中 $\bar{\boldsymbol{x}}^k$ 的动力学以及引理 1，对于 $F(\bar{\boldsymbol{x}}^{k+1})$，我们有以下不等式：

$$\begin{aligned}F(\bar{\boldsymbol{x}}^{k+1})\leqslant&F(\bar{\boldsymbol{x}}^k)-\big\langle\nabla F(\bar{\boldsymbol{x}}^k),\bar{\boldsymbol{x}}^{k+1}-\bar{\boldsymbol{x}}^k\big\rangle\\&+\frac{AL_0+BL_1\|\nabla F(\bar{\boldsymbol{x}}^k)\|}{2}\|\bar{\boldsymbol{x}}^{k+1}-\bar{\boldsymbol{x}}^k\|^2\\\leqslant&F(\bar{\boldsymbol{x}}^k)-\left\langle\nabla F(\bar{\boldsymbol{x}}^k),\frac{1}{N}\boldsymbol{u}^\top\boldsymbol{\alpha}_k\boldsymbol{y}^k\right\rangle\\&+\frac{L}{2}\|\frac{1}{N}\boldsymbol{u}^\top\boldsymbol{\alpha}_k\boldsymbol{y}^k\|^2,\end{aligned}\tag{23}$$

其中 $L=AL_0+BL_1G$。

对于内积项，我们有

$$\begin{aligned}&\big\langle\nabla F(\bar{\boldsymbol{x}}^k),\tfrac{1}{N}\boldsymbol{u}^\top\boldsymbol{\alpha}_k\boldsymbol{y}^k\big\rangle\\=&\big\langle\nabla F(\bar{\boldsymbol{x}}^k),\tfrac{1}{N}\boldsymbol{u}^\top\left(\boldsymbol{\alpha}_k\boldsymbol{y}^k-\tilde{\boldsymbol{\alpha}}_k\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right)\big\rangle\\&+\big\langle\nabla F(\bar{\boldsymbol{x}}^k),\tfrac{1}{N}\boldsymbol{u}^\top\tilde{\boldsymbol{\alpha}}_k\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\big\rangle,\end{aligned}\tag{24}$$

其中

$$\begin{aligned}&\tilde{\boldsymbol{\alpha}}_k\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\\=&[(\bar{\alpha}_1^kv_1\nabla F(\bar{\boldsymbol{x}}^k))^\top;\cdots;(\bar{\alpha}_N^kv_N\nabla F(\bar{\boldsymbol{x}}^k))^\top].\end{aligned}$$

我们首先分析 (24) 右侧的第一项，可以验证它满足

$$\begin{aligned}&\big\langle\nabla F(\bar{\boldsymbol{x}}^k),\tfrac{1}{N}\boldsymbol{u}^\top\left(\boldsymbol{\alpha}_k\boldsymbol{y}^k-\tilde{\boldsymbol{\alpha}}_k\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right)\big\rangle\\=&\big\langle\nabla F(\bar{\boldsymbol{x}}^k),\tfrac{1}{N}\boldsymbol{u}^\top\left(\boldsymbol{\alpha}_k\boldsymbol{y}^k-\tilde{\boldsymbol{\alpha}}_k\boldsymbol{y}^k\right)\big\rangle\\&+\big\langle\nabla F(\bar{\boldsymbol{x}}^k),\tfrac{1}{N}\boldsymbol{u}^\top\left(\tilde{\boldsymbol{\alpha}}_k\boldsymbol{y}^k-\tilde{\boldsymbol{\alpha}}_k\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right)\big\rangle,\end{aligned}\tag{25}$$

其中 $\tilde{\boldsymbol{\alpha}}_k\boldsymbol{y}^k=[(\bar{\alpha}_1^k\boldsymbol{y}_1^k)^\top;\ \cdots;\ (\bar{\alpha}_N^k\boldsymbol{y}_N^k)^\top]$。

对于 (25) 中的第一项，根据引理 8，我们有

$$\begin{aligned}&\left|\big\langle\nabla F(\bar{\boldsymbol{x}}^k),\tfrac{1}{N}\boldsymbol{u}^\top\left(\boldsymbol{\alpha}_k\boldsymbol{y}^k-\tilde{\boldsymbol{\alpha}}_k\boldsymbol{y}^k\right)\big\rangle\right|\\\leqslant&\frac{1}{N}\|\nabla F(\bar{\boldsymbol{x}}^k)\|\sum_{i=1}^N u_i\left|\alpha_i^k-\bar{\alpha}_i^k\right|\|\boldsymbol{y}_i^k\|\\\leqslant&\frac{1}{N}\|\nabla F(\bar{\boldsymbol{x}}^k)\|\sum_{i=1}^N u_i\bar{\alpha}_i^k\|\boldsymbol{y}_i^k-v_i\nabla F(\bar{\boldsymbol{x}}^k)\|\\\leqslant&\frac{\|\boldsymbol{v}\|}{N}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|\sum_{i=1}^N\frac{u_i}{v_i}\|\boldsymbol{y}_i^k-v_i\nabla F(\bar{\boldsymbol{x}}^k)\|,\end{aligned}\tag{26}$$

其中，最后一个不等式用到了关系 $\bar{\alpha}_k\leqslant\bar{\alpha}_i^k\leqslant\frac{\|\boldsymbol{v}\|}{v_i}\bar{\alpha}_k$。记 $\kappa_{uv}:=\sup_i\frac{u_i}{v_i}$，则 (26) 中的不等式可表示如下：

$$\begin{aligned}&\left|\big\langle\nabla F(\bar{\boldsymbol{x}}^k),\tfrac{1}{N}\boldsymbol{u}^\top\left(\boldsymbol{\alpha}_k\boldsymbol{y}^k-\tilde{\boldsymbol{\alpha}}_k\boldsymbol{y}^k\right)\big\rangle\right|\\\leqslant&\frac{\kappa_{uv}\|\boldsymbol{v}\|}{N}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|\sum_{i=1}^N\|\boldsymbol{y}_i^k-v_i\nabla F(\bar{\boldsymbol{x}}^k)\|.\end{aligned}\tag{27}$$

类似地，对于 (25) 右端的第二项，我们有

$$\begin{aligned}&\left|\big\langle\nabla F(\bar{\boldsymbol{x}}^k),\tfrac{1}{N}\boldsymbol{u}^\top\left(\tilde{\boldsymbol{\alpha}}_k\boldsymbol{y}^k-\tilde{\boldsymbol{\alpha}}_k\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right)\big\rangle\right|\\=&\frac{1}{N}\|\nabla F(\bar{\boldsymbol{x}}^k)\|\sum_{i=1}^N u_i\bar{\alpha}_i^k\|\boldsymbol{y}_i^k-v_i\nabla F(\bar{\boldsymbol{x}}^k)\|\\\leqslant&\frac{\kappa_{uv}\|\boldsymbol{v}\|}{N}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|\sum_{i=1}^N\|\boldsymbol{y}_i^k-v_i\nabla F(\bar{\boldsymbol{x}}^k)\|.\end{aligned}\tag{28}$$

结合 (27) 和 (28)，我们可以对 (23) 中的内积项作如下界定：

$$\begin{aligned}-\left\langle\nabla F(\bar{\boldsymbol{x}}^k),\frac{1}{N}\boldsymbol{u}^\top\boldsymbol{\alpha}_k\boldsymbol{y}^k\right\rangle\leqslant&-\frac{\boldsymbol{u}^\top\boldsymbol{v}}{2N}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2\\&+\frac{2\kappa_{uv}^2\|\boldsymbol{v}\|^2}{N\boldsymbol{u}^\top\boldsymbol{v}}\bar{\alpha}_k\|\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\|^2.\end{aligned}\tag{29}$$

利用引理 9 和引理 10 中的误差界，我们得到

$$\begin{aligned}\|\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\|^2=&\|\boldsymbol{y}^k-\boldsymbol{v}\bar{\boldsymbol{y}}^k+\boldsymbol{v}\bar{\boldsymbol{y}}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\|^2\\\leqslant&2\left\|\boldsymbol{y}^k-\boldsymbol{v}\bar{\boldsymbol{y}}^k\right\|^2+2\left\|\boldsymbol{v}\bar{\boldsymbol{y}}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\\\leqslant&2\mathcal{C}_y\alpha^2c_0^2+2L^2\|\boldsymbol{v}\|^2\mathcal{C}_x\alpha^2c_0^2.\end{aligned}\tag{30}$$

对于项 $\frac{L}{2N^2}\|\boldsymbol{u}^\top\boldsymbol{\alpha}_k\boldsymbol{y}^k\|^2$，结合恒等式 $\boldsymbol{u}^\top\boldsymbol{\alpha}_k\boldsymbol{y}^k = \sum_{i=1}^N u_i\alpha_i^k\boldsymbol{y}_i^k$ 和詹森不等式 $\left\|\sum_i z_i\right\|^2 \leqslant N\sum_i\|z_i\|^2$，我们可以对其进行界定，得到

$$\frac{L}{2N^2}\Big\|\sum_{i=1}^N u_i\alpha_i^k\boldsymbol{y}_i^k\Big\|^2 \leqslant \frac{L}{2N}\sum_{i=1}^N u_i^2\big\|\alpha_i^k\boldsymbol{y}_i^k\big\|^2. \tag{31}$$

对于（31）右端的 $\alpha_i^k\boldsymbol{y}_i^k$，我们可以加上并减去 $\bar{\alpha}_i^k\boldsymbol{y}_i^k$ 和 $v_i\bar{\alpha}_i^k\nabla F(\bar{\boldsymbol{x}}^k)$，得到

$$\alpha_i^k\boldsymbol{y}_i^k = (\alpha_i^k-\bar{\alpha}_i^k)\boldsymbol{y}_i^k + \bar{\alpha}_i^k(\boldsymbol{y}_i^k - v_i\nabla F(\bar{\boldsymbol{x}}^k)) + v_i\bar{\alpha}_i^k\nabla F(\bar{\boldsymbol{x}}^k). \tag{32}$$

进一步利用不等式 $\|a+b+c\|^2\leqslant 3$（$\|a\|^2+\|b\|^2+\|c\|^2$），得到

$$\begin{aligned}\|\alpha_i^k\boldsymbol{y}_i^k\|^2 \leqslant& 3\Big(\sum_{i=1}^N u_i^2\left\|\alpha_i^k\boldsymbol{y}_i^k-\bar{\alpha}_i^k\boldsymbol{y}_i^k\right\|^2\\ &+\sum_{i=1}^N u_i^2(\bar{\alpha}_i^k)^2\left\|\boldsymbol{y}_i^k - v_i\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\\ &+\sum_{i=1}^N\left\|v_i\bar{\alpha}_i^k\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\Big).\end{aligned} \tag{33}$$

根据引理 8，我们有：

$$\begin{aligned}&\frac{L}{2N^2}\|\boldsymbol{u}^\top\boldsymbol{\alpha}_k\boldsymbol{y}^k\|^2\\ \leqslant&\frac{3L}{2N}\Big(2\bar{\alpha}_k^2\sum_{i=1}^N\frac{u_i^2}{v_i^2}\|\boldsymbol{v}\|^2\left\|\boldsymbol{y}_i^k - v_i\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\\ &+\|\boldsymbol{v}\|^2\bar{\alpha}_k^2N\left\|\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\Big)\\ \leqslant&\frac{3L}{2}\|\boldsymbol{v}\|^2\bar{\alpha}_k^2\left\|\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\\ &+\frac{3L\|\boldsymbol{v}\|^2\kappa_{uv}^2}{N}\bar{\alpha}_k^2\left\|\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2.\end{aligned} \tag{34}$$

结合（29）、（30）和（34），我们在 $\alpha\leqslant\frac{\boldsymbol{u}^\top\boldsymbol{v}}{9LN\|\boldsymbol{v}\|^2}$ 下得到以下关系：

$$\begin{aligned}&F(\bar{\boldsymbol{x}}^{k+1})-\underline{f}+\frac{\boldsymbol{u}^\top\boldsymbol{v}}{3N}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2\leqslant F(\bar{\boldsymbol{x}}^k)-\underline{f}\\ &+\left(\frac{3L\kappa_{uv}^2\|\boldsymbol{v}\|^2\bar{\alpha}_k}{N}+\frac{2\kappa_{uv}^2\|\boldsymbol{v}\|^2}{N\boldsymbol{u}^\top\boldsymbol{v}}\right)\bar{\alpha}_k\|\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\|^2\\ \leqslant&F(\bar{\boldsymbol{x}}^k)-\underline{f}+\mathcal{C}_f\alpha^3c_0^2,\end{aligned} \tag{35}$$

其中

$$\mathcal{C}_f=\left(\frac{3L\kappa_{uv}^2\|\boldsymbol{v}\|^2\alpha}{N}+\frac{2\kappa_{uv}^2\|\boldsymbol{v}\|^2}{N\boldsymbol{u}^\top\boldsymbol{v}}\right)\left(2\mathcal{C}_y+2L^2\|\boldsymbol{v}\|^2\mathcal{C}_x\right).$$

在 $c_0=\frac{1}{\sqrt{K}}$ 下对 $s\leqslant k+1\leqslant K$ 求和，我们有

$$\begin{aligned}&F(\bar{\boldsymbol{x}}^{k+1})-\underline{f}+\frac{\boldsymbol{u}^\top\boldsymbol{v}}{3N}\sum_{s=0}^k\bar{\alpha}^s\|\nabla F(\bar{\boldsymbol{x}}^s)\|^2\\ \leqslant&F(\bar{\boldsymbol{x}}^0)-\underline{f}+K\mathcal{C}_f\alpha^3c_0^2\\ \leqslant&F(\bar{\boldsymbol{x}}^0)-\underline{f}+\mathcal{C}_f\alpha^3.\end{aligned} \tag{36}$$

由引理 16 可得 $\|\nabla F(\bar{\boldsymbol{x}}^{k+1})\|\leqslant G$。
因此，对于所有 $k\leqslant K$，我们有 $\|\nabla F(\bar{\boldsymbol{x}}^k)\|\leqslant G$。 □

在确立梯度有界性后，我们现在可以推导误差动态的递归界。

引理 12. 假设 1、2、3 和 4 成立。若步长 α 满足

$$0<\alpha\leqslant\min\left\{\frac{(1-\sigma_R^2)\sqrt{N}}{6\sqrt{2}\|\boldsymbol{v}\|\kappa_v\|\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{u}^\top}{N}\|_R\delta_{R,2}(AL_0+BL_1b+BL_1\ell G)},\right.$$
$$\left.\frac{1-\sigma_C^2}{12\sqrt{2}\kappa_v\delta_{C,2}\|\boldsymbol{I}-\frac{\boldsymbol{v}\boldsymbol{1}^\top}{N}\|_C(AL_0+BL_1b+BL_1\ell G)}\right\}, \tag{37}$$

那么，一致性误差 $\boldsymbol{e}_{x,\,k}$ 和梯度跟踪误差 $\boldsymbol{e}_{y,\,k}$ 对所有 $k\geqslant 0$ 均满足以下关系：

$$\begin{aligned}\|\boldsymbol{e}_{x,k+1}\|_R^2\leqslant&\tfrac{1+\sigma_R^2}{2}\|\boldsymbol{e}_{x,k}\|_R^2+\alpha^2\mathcal{C}_{x,1}\|\boldsymbol{e}_{y,k}\|_C^2\\ &+\alpha\,\mathcal{C}_{x,2}\,\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2,\\ \|\boldsymbol{e}_{y,k+1}\|_C^2\leqslant&\mathcal{C}_{y,1}\|\boldsymbol{e}_{x,k}\|_R^2+\tfrac{1+\sigma_C^2}{2}\|\boldsymbol{e}_{y,k}\|_C^2\\ &+\alpha\,\mathcal{C}_{y,2}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2,\end{aligned} \tag{38}$$

其中

$$\mathcal{C}_{x,1}=\frac{12(1+2\sigma_R^2)\|\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{u}^\top}{N}\|_R^2\delta_{R,2}^2\kappa_v^2}{1-\sigma_R^2}, \tag{39}$$

$$\mathcal{C}_{x,2}=\frac{3N\left(1+\sigma_R^2\right)\|\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{u}^\top}{N}\|_R^2\delta_{R,2}^2\|\boldsymbol{v}\|^2}{1-\sigma_R^2}, \tag{40}$$

$$\mathcal{C}_{y,1}=\frac{(1+2\sigma_C^2)\|\boldsymbol{I}-\frac{\boldsymbol{v}\boldsymbol{1}^\top}{N}\|_C^2\delta_{C,2}^2}{1-\sigma_C^2}(2\sigma_R^2+\frac{(1+2\sigma_C^2)(1-\sigma_C^2)\|\boldsymbol{v}\|^2}{12N}), \tag{41}$$

$$\mathcal{C}_{y,2}=\frac{12N(1+\sigma_C^2)\left\|\boldsymbol{I}-\frac{\boldsymbol{v}\boldsymbol{1}^\top}{N}\right\|_C^2\|\boldsymbol{v}\|^2\kappa_v^2\delta_{C,2}^2\left(AL_0+BL_1b+BL_1\ell G\right)^2}{1-\sigma_C^2}. \tag{42}$$

证明。见附录 F。 □

在刻画了一致性误差和梯度跟踪误差后，我们现在来刻画 $\nabla F(\bar{\boldsymbol{x}}^k)$ 的下降行为。这一下降性质将成为建立算法 1 收敛保证的关键要素。

引理 13. 假设假设 2、3 和 4 成立。若步长满足 $\alpha\leqslant\frac{\boldsymbol{u}^\top\boldsymbol{v}}{9LN\|\boldsymbol{v}\|^2}$，则算法1生成的迭代序列满足以下不等式：

$$\begin{aligned}\frac{\boldsymbol{u}^\top\boldsymbol{v}}{3N}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2\leqslant&F(\bar{\boldsymbol{x}}^k)-F(\bar{\boldsymbol{x}}^{k+1})\\ &+(\frac{6L\kappa_{uv}\|\boldsymbol{v}\|}{N}+\frac{4\kappa_{uv}^2\|\boldsymbol{v}\|^2}{N\boldsymbol{u}^\top\boldsymbol{v}})\bar{\alpha}_k\|\boldsymbol{e}_{y,k}\|_C^2\\ &+(\frac{6L\kappa_{uv}\|\boldsymbol{v}\|^3}{N^2}+\frac{4\kappa_v^2\|\boldsymbol{v}\|^4}{N^2\boldsymbol{u}^\top\boldsymbol{v}})\times\\ &\left(AL_0+BL_1(b+\ell G)\right)^2\bar{\alpha}_k\|\boldsymbol{e}_{x,k}\|_R^2.\end{aligned} \tag{43}$$

证明. 见附录 G。 □

引理 13 为平均优化变量的梯度范数建立了一个下降不等式。结合引理 12 中推导出的递归误差界，该结果使我们能够刻画随时间累积的共识误差和梯度跟踪误差。现在，我们可以证明算法 1 的收敛性。

C. 收敛结果

引理 14. 假设引理 12 和引理 13 中的条件成立。若步长还满足

$$0<\alpha \leqslant \min\left\{\sqrt{\tfrac{(1-\sigma_R^2)(1-\sigma_C^2)}{8\mathcal{C}_{x,1}\mathcal{C}_{y,1}}},\ \sqrt{\tfrac{N(\boldsymbol{u}^\top\boldsymbol{v})^2(1-\sigma_R^2)}{\mathcal{C}_{x,2}\|\boldsymbol{v}\|^3(144L\kappa_{uv}\boldsymbol{u}^\top\boldsymbol{v}+96\kappa_{uv}^2\|\boldsymbol{v}\|)(AL_0+BL_1(b+\ell G))^2}}\right\}, \tag{44}$$

那么，对于所有 $K\geqslant 0$，以下结果成立：

$$\begin{aligned}
&\tfrac{1}{K}\sum_{k=0}^{K}\|\boldsymbol{e}_{x,k}\|_R^2\leqslant\mathcal{O}\Big(\tfrac{N}{K(1-\sigma_R^2)}\Big),\\
&\tfrac{1}{K}\sum_{k=0}^{K}\|\boldsymbol{e}_{y,k}\|_C^2\leqslant\mathcal{O}\Big(\tfrac{1}{K(1-\sigma_R^2)(1-\sigma_C^2)}\Big),\\
&\tfrac{1}{K}\sum_{k=0}^{K-1}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2\leqslant\mathcal{O}\big(\tfrac{N}{K\boldsymbol{u}^\top\boldsymbol{v}}\big).
\end{aligned} \tag{45}$$

证明。由引理 12，$\|\boldsymbol{e}_{x,\ k+1}\|_R^2$ 和 $\|\boldsymbol{e}_{y,\ k+1}\|_C^2$ 满足以下不等式系统：

$$\begin{aligned}
\begin{bmatrix}\|\boldsymbol{e}_{x,k+1}\|_R^2\\ \|\boldsymbol{e}_{y,k+1}\|_C^2\end{bmatrix}\leqslant&\begin{bmatrix}\frac{1+\sigma_R^2}{2} & \alpha^2\mathcal{C}_{x,1}\\ \mathcal{C}_{y,1} & \frac{1+\sigma_C^2}{2}\end{bmatrix}\begin{bmatrix}\|\boldsymbol{e}_{x,k}\|_R^2\\ \|\boldsymbol{e}_{y,k}\|_C^2\end{bmatrix}\\
&+\begin{bmatrix}\alpha\mathcal{C}_{x,2}\\ \alpha\mathcal{C}_{y,2}\end{bmatrix}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2.
\end{aligned} \tag{46}$$

定义堆叠误差向量为

$$\boldsymbol{u}_k=\begin{bmatrix}\|\boldsymbol{e}_{x,k}\|_R^2\\ \|\boldsymbol{e}_{y,k}\|_C^2\end{bmatrix},$$

系统矩阵为

$$\boldsymbol{G}=\begin{bmatrix}\frac{1+\sigma_R^2}{2} & \alpha^2\mathcal{C}_{x,1}\\ \mathcal{C}_{y,1} & \frac{1+\sigma_C^2}{2}\end{bmatrix},$$

输入项为

$$\boldsymbol{b}_k=\begin{bmatrix}\alpha\mathcal{C}_{x,2}\\ \alpha\mathcal{C}_{y,2}\end{bmatrix}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2,$$

则（46）中的不等式可以简洁地写成

$$\boldsymbol{u}_{k+1}\leqslant\boldsymbol{G}\boldsymbol{u}_k+\boldsymbol{b}_k. \tag{47}$$

在步长条件 $\alpha\leqslant\sqrt{\frac{(1-\sigma_R^2)(1-\sigma_C^2)}{8\mathcal{C}_{x,1}\mathcal{C}_{y,1}}}$ 下，矩阵（$I_2-\boldsymbol{G}$）是可逆的，因为 $|I_2-G|\geqslant(1-\sigma_R^2)(1-\sigma_C^2)8$。此外，基于矩阵求逆，我们有以下关系：

$$(I_2-\boldsymbol{G})^{-1}\leqslant\begin{bmatrix}\frac{4}{1-\sigma_R^2} & \frac{8\alpha^2\mathcal{C}_{x,1}}{(1-\sigma_R^2)(1-\sigma_C^2)}\\ \frac{8\mathcal{C}_{y,1}}{(1-\sigma_R^2)(1-\sigma_C^2)} & \frac{4}{1-\sigma_C^2}\end{bmatrix}.$$

因此，从 $k=0$ 到 K 递归应用式 (47)，可得

$$\sum_{k=0}^{K}\boldsymbol{u}_k\leqslant(I_2-\boldsymbol{G})^{-1}\boldsymbol{u}_0+(I_2-\boldsymbol{G})^{-1}\sum_{k=0}^{K-1}\boldsymbol{b}_k. \tag{48}$$

根据式 (48)，我们进一步逐项计算一致性误差的上界：

$$\begin{aligned}
\sum_{k=0}^{K}\|\boldsymbol{e}_{x,k}\|_R^2\leqslant&\frac{8\alpha^2\mathcal{C}_{x,1}}{(1-\sigma_R^2)(1-\sigma_C^2)}\|\boldsymbol{e}_{y,0}\|_C^2\\
&+\left(\frac{4\alpha\mathcal{C}_{x,2}}{1-\sigma_R^2}+\frac{8\alpha^2\mathcal{C}_{x,1}}{(1-\sigma_R^2)(1-\sigma_C^2)}\right)\sum_{k=0}^{K-1}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2,
\end{aligned} \tag{49}$$

$$\begin{aligned}
\sum_{k=0}^{K}\|\boldsymbol{e}_{y,k}\|_C^2\leqslant&\frac{4}{1-\sigma_C^2}\|\boldsymbol{e}_{y,0}\|_C^2\\
&+\left[\frac{8\alpha\mathcal{C}_{x,2}\mathcal{C}_{y,1}}{(1-\sigma_R^2)(1-\sigma_C^2)}+\frac{4\alpha\mathcal{C}_{y,2}}{1-\sigma_C^2}\right]\sum_{k=0}^{K-1}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2.
\end{aligned} \tag{50}$$

Finally, combining (43), (49), and (50), we obtain the following results under $\alpha\leqslant\sqrt{\frac{\boldsymbol{u}^\top\boldsymbol{v}(1-\sigma_R^2)}{(\frac{144L\kappa_{uv}\|\boldsymbol{v}\|^3}{N}+\frac{96\kappa_{uv}^2\|\boldsymbol{v}\|^4}{N\boldsymbol{u}^\top\boldsymbol{v}})\mathcal{C}_{x,2}(AL_0+BL_1(b+\ell G))^2}}$:

$$\begin{aligned}
\sum_{k=0}^{K}\|\boldsymbol{e}_{x,k}\|_R^2\leqslant&\mathcal{R}_{x,1}\big(F(\bar{\boldsymbol{x}}^0)-\underline{f}\big)+\mathcal{R}_{x,2}\|\boldsymbol{e}_{y,0}\|_C^2,\\
\sum_{k=0}^{K}\|\boldsymbol{e}_{y,k}\|_C^2\leqslant&\mathcal{R}_{y,1}\big(F(\bar{\boldsymbol{x}}^0)-\underline{f}\big)+\mathcal{R}_{y,2}\|\boldsymbol{e}_{y,0}\|_C^2,\\
\sum_{k=0}^{K-1}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2\leqslant&\tfrac{6N}{\boldsymbol{u}^\top\boldsymbol{v}}\big(F(\bar{\boldsymbol{x}}^0)-\underline{f}\big)+\mathcal{R}_{\nabla,1}\|\boldsymbol{e}_{y,0}\|_C^2,
\end{aligned} \tag{51}$$

其中常数为

$$\begin{aligned}
\mathcal{R}_{\nabla,1}=&\tfrac{6\kappa_{uv}\|\boldsymbol{v}\|}{\boldsymbol{u}^\top\boldsymbol{v}}\Big[(6L+\tfrac{4\kappa_{uv}\|\boldsymbol{v}\|}{\boldsymbol{u}^\top\boldsymbol{v}})\tfrac{4\alpha}{1-\sigma_C^2}\\
&+(6L\|\boldsymbol{v}\|^2+\tfrac{4\kappa_{uv}\|\boldsymbol{v}\|^3}{\boldsymbol{u}^\top\boldsymbol{v}})\tfrac{8\alpha^5\mathcal{C}_{x,1}L^2}{(1-\sigma_R^2)(1-\sigma_C^2)}\Big]\\
=&\mathcal{O}(\tfrac{1}{N}),\\
\mathcal{R}_{x,1}=&\tfrac{24N\mathcal{C}_{x,2}}{\boldsymbol{u}^\top\boldsymbol{v}(1-\sigma_R^2)}\alpha=\mathcal{O}\Big(\tfrac{N}{1-\sigma_R^2}\Big),\\
\mathcal{R}_{x,2}=&\tfrac{4\mathcal{R}_{\nabla,1}\mathcal{C}_{x,2}}{1-\sigma_R^2}\alpha+\tfrac{8\mathcal{C}_{x,1}}{(1-\sigma_R^2)(1-\sigma_C^2)}\alpha^2=\mathcal{O}\big(\tfrac{1}{N^2}\big),\\
\mathcal{R}_{y,1}=&\tfrac{6N}{u^\top v}\Big(\tfrac{8C_{x,2}C_{y,1}}{(1-\sigma_R^2)(1-\sigma_C^2)}+\tfrac{4C_{y,2}}{1-\sigma_C^2}\Big)\alpha\\
=&\mathcal{O}\Big(\tfrac{1}{(1-\sigma_R^2)(1-\sigma_C^2)}\Big),\\
\mathcal{R}_{y,2}=&\mathcal{R}_{\nabla,1}\left(\tfrac{8\mathcal{C}_{x,2}\mathcal{C}_{y,1}}{(1-\sigma_R^2)(1-\sigma_C^2)}+\tfrac{4\mathcal{C}_{y,2}}{1-\sigma_C^2}\right)\alpha+\tfrac{4}{1-\sigma_C^2}\\
=&\mathcal{O}\Big(\tfrac{1}{1-\sigma_C^2}\Big).
\end{aligned}$$

□

现在我们可以给出主要的收敛结果，该结果保证收敛到一个 ϵ-驻点。

定理 1. 假设 1、2、3 和 4 成立。在裁剪阈值 $c_0=\frac{1}{\sqrt{K}}$ 下，对于任意 $\epsilon>0$，存在 $k^\star\in\{0,1,\ldots,K-1\}$，使得算法 1 生成的迭代序列满足：

1) $\|\nabla F(\bar{\boldsymbol{x}}^{k^\star})\|\leqslant\epsilon$,
2) $\max_{1\leqslant i\leqslant N}\|\boldsymbol{x}_i^{k^\star}-\bar{\boldsymbol{x}}^{k^\star}\|_R^2\leqslant\epsilon$,

之后

$$K=\mathcal{O}\left(\frac{1}{\alpha^2\epsilon^2}\right)$$

迭代，其中步长 α 满足 $0<\alpha\leqslant$min $\{C_1,C_2,C_3,C_4,C_5\}$ with

$$C_1=\frac{(1-\sigma_R^2)\sqrt{N}}{6\sqrt{2}\|\boldsymbol{v}\|\kappa_v\|\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{u}^\top}{N}\|_R\delta_{R,2}(AL_0+BL_1b+BL_1\ell G)}, \tag{52}$$

$$C_2=\frac{1-\sigma_C^2}{12\sqrt{2}\kappa_v\delta_{C,2}\|\boldsymbol{I}-\frac{\boldsymbol{v}\boldsymbol{1}^\top}{N}\|_C(AL_0+BL_1b+BL_1\ell G)}, \tag{53}$$

$$C_3=\sqrt{\frac{(1-\sigma_R^2)(1-\sigma_C^2)}{8\mathcal{C}_{x,1}\mathcal{C}_{y,1}}},\quad C_4=\frac{\boldsymbol{u}^\top\boldsymbol{v}}{9LN\|\boldsymbol{v}\|^2}, \tag{54}$$

$$C_5=\sqrt{\frac{N(\boldsymbol{u}^\top\boldsymbol{v})^2(1-\sigma_R^2)}{\mathcal{C}_{x,2}\|\boldsymbol{v}\|^3(144L\kappa_{uv}\boldsymbol{u}^\top\boldsymbol{v}+96\kappa_{uv}^2\|\boldsymbol{v}\|)(AL_0+BL_1(b+\ell G))^2}}, \tag{55}$$

其中常数$\mathcal{C}_{x,1}$, $\mathcal{C}_{x,2}$ 和 $\mathcal{C}_{y,1}$ 定义于 (40) 和 (41) 中。 (39),

证明。由引理 14 可得

$$\sum_{k=0}^{K-1}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2\leqslant\frac{6N}{\boldsymbol{u}^\top\boldsymbol{v}}\big(F(\bar{\boldsymbol{x}}^0)-\underline{f}\big)+\mathcal{R}_{\nabla,1}\|\boldsymbol{e}_{y,0}\|_C^2. \tag{56}$$

我们将步长 $\bar{\alpha}_k=\alpha\min\left\{1,\ \frac{c_0}{\|\boldsymbol{v}\|\|\nabla F(\bar{\boldsymbol{x}}^k)\|}\right\}$ 代入 (56), 并根据梯度幅值将迭代分为两组:

$$\mathcal{S}=\left\{0\leqslant k\leqslant K-1\,\middle|\,\|\boldsymbol{v}\|\|\nabla F(\bar{\boldsymbol{x}}^k)\|<c_0\right\},$$

and

$$\mathcal{S}^C=\left\{0\leqslant k\leqslant K-1\,\middle|\,\|\boldsymbol{v}\|\|\nabla F(\bar{\boldsymbol{x}}^k)\|\geqslant c_0\right\}.$$

因此，不等式 (56) 可改写为

$$\begin{aligned}\sum_{k\in\mathcal{S}}\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2&\leqslant\mathcal{O}\left(\frac{1}{\alpha}\right),\\ \sum_{k\in\mathcal{S}^C}\|\nabla F(\bar{\boldsymbol{x}}^k)\|&\leqslant\mathcal{O}\left(\frac{1}{\alpha c_0}\right).\end{aligned} \tag{57}$$

接下来，利用柯西–施瓦茨不等式，可得

$$\left(\sum_{k\in\mathcal{S}}\|\nabla F(\bar{\boldsymbol{x}}^k)\|\right)^2\leqslant|\mathcal{S}|\sum_{k\in\mathcal{S}}\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2,$$

这进一步根据以下内容得出不等式 (57)

$$\sum_{k\in\mathcal{S}}\|\nabla F(\bar{\boldsymbol{x}}^k)\|\leqslant\mathcal{O}\left(\sqrt{\frac{|\mathcal{S}|}{\alpha}}\right). \tag{58}$$

结合（57）和（58），我们可以将所有迭代中的平均梯度范数界定为

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla F(\bar{\boldsymbol{x}}^k)\|&\leqslant\frac{1}{K}\left(\sum_{k\in\mathcal{S}}\|\nabla F(\bar{\boldsymbol{x}}^k)\|+\sum_{k\in\mathcal{S}^C}\|\nabla F(\bar{\boldsymbol{x}}^k)\|\right)\\&\leqslant\mathcal{O}\left(\frac{\sqrt{\frac{|\mathcal{S}|}{\alpha}}+\frac{1}{\alpha c_0}}{K}\right).\end{aligned} \tag{59}$$

由于 $|\mathcal{S}|\leqslant K$ 和 $c_0=\frac{1}{\sqrt{K}}$，不等式（59）可简化为

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla F(\bar{\boldsymbol{x}}^k)\|\leqslant\mathcal{O}\left(\frac{\sqrt{\frac{1}{\alpha}}+\frac{1}{\alpha}}{\sqrt{K}}\right). \tag{60}$$

此外，由引理 14 可知

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\boldsymbol{e}_{x,k}\|_R^2\leqslant\mathcal{O}\left(\frac{1}{K}\right). \tag{61}$$

结合 (60) 和 (61)，我们得到

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(\|\nabla F(\bar{\boldsymbol{x}}^k)\|+\|\boldsymbol{e}_{x,k}\|_R^2\right)\leqslant\mathcal{O}\left(\frac{1}{\sqrt{\alpha K}}+\frac{1}{K}\right). \tag{62}$$

因此，对于充分大的 $K=\mathcal{O}\left(\frac{1}{\alpha^2\epsilon^2}\right)$，不等式 (62) 意味着

$$\min_{0\leqslant k\leqslant K-1}\left(\|\nabla F(\bar{\boldsymbol{x}}^k)\|+\|\boldsymbol{e}_{x,k}\|_R^2\right)\leqslant\epsilon. \tag{63}$$

由于 (63) 左侧的两项均非负，存在一次迭代 $k^\star\in\{0,1,\ldots,K-1\}$ 使得

$$\|\nabla F(\bar{\boldsymbol{x}}^{k^\star})\|\leqslant\epsilon\quad\text{and}\quad\|\boldsymbol{e}_{x,k^\star}\|_R^2\leqslant\epsilon. \tag{64}$$

(64) 中的第二个不等式结合定义$\|\boldsymbol{e}_{x,k}\|_R^2=\sum_{i=1}^N\|\boldsymbol{x}_i^k-\bar{\boldsymbol{x}}^k\|_R^2$，可得

$$\max_{1\leqslant i\leqslant N}\|\boldsymbol{x}_i^{k^\star}-\bar{\boldsymbol{x}}^{k^\star}\|_R^2\leqslant\epsilon. \tag{65}$$

□

注 1. 定理 1 表明，算法 1 在 $\mathcal{O}(1/\epsilon^2)$ 次迭代内可达到 $\min_{0\leqslant k\leqslant K-1}\|\nabla F(\bar{\boldsymbol{x}}^k)\|\leqslant\epsilon$ 的优化误差，同时保持智能体之间的共识。这与 *[13]* 中关于 (L_0,L_1)-光滑性的集中式优化已有结果一致。

V. 数值实验

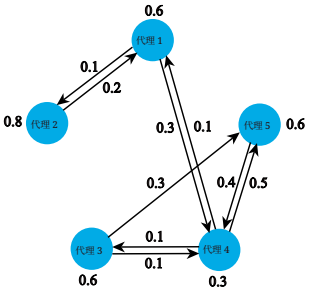
本节通过在基准数据集上使用正则化逻辑回归和卷积神经网络（CNN）进行实验，评估所提算法的有效性。两项实验分别采用图 1a 和图 1b 所示的通信矩阵 $\boldsymbol{R}$ 和 $\boldsymbol{C}$ 。

A. 正则化逻辑回归

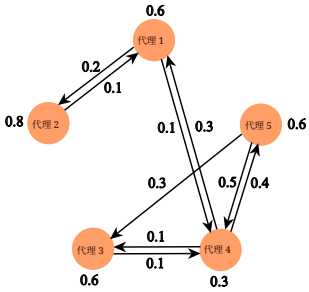
在本实验中，我们采用非凸正则化逻辑回归，使用来自 LIBSVM [53], 的真实世界数据集，具体而言是 a9a数据集，来解决二分类问题。训练样本的特征向量记为 $\boldsymbol{h}\in\mathbb{R}^d$，其中 $d=123$，类别标签为 $y\in\{0,1\}$。

损失函数定义为

$$\begin{aligned}&f_i(\boldsymbol{x}_i;\{\boldsymbol{h},y\})\\&=-y\log(\frac{1}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{h})})+(1-y)\log(\frac{\exp(\boldsymbol{x}_i^\top\boldsymbol{h})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{h})})\\&\quad+\lambda_i\|\boldsymbol{x}_i\|^{p_i},\end{aligned} \tag{66}$$



(a) 行随机混合 matrix $\boldsymbol{R}$。



(b) 列随机混合矩阵 $\boldsymbol{C}$。

图 1：评估中使用的有向通信图。

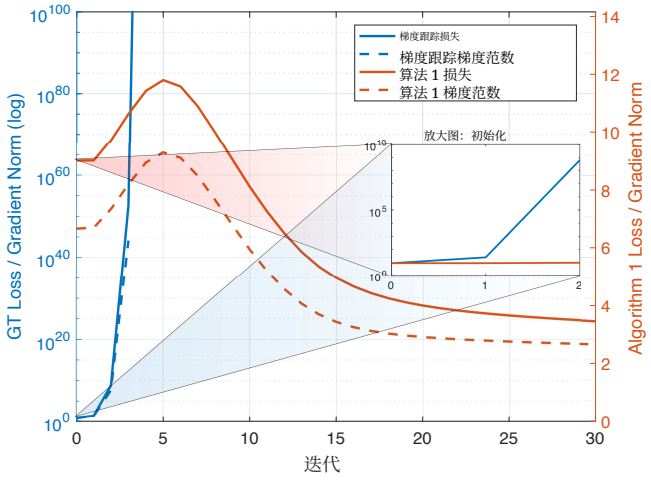


图 2： [54] 中算法 1 与梯度跟踪算法在 a9a数据集上的损失和梯度范数比较。 [54] 中的标准梯度跟踪（GT）方法（蓝色曲线，对应左轴）在初始迭代期间表现出严重的不稳定性，损失和梯度范数均迅速增大。相比之下，算法 1 （红色曲线，对应右轴）确保损失值和梯度范数均平稳下降。放大子图显示，两种算法从相同的初始化开始。

其中， $\{\boldsymbol{h}, y\}$ 表示训练元组， λ_i 表示智能体 i 的正则化系数。

在实验中，为了体现智能体之间本地数据分布和模型偏好的异质性，我们为五个智能体指定以下数值： $\lambda_1 = 5 \times 10^{-4}$， $\lambda_2 = 1 \times 10^{-3}$， $\lambda_3 = 2 \times 10^{-3}$， $\lambda_4 = 1 \times 10^{-3}$， $\lambda_5 = 1 \times 10^{-3}$， $p_1 = 4$， $p_2 = 5$， $p_3 = 6$， $p_4 = 5$， $p_5 = 4$.正则化项 $\|\boldsymbol{x}_i\|^{p_i}$ 使损失函数满足（L_0， L_1）-光滑性条件，但不满足传统光滑性条件，详见 [18]。

我们将所提出的算法 1 与标准梯度跟踪方法 [54] 以及带裁剪的去中心化梯度下降（DGD） [29]进行性能比较。在所有算法中，批量大小设为 32，步长固定为 $\alpha = 0.05$。对于基于裁剪的方法，裁剪阈值设为 $c_0 = 5$。

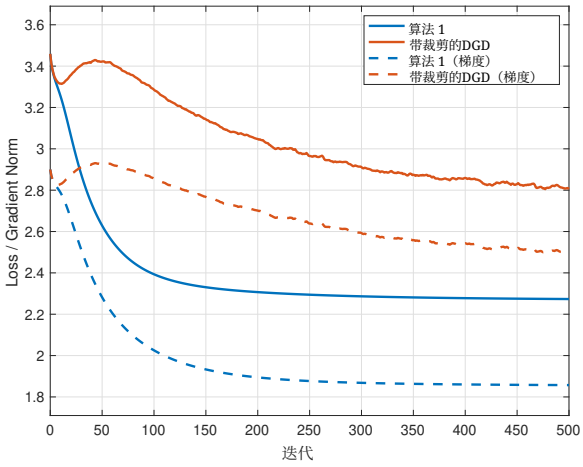


图 3：算法 1 与带梯度裁剪的DGD在 a9a数据集上的损失和梯度范数变化对比 [29] 。

图 2 展示了标准梯度跟踪算法 [54]和所提出的算法 1 的损失函数与梯度范数的变化情况。标准梯度跟踪方法（蓝色曲线）在最初几次迭代中表现出严重的不稳定性。这种不稳定性源于 (L_0, L_1)-光滑性所导致的较大梯度幅值。相比之下，算法 1（红色曲线）在整个训练过程中保持稳定。裁剪机制在最初几次迭代中有效控制了梯度更新的幅值，避免了标准梯度跟踪算法 [54]中出现的梯度爆炸。

图 3 展示了所提出的算法 1 与基于 DGD 并采用梯度裁剪的算法 [29],在损失和梯度范数方面的变化情况。显然，我们提出的算法能够快速且稳定地收敛，而采用梯度裁剪的 DGD 则表现出明显的振荡，且收敛速度显著较慢。这凸显了我们算法设计的优势，并证实了前文讨论的问题：当不同智能体的目标函数存在异质性时，直接裁剪局部梯度可能会引发问题。

B. 卷积神经网络

在该实验中，我们研究卷积神经网络（CNN）的训练，用于对 CIFAR-10 数据集 [55], 进行分类。该数据集包含 10 个不同类别的 50,000 张训练图像。我们将 CIFAR-10 数据集平均分配给五个智能体，并将批量大小设为 32。作为基线的 CNN 架构是深度网络 ResNet-18，其训练是一个高度非凸且非Lipschitz连续的问题。

在实验中，我们使用所提出的算法 1 训练 CNN，并将其性能与几种有代表性的分布式优化算法进行比较，包括标准梯度跟踪（DGT） [54]，CDSGD [56],带Polyak动量的 CDSGD（CDSGD-P） [56]，CDSGD

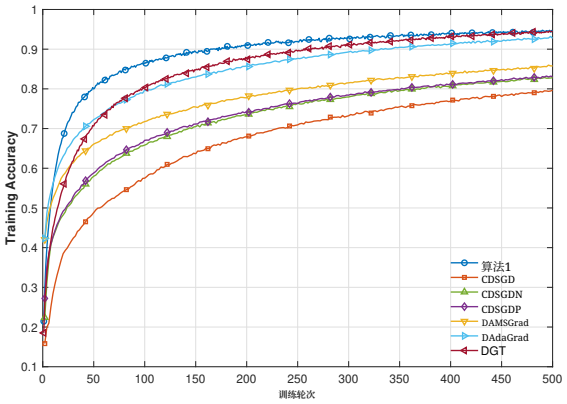


图4：在 CIFAR-10 数据集上，所提算法与最先进方法的训练准确率比较。

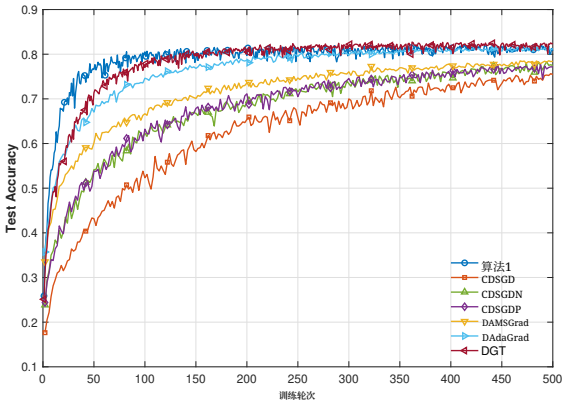


图5：所提算法与最先进方法在 CIFAR-10数据集上的测试准确率比较。

带 Nesterov动量的 CDSGD（CDSGD-N） [57], DAMSGrad [58],和 DAdaGrad [58]。对于所提出的算法，步长和裁剪阈值分别设为 $\alpha = 0.05$ 和 $c_0 = 10$。对于所有基线算法，均采用能够确保收敛的最大步长，以便进行公平比较。

训练准确率和测试准确率的变化分别如图4和图5所示。如图4所示，与现有最先进的分布式优化方法相比，所提算法收敛速度更快，且训练准确率更高。此外，图5表明，所提算法的测试准确率始终优于相应算法，凸显了其强大的泛化能力。这些结果证实了所提方法在非凸和广义光滑性问题上的去中心化深度学习中的有效性。

VI. 结论

本文提出了一种新的分布式优化算法，可在有向通信图和不一定满足传统光滑性条件的 (L_0, L_1)-光滑目标函数下确保精确收敛。不同于依赖有界梯度差异性的现有 (L_0, L_1)-光滑性研究结果，我们的方法即使在梯度差异性无界时也能确保精确收敛。一项关键创新是对全局梯度的局部估计进行裁剪，而不是直接裁剪局部梯度。然而，这会引入显著的非线性和复杂性，使传统分析技术无法适用。为此，我们建立了一个新的理论框架，提供了严格的收敛保证。事实上，我们的分析表明，该算法达到 $\mathcal{O}(\epsilon^{-2})$ 收敛速度，与相同光滑性条件下集中式方法的现有结果相匹配。在真实世界数据集上进行的数值实验进一步证实了所提方法的有效性。

附录A：引理3的证明

根据假设 2，我们有

$$\begin{aligned}\left\|\nabla^2 F(\boldsymbol{\theta})\right\| &= \frac{1}{N}\left\|\sum_{i=1}^{N}\nabla^2 f_i(\boldsymbol{\theta})\right\| \\ &\leqslant \frac{1}{N}\sum_{i=1}^{N}\left(L_0^i + L_1^i\left\|\nabla f_i(\boldsymbol{\theta})\right\|\right).\end{aligned} \tag{67}$$

接下来，利用假设 3，我们得到

$$\begin{aligned}\left\|\nabla^2 F(\boldsymbol{\theta})\right\| \leqslant & \frac{1}{N}\sum_{i=1}^{N}\left(L_0^i + L_1^i\left(\ell\left\|\nabla F(\boldsymbol{\theta})\right\| + b\right)\right) \\ = & \frac{1}{N}\sum_{i=1}^{N}\left(L_0^i + bL_1^i\right) \\ & + \left(\frac{\ell}{N}\sum_{i=1}^{N}L_1^i\right)\left\|\nabla F(\boldsymbol{\theta})\right\|.\end{aligned} \tag{68}$$

因此，全局目标函数 $F(\boldsymbol{\theta})$ 也是(L_0, L_1)-光滑的，其中

$$L_0 = \frac{1}{N}\sum_{i=1}^{N}\left(L_0^i + bL_1^i\right), \qquad L_1 = \frac{\ell}{N}\sum_{i=1}^{N}L_1^i.$$

附录 B：一些有用引理

引理 15（[14], 引理 3.5）。如果 f 是(L_0, L_1)-光滑的，那么对于任意 $\boldsymbol{x} \in \mathbb{R}^d$，我们有

$$\|\nabla f(\boldsymbol{x})\|^2 \leqslant 2\left(L_0 + 2L_1\|\nabla f(\boldsymbol{x})\|\right)\left(f(\boldsymbol{x}) - \underline{f}\right). \tag{69}$$

引理 16（[14], 推论 3.6）。假设 f 是（L_0, L_1）-光滑的。如果对于某个 $\boldsymbol{x} \in \mathcal{X}$，我们有 $f(\boldsymbol{x}) - \underline{f} \leqslant \Delta_f$且$\Delta_f \geqslant 0$，那么我们有

$$G^2 = 2\left(L_0 + 2L_1\|\nabla f(\boldsymbol{x})\|\right)\Delta_f$$

and

$$\|\nabla f(\boldsymbol{x})\| \leqslant G < \infty,$$

其中

$$G = \sup\left\{u \geqslant 0 \;\middle|\; u^2 \leqslant 2\left(L_0 + 2L_1\|\nabla f(\boldsymbol{x})\|\right)\Delta_f\right\}.$$

附录 C：引理 8 的证明

根据 α_i^k 和 $\bar{\alpha}_i^k$ 的定义，我们逐一讨论并建立方程（20）中的关系如下：

情形 1： $v_i\|\nabla F(\bar{\boldsymbol{x}}^k)\| \leqslant c_0$, $\|\boldsymbol{y}_i^k\| \leqslant c_0$:
在这种情形下，我们有 $\bar{\alpha}_i^k = \alpha_i^k = \alpha$，这意味着

$$\left|\alpha_i^k - \bar{\alpha}_i^k\right| \|\boldsymbol{y}_i^k\| = 0.$$

情形 2： $v_i\|\nabla F(\bar{\boldsymbol{x}}^k)\| \leqslant c_0$, $\|\boldsymbol{y}_i^k\| > c_0$:
在这种情形下，我们有 $\alpha_i^k < \alpha = \bar{\alpha}_i^k$，这意味着

$$\begin{aligned}\left|\alpha_i^k - \bar{\alpha}_i^k\right| \|\boldsymbol{y}_i^k\| &= \alpha c_0 \left|\tfrac{1}{\|\boldsymbol{y}_i^k\|} - \tfrac{1}{c_0}\right| \|\boldsymbol{y}_i^k\| \\ &= \alpha c_0 \left(\tfrac{1}{c_0} - \tfrac{1}{\|\boldsymbol{y}_i^k\|}\right) \|\boldsymbol{y}_i^k\| \\ &= \alpha \left(1 - \tfrac{c_0}{\|\boldsymbol{y}_i^k\|}\right) \|\boldsymbol{y}_i^k\|.\end{aligned} \tag{70}$$

在上面的第二个等式中，我们在这种情形下使用了条件 $\|\boldsymbol{y}_i^k\| > c_0$。

将 α 代入 $\bar{\alpha}_i^k$ 并利用关系式$v_i\|\nabla F(\bar{\boldsymbol{x}}^k)\| \leqslant c_0$，我们得到

$$\begin{aligned}\left|\alpha_i^k - \bar{\alpha}_i^k\right| \|\boldsymbol{y}_i^k\| \leqslant& \bar{\alpha}_i^k \left(\|\boldsymbol{y}_i^k\| - \|v_i \nabla F(\bar{\boldsymbol{x}}^k)\|\right) \\ \leqslant& \bar{\alpha}_i^k \|\boldsymbol{y}_i^k - v_i \nabla F(\bar{\boldsymbol{x}}^k)\|,\end{aligned}$$

其中，最后一个不等式由 $|\|a\| - \|b\|| \leqslant \|a-b\|$推出。

情形 3： $v_i\|\nabla F(\bar{\boldsymbol{x}}^k)\| > c_0$, $\left\|\boldsymbol{y}_i^k\right\| \leqslant c_0$:
在此情形下，我们有 $\alpha_i^k = \alpha > \bar{\alpha}_i^k$，这意味着

$$\begin{aligned}\left|\alpha_i^k - \bar{\alpha}_i^k\right| \left\|\boldsymbol{y}_i^k\right\| &= \alpha c_0 \left|\frac{1}{c_0} - \frac{1}{v_i \left\|\nabla F\left(\bar{\boldsymbol{x}}^k\right)\right\|}\right| \left\|\boldsymbol{y}_i^k\right\| \\ &= \alpha \left(1 - \frac{c_0}{v_i \left\|\nabla F\left(\bar{\boldsymbol{x}}^k\right)\right\|}\right) \left\|\boldsymbol{y}_i^k\right\| \\ &= \frac{\alpha \left\|\boldsymbol{y}_i^k\right\|}{v_i \left\|\nabla F\left(\bar{\boldsymbol{x}}^k\right)\right\|} \left(\left\|v_i \nabla F\left(\bar{\boldsymbol{x}}^k\right)\right\| - c_0\right),\end{aligned} \tag{71}$$

其中，在第二个等式中，我们使用了此情形下的条件 $v_i\|\nabla F(\bar{\boldsymbol{x}}^k)\| > c_0$。利用 $\bar{\alpha}_i^k$的定义和三角不等式 $|\|a\| - \|b\|| \leqslant \|a-b\|$，我们得到以下不等式：

$$\begin{aligned}\left|\alpha_i^k - \bar{\alpha}_i^k\right| \left\|\boldsymbol{y}_i^k\right\| \leqslant& \frac{\alpha c_0}{v_i \left\|\nabla F\left(\bar{\boldsymbol{x}}^k\right)\right\|} \left(v_i \left\|\nabla F\left(\bar{\boldsymbol{x}}^k\right)\right\| - \left\|\boldsymbol{y}_i^k\right\|\right) \\ \leqslant& \bar{\alpha}_i^k \left\|\boldsymbol{y}_i^k - v_i \nabla F\left(\bar{\boldsymbol{x}}^k\right)\right\|.\end{aligned} \tag{72}$$

情形 4： $v_i\|\nabla F(\bar{\boldsymbol{x}}^k)\| > c_0$, $\|\boldsymbol{y}_i^k\| > c_0$:
在这种情况下，我们有 $\alpha_i^k < \alpha$、 $\bar{\alpha}_i^k < \alpha$，从而得到

$$\begin{aligned}\left|\alpha_i^k - \bar{\alpha}_i^k\right| \|\boldsymbol{y}_i^k\| =& \alpha c_0 \left|\tfrac{1}{\|\boldsymbol{y}_i^k\|} - \tfrac{1}{v_i\|\nabla F(\bar{\boldsymbol{x}}^k)\|}\right| \|\boldsymbol{y}_i^k\| \\ =& \alpha c_0 \frac{\left|v_i\|\nabla F(\bar{\boldsymbol{x}}^k)\| - \|\boldsymbol{y}_i^k\|\right|}{v_i\|\nabla F(\bar{\boldsymbol{x}}^k)\|}.\end{aligned}$$

利用 $\bar{\alpha}_i^k = \frac{\alpha c_0}{v_i\|\nabla F\ \bar{\boldsymbol{x}}^k\|}$ $|\|a\| - _{()}$和三角不等式 $\|b\|| \leqslant \|a-b\|$，我们得到

$$\left|\alpha_i^k - \bar{\alpha}_i^k\right| \|\boldsymbol{y}_i^k\| \leqslant \bar{\alpha}_i^k \|\boldsymbol{y}_i^k - v_i \nabla F\left(\bar{\boldsymbol{x}}^k\right)\|.$$

因此，方程（20）中的关系在所有情况下都成立，由此完成证明。

附录 D：引理 9 的证明

根据（16）中优化误差 $\boldsymbol{e}_{x,\ k}$的推导，对（16）两边取范数并利用不等式 $(a+b)^{\ 2} \leqslant\ (1+\eta)\ a^2 +\ (1+\frac{1}{\eta})\ b^2$，可得

$$\begin{aligned}\|\boldsymbol{e}_{x,k+1}\|_R^2 \leqslant& (1+\eta)\|\big(\boldsymbol{R} - \tfrac{\mathbf{1}\boldsymbol{u}^\top}{N}\big)\boldsymbol{e}_{x,k}\|_R^2 \\ &+ (1+\tfrac{1}{\eta})\|\big(\boldsymbol{I} - \tfrac{\mathbf{1}\boldsymbol{u}^\top}{N}\big)\boldsymbol{\alpha}_k \boldsymbol{y}^k\|_R^2, \\ \leqslant& (1+\eta)\sigma_R^2 \delta_{R,2}^2 \|\boldsymbol{e}_{x,k}\|^2 \\ &+ \Big(1+\tfrac{1}{\eta}\Big)\|\boldsymbol{I} - \tfrac{\mathbf{1}\boldsymbol{u}^\top}{N}\|_R^2 \delta_{R,2}^2 \|\boldsymbol{\alpha}_k \boldsymbol{y}^k\|^2.\end{aligned} \tag{73}$$

根据定义 $\boldsymbol{\alpha}^k\boldsymbol{y}^k = [(\alpha_1^k \boldsymbol{y}_1^k)^\top;\ \cdots;\ (\alpha_N^k \boldsymbol{y}_N^k)^\top] \in \mathbb{R}^{N\times d}$，我们得到

$$\begin{aligned}\|\boldsymbol{\alpha}_k \boldsymbol{y}^k\|^2 &= \sum_{i=1}^N (\alpha_i^k)^2 \|\boldsymbol{y}_i^k\|^2 \\ &= \sum_{i=1}^N \alpha^2 \min\left\{1, \tfrac{c_0^2}{\|\boldsymbol{y}_i^k\|^2}\right\} \|\boldsymbol{y}_i^k\|^2 \\ &\leqslant N\alpha^2 c_0^2.\end{aligned} \tag{74}$$

结合 (73) 和 (74)，我们有

$$\begin{aligned}&\|\boldsymbol{e}_{x,k+1}\|_R^2 \\ =&(1+\eta)\sigma_R^2 \delta_{R,2}^2 \|\boldsymbol{e}_{x,k}\|^2 + (1+\tfrac{1}{\eta})\|\boldsymbol{I} - \tfrac{\mathbf{1}\boldsymbol{u}^\top}{N}\|_R^2 N\alpha^2 c_0^2 \delta_{R,2}^2 \\ \leqslant&\tfrac{1+\sigma_R^2}{2}\|\boldsymbol{e}_{x,k}\|_R^2 + \tfrac{\sigma_R^2(1+\sigma_R^2)}{1-\sigma_R^2}\|\boldsymbol{I} - \tfrac{\mathbf{1}\boldsymbol{u}^\top}{N}\|_R^2 N\alpha^2 c_0^2 \delta_{R,2}^2,\end{aligned} \tag{75}$$

其中，最后一个不等式使用了 $\eta = \frac{1-\sigma_R^2}{2\sigma_R^2}$。

式（75）中的不等式是如下形式的线性递推

$$z_{k+1} \leqslant a' z_k + b',$$

其中

$$z_k = \|\boldsymbol{e}_{x,k}\|_R^2, \qquad a' = \tfrac{1+\sigma_R^2}{2}, \qquad b' = \tfrac{\sigma_R^2(1+\sigma_R^2)}{1-\sigma_R^2} N\alpha^2 c_0^2 \delta_{R,2}^2.$$

对式（75）进行迭代，可得

$$\|\boldsymbol{e}_{x,k+1}\|_R^2 \leqslant a'^{k+1}\|\boldsymbol{e}_{x,0}\|_R^2 + \frac{b'}{1-a'}\big(1 - a'^{k+1}\big). \tag{76}$$

特别地，如果 $a' < 1$，一致性误差是一致有界的，即

$$\|\boldsymbol{e}_{x,k}\|_R^2 \ \leqslant\ \frac{2N\sigma_R^2(1+\sigma_R^2)\delta_{R,2}^2 \|\boldsymbol{I} - \frac{\mathbf{1}\boldsymbol{u}^\top}{N}\|_R^2}{(1-\sigma_R^2)^2}\alpha^2 c_0^2, \quad \forall k \geqslant 0. \tag{77}$$

附录 E：引理 10 的证明

根据 (19)，对不等式两边取范数，可得

$$\begin{aligned}
&\|\boldsymbol{e}_{y,k+1}\|_C^2\\
\leqslant&(1+\eta)\|\big(\boldsymbol{C}-\tfrac{\boldsymbol{v}\mathbf{1}^\top}{N}\big)\boldsymbol{e}_{y,k}\|_C^2\\
&+(1+\tfrac{1}{\eta})\|\big(\boldsymbol{I}-\tfrac{\boldsymbol{v}\mathbf{1}^\top}{N}\big)\big(\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\boldsymbol{x}^k)\big)\|_C^2\\
\leqslant&(1+\eta)\sigma_C^2\|\boldsymbol{e}_{y,k}\|_C^2\\
&+(1+\tfrac{1}{\eta})\left\|\boldsymbol{I}-\tfrac{\boldsymbol{v}\mathbf{1}^\top}{N}\right\|_C^2\|\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\boldsymbol{x}^k)\|_C^2\\
\leqslant&\tfrac{1+\sigma_C^2}{2}\|\boldsymbol{e}_{y,k}\|_C^2+\tfrac{1+\sigma_C^2}{1-\sigma_C^2}\delta_{C,2}^2\left\|\boldsymbol{I}-\tfrac{\boldsymbol{v}\mathbf{1}^\top}{N}\right\|_C^2\\
&\|\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\boldsymbol{x}^k)\|^2,
\end{aligned}\tag{78}$$

其中，最后一个不等式由 $\eta=\frac{1-\sigma_C^2}{2\sigma_C^2}$ 的选取可得。

对于第二项 $\|\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\boldsymbol{x}^k)\|^2$，我们可以将其分为以下两部分：

$$\begin{aligned}
&\|\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\boldsymbol{x}^k)\|^2\\
\leqslant&2\|\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\mathbf{1}\bar{\boldsymbol{x}}^k)\|^2+2\|\nabla f(\mathbf{1}\bar{\boldsymbol{x}}^k)-\nabla f(\boldsymbol{x}^k)\|^2.
\end{aligned}\tag{79}$$

为了利用引理 2 中 (L_0,L_1)-光滑函数的性质，我们首先分析项 $\|\boldsymbol{x}^{k+1}-\mathbf{1}\bar{\boldsymbol{x}}^k\|^2$. 。利用 (12)，可得

$$\begin{aligned}
&\|\boldsymbol{x}^{k+1}-\mathbf{1}\bar{\boldsymbol{x}}^k\|^2\\
=&\|\boldsymbol{R}\boldsymbol{x}^k+\boldsymbol{\alpha}_k\boldsymbol{y}^k-\mathbf{1}\bar{\boldsymbol{x}}^k\|^2\\
\leqslant&2\|\boldsymbol{R}\boldsymbol{x}^k-\mathbf{1}\bar{\boldsymbol{x}}^k\|_R^2+2\|\boldsymbol{\alpha}_k\boldsymbol{y}^k\|^2\\
\leqslant&2\left\|\boldsymbol{R}-\frac{\mathbf{1}u^\top}{N}\right\|_R^2\|\boldsymbol{x}^k-\mathbf{1}\bar{\boldsymbol{x}}^k\|_R^2+2\|\boldsymbol{\alpha}_k\boldsymbol{y}^k\|^2.
\end{aligned}\tag{80}$$

根据引理 6，我们可以得到 $\|\boldsymbol{x}^{k+1}-\mathbf{1}\bar{\boldsymbol{x}}^k\|^2$ 的上界如下：

$$\|\boldsymbol{x}^{k+1}-\mathbf{1}\bar{\boldsymbol{x}}^k\|^2\leqslant 2\sigma_R^2\|\boldsymbol{e}_{x,k}\|_R^2+2N\alpha^2c_0^2.\tag{81}$$

然后，根据引理 2，我们可以得到

$$\begin{aligned}
&\|\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\boldsymbol{x}^k)\|^2\\
\leqslant&2\left(AL_0+BL_1\|\nabla f_i(\bar{\boldsymbol{x}}^k)\|\right)^2\|\boldsymbol{x}^{k+1}-\mathbf{1}\bar{\boldsymbol{x}}^k\|^2\\
&+2\left(AL_0+BL_1\|\nabla f_i(\bar{\boldsymbol{x}}^k)\|\right)^2\|\boldsymbol{x}^k-\mathbf{1}\bar{\boldsymbol{x}}^k\|^2.
\end{aligned}\tag{82}$$

根据假设 3，将 $\|\nabla f_i(\bar{\boldsymbol{x}}^k)\|$ 替换为 $\|\nabla f(\bar{\boldsymbol{x}}^k)\|$，我们得到：

$$\begin{aligned}
&\|\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\boldsymbol{x}^k)\|^2\\
\leqslant&2\left(AL_0+BL_1b+BL_1\ell\|\nabla f(\bar{\boldsymbol{x}}^k)\|\right)^2\|\boldsymbol{x}^{k+1}-\mathbf{1}\bar{\boldsymbol{x}}^k\|^2\\
&+2\left(AL_0+BL_1b+BL_1\ell\|\nabla f(\bar{\boldsymbol{x}}^k)\|\right)^2\|\boldsymbol{x}^k-\mathbf{1}\bar{\boldsymbol{x}}^k\|^2\\
\leqslant&2\left(AL_0+BL_1b+BL_1\ell G\right)^2\\
&\Big(\|\boldsymbol{x}^{k+1}-\mathbf{1}\bar{\boldsymbol{x}}^k\|^2+\|\boldsymbol{x}^k-\mathbf{1}\bar{\boldsymbol{x}}^k\|^2\Big).
\end{aligned}\tag{83}$$

根据引理 9 和 (81)，我们得到

$$\begin{aligned}
&\|\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\boldsymbol{x}^k)\|^2\\
\leqslant&2\left(AL_0+BL_1b+BL_1\ell G\right)^2\big(2N+(1+2\sigma_R^2)\mathcal{C}_x\big)\alpha^2c_0^2\\
=&\mathcal{C}_1\alpha^2c_0^2,
\end{aligned}\tag{84}$$

其中 $\mathcal{C}_1=2(AL_0+BL_1b+BL_1\ell G)^2\big(2N+(1+2\sigma_R^2)\mathcal{C}_x\big)$。结合（78）和（84），我们得到梯度跟踪误差 $\|e_{y,\ k}\|_C^2$ 的递推关系如下：

$$\begin{aligned}
&\|\boldsymbol{e}_{y,k+1}\|_C^2\\
\leqslant&\frac{1+\sigma_C^2}{2}\|\boldsymbol{e}_{y,k}\|_C^2+\frac{1+\sigma_C^2}{1-\sigma_C^2}\delta_{C,2}^2\left\|\boldsymbol{I}-\tfrac{\boldsymbol{v}\mathbf{1}^\top}{N}\right\|_C^2\mathcal{C}_1\alpha^2c_0^2.
\end{aligned}\tag{85}$$

然后，我们可以得到梯度跟踪误差$\|\boldsymbol{e}_{y,\ k}\|_C^2$的一致有界：

$$\|\boldsymbol{e}_{y,k}\|_C^2\ \leqslant\ \frac{2(1+\sigma_C^2)}{(1-\sigma_C^2)^2}\,\delta_{C,2}^2\left\|\boldsymbol{I}-\tfrac{\boldsymbol{v}\mathbf{1}^\top}{N}\right\|_C^2\,\mathcal{C}_1\,\alpha^2c_0^2.\tag{86}$$

附录 F：引理 12 的证明

A. 预备结果

为证明引理 12，我们首先给出一个有用的预备结果。

引理 17. 在假设 2 和假设 4 下，利用引理 8，可以验证算法 1 的迭代满足

$$\begin{aligned}
&\left\|\boldsymbol{\alpha}_k\boldsymbol{y}^k\right\|^2\\
\leqslant&6\kappa_v^2\bar{\alpha}_k^2\left\|\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2+3N\bar{\alpha}_k^2\|\boldsymbol{v}\|^2\left\|\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2.
\end{aligned}\tag{87}$$

$$\begin{aligned}
&\left\|\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\\
\leqslant&2\|\boldsymbol{e}_{y,k}\|^2+\frac{2\|\boldsymbol{v}\|^2}{N}(AL_0+BL_1b+BL_1\ell G)^2\|\boldsymbol{e}_{x,k}\|^2.
\end{aligned}\tag{88}$$

证明。我们首先证明 (87)。根据定义，

$$\left\|\boldsymbol{\alpha}_k\boldsymbol{y}^k\right\|^2=\sum_{i=1}^N\left\|\alpha_i^k\boldsymbol{y}_i^k\right\|^2.$$

在上述等式右端的每一项中分别加上并减去 $\bar{\alpha}_k$ 和 $\bar{\alpha}_k\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)$，得到

$$\begin{aligned}
&\left\|\alpha_i^k\boldsymbol{y}_i^k\right\|^2\\
=&\left\|(\alpha_i^k-\bar{\alpha}_k)\boldsymbol{y}_i^k+\bar{\alpha}_k\big(\boldsymbol{y}_i^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\big)+\bar{\alpha}_k\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\\
\leqslant&3\left\|(\alpha_i^k-\bar{\alpha}_k)\boldsymbol{y}_i^k\right\|^2+3\left\|\bar{\alpha}_k(\boldsymbol{y}_i^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k))\right\|^2\\
&+3\left\|\bar{\alpha}_k\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2,
\end{aligned}$$

其中我们使用了不等式 $\|\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}\|^2\leqslant 3(\|\boldsymbol{a}\|^2+\|\boldsymbol{b}\|^2+\|\boldsymbol{c}\|^2)$。对 i 求和可得

$$\begin{aligned}
&\left\|\boldsymbol{\alpha}_k\boldsymbol{y}^k\right\|^2\\
\leqslant&3\sum_{i=1}^N\left\|(\alpha_i^k-\bar{\alpha}_k)\boldsymbol{y}_i^k\right\|^2+3\bar{\alpha}_k^2\sum_{i=1}^N\left\|\boldsymbol{y}_i^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\\
&+3N\bar{\alpha}_k^2\|\boldsymbol{v}\|^2\left\|\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2.
\end{aligned}$$

根据引理 8，我们有

$$\sum_{i=1}^{N}\left\|(\alpha_i^k-\bar{\alpha}_i^k)\boldsymbol{y}_i^k\right\|^2\leqslant\bar{\alpha}_k^2\sum_{i=1}^{N}\frac{\|\boldsymbol{v}\|}{v_i}\|\boldsymbol{y}_i^k-v_i\nabla F(\bar{\boldsymbol{x}}^k)\|^2.$$

前一个不等式中的界意味着

$$\begin{aligned}\left\|\boldsymbol{\alpha}_k\boldsymbol{y}^k\right\|^2\leqslant&6\bar{\alpha}_k^2\sum_{i=1}^{N}\frac{\|\boldsymbol{v}\|^2}{v_i^2}\left\|\boldsymbol{y}_i^k-v_i\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\\&+3N\bar{\alpha}_k^2\|\boldsymbol{v}\|^2\left\|\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\\\leqslant&6\kappa_v^2\bar{\alpha}_k^2\left\|\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\\&+3N\bar{\alpha}_k^2\|\boldsymbol{v}\|^2\left\|\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2,\end{aligned}\tag{89}$$

其中 $\kappa_v=\max\left\{\frac{\|\boldsymbol{v}\|}{v_i}\middle|i\in N\right\}$。

注意到 $\sum_{i=1}^{N}\|\boldsymbol{y}_i^k-v_i\nabla F(\bar{\boldsymbol{x}}^k)\|^2=\|\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\|^2$和$\bar{\alpha}_k\leqslant\alpha$，可得（87）。

现在证明（88）。首先，我们将 $\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)$ 分解为

$$\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)=\left(\boldsymbol{y}^k-\boldsymbol{v}\bar{\boldsymbol{y}}^k\right)+\boldsymbol{v}\left(\bar{\boldsymbol{y}}^k-\nabla F(\bar{\boldsymbol{x}}^k)\right).$$

应用不等式 $\|\boldsymbol{a}+\boldsymbol{b}\|^2\leqslant2\|\boldsymbol{a}\|^2+2\|\boldsymbol{b}\|^2$，得到

$$\begin{aligned}&\left\|\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\\\leqslant&2\left\|\boldsymbol{y}^k-\boldsymbol{v}\bar{\boldsymbol{y}}^k\right\|^2+2\|\boldsymbol{v}\|^2\left\|\bar{\boldsymbol{y}}^k-\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2.\end{aligned}\tag{90}$$

根据定义，$\boldsymbol{e}_{y,k}=\boldsymbol{y}^k-\boldsymbol{v}\bar{\boldsymbol{y}}^k$，因此（90）右端的第一项等于$2\|\boldsymbol{e}_{y,k}\|^2$。对于（90）右端的第二项，注意到 $\bar{\boldsymbol{y}}^k=\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_i^k)$，并写为

$$\left\|\bar{\boldsymbol{y}}^k-\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\leqslant\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(\boldsymbol{x}_i^k)-\nabla f_i(\bar{\boldsymbol{x}}^k)\right\|^2.$$

利用假设 2 和假设 3，结合引理 11，我们可以得到

$$\left\|\bar{\boldsymbol{y}}^k-\nabla F(\bar{\boldsymbol{x}}^k)\right\|^2\leqslant\frac{\left(AL_0+BL_1b+BL_1\ell G\right)^2}{N}\|\boldsymbol{e}_{x,k}\|^2.\tag{91}$$

将 (91) 代入 (90) 可得 (88)，证明完毕。 □

B. 引理 12 的证明

由 (73) 中的不等式可得

$$\begin{aligned}\|\boldsymbol{e}_{x,k+1}\|_R^2\leqslant&(1+\eta)\sigma_R^2\|\boldsymbol{e}_{x,k}\|_R^2\\&+(1+\tfrac{1}{\eta})\|\boldsymbol{I}-\tfrac{\boldsymbol{1}\boldsymbol{u}^\top}{N}\|_R^2\delta_{R,2}^2\|\boldsymbol{\alpha}_k\boldsymbol{y}^k\|^2.\end{aligned}\tag{92}$$

接下来，利用引理 17，在 $\alpha\leqslant\left(\sqrt{1-\sigma_R^2}\sqrt{N}\right)\cdot\left(\frac{1-\sigma_C^2}{6\sqrt{2}\|\boldsymbol{v}\|\kappa_v\|\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{u}^\top}{N}\|_R\delta_{R,2}AL_0+BL_1b+BL_1\ell G)}\right)$ 和 $\eta=\frac{1-\sigma_R^2}{3\sigma_R^2}$的条件下，我们可以进一步将 (92) 上界化为

$$\begin{aligned}\|\boldsymbol{e}_{x,k+1}\|_R^2\leqslant&\tfrac{1+\sigma_R^2}{2}\|\boldsymbol{e}_{x,k}\|_R^2\\&+\frac{12(1+2\sigma_R^2)\|\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{u}^\top}{N}\|_R^2\delta_{R,2}^2\kappa_v^2}{1-\sigma_R^2}\bar{\alpha}_k^2\|\boldsymbol{e}_{y,k}\|_C^2\\&+\frac{3N(1+2\sigma_R^2)\|\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{u}^\top}{N}\|_R^2\delta_{R,2}^2\|\boldsymbol{v}\|^2}{1-\sigma_R^2}\bar{\alpha}_k^2\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2.\end{aligned}\tag{93}$$

为方便记号，不等式 (93) 可简洁地写为

$$\begin{aligned}\|\boldsymbol{e}_{x,k+1}\|_R^2\leqslant&\tfrac{1+\sigma_R^2}{2}\|\boldsymbol{e}_{x,k}\|_R^2+\alpha^2\mathcal{C}_{x,1}\|\boldsymbol{e}_{y,k}\|_C^2\\&+\alpha\,\mathcal{C}_{x,2}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2,\end{aligned}\tag{94}$$

其中，常数 $\mathcal{C}_{x,1}$ 和 $\mathcal{C}_{x,2}$ 定义于 (39) 和 (40) 中。接下来，我们分析$\|\boldsymbol{e}_{y,k}\|_C^2$。

由 (78) 可得

$$\begin{aligned}&\|\boldsymbol{e}_{y,k+1}\|_C^2\\\leqslant&(1+\eta)\sigma_C^2\|\boldsymbol{e}_{y,k}\|_C^2\\&+(1+\tfrac{1}{\eta})\left\|\boldsymbol{I}-\tfrac{\boldsymbol{v}\boldsymbol{1}^\top}{N}\right\|_C^2\|\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\boldsymbol{x}^k)\|_C^2.\end{aligned}\tag{95}$$

对于（95）右端的第二项，由（84）可得

$$\begin{aligned}&\|\nabla f(\boldsymbol{x}^{k+1})-\nabla f(\boldsymbol{x}^k)\|^2\\\leqslant&2\left(AL_0+BL_1b+BL_1\ell G\right)^2\times\\&\left(\|\boldsymbol{x}^{k+1}-1\bar{\boldsymbol{x}}^k\|^2+\|\boldsymbol{x}^k-1\bar{\boldsymbol{x}}^k\|^2\right),\end{aligned}\tag{96}$$

结合（80）和引理 17，可得

$$\begin{aligned}&\|\boldsymbol{x}^{k+1}-1\bar{\boldsymbol{x}}^k\|^2\\\leqslant&2\left\|\boldsymbol{R}-\frac{1u^\top}{N}\right\|_R^2\|\boldsymbol{x}^k-1\bar{\boldsymbol{x}}^k\|_R^2+2\|\boldsymbol{\alpha}_k\boldsymbol{y}^k\|^2\\\leqslant&\left(2\sigma_R^2+\frac{24\bar{\alpha}_k^2\|\boldsymbol{v}\|^2\kappa_v^2\left(AL_0+BL_1b+BL_1\ell G\right)^2}{N}\right)\|\boldsymbol{e}_{x,k}\|_R^2\\&+24\kappa_v^2\bar{\alpha}_k^2\|\boldsymbol{e}_{y,k}\|_C^2+6N\|\boldsymbol{v}\|^2\bar{\alpha}_k^2\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2.\end{aligned}\tag{97}$$

Substituting (96) and (97) into (95), and using the condition $\alpha\leqslant\frac{1-\sigma_C^2}{12\sqrt{2}\kappa_v\delta_{C,2}\|\boldsymbol{I}-\frac{\boldsymbol{v}\boldsymbol{1}^\top}{N}\|_C\left(AL_0+BL_1b+BL_1\ell G\right)}$, we obtain

$$\begin{aligned}\|\boldsymbol{e}_{y,k+1}\|_C^2\leqslant&\tfrac{1+\sigma_C^2}{2}\|\boldsymbol{e}_{y,k}\|_C^2\\&+\mathcal{C}_{y,1}\|\boldsymbol{e}_{x,k}\|_R^2+\alpha\,\mathcal{C}_{y,2}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2,\end{aligned}\tag{98}$$

其中，常数 $\mathcal{C}_{y,\ 1}$ 和 $\mathcal{C}_{y,\ 2}$ 定义于（41）和（42）中。

附录 G：引理 13 的证明

现在，我们为全局目标函数 $F(\bar{\boldsymbol{x}}^k)$ 推导递归下降关系。当步长满足 $\alpha\leqslant\frac{\boldsymbol{u}^\top\boldsymbol{v}}{9LN\|\boldsymbol{v}\|^2}$ 时，由（35）可得

$$\begin{aligned}&\frac{\boldsymbol{u}^\top\boldsymbol{v}}{3N}\bar{\alpha}_k\|\nabla F(\bar{\boldsymbol{x}}^k)\|^2\leqslant F(\bar{\boldsymbol{x}}^k)-F(\bar{\boldsymbol{x}}^{k+1})\\&+\left(\frac{6L\kappa_{uv}\|\boldsymbol{v}\|}{2N}+\frac{2\kappa_{uv}^2\|\boldsymbol{v}\|^2}{N\boldsymbol{u}^\top\boldsymbol{v}}\right)\bar{\alpha}_k\|\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\|^2.\end{aligned}\tag{99}$$

根据引理 17，可得

$$\begin{aligned}&\|\boldsymbol{y}^k-\boldsymbol{v}\nabla F(\bar{\boldsymbol{x}}^k)\|^2\\\leqslant&2\|\boldsymbol{e}_{y,k}\|_C^2+\frac{2\|\boldsymbol{v}\|^2}{N}(AL_0+BL_1(b+\ell G))^2\|\boldsymbol{e}_{x,k}\|_R^2.\end{aligned}\tag{100}$$

将 (100) 代入 (99) 即可得到所需结果。

参考文献

[1] J. Dean、G. Corrado、R. Monga、K. Chen、M. Devin、M. Mao、M. Ran- zato、A. Senior、P. Tucker、K. Yang 等人，《大规模分布式深度网络》，《神经信息处理系统进展》，第 25 卷，2012年。[2] F. Chen、W. Ren 等人，《多智能体系统控制综述》，《系统与控制基础与趋势》，第 6 卷，第 4 期，第 339–499 页，2019年。[3] M. Rabbat 和 R. Nowak，《传感器网络中的分布式优化》，载于《第三届传感器网络信息处理国际研讨会》，第 20–27 页，2004年。[4] A.-H. Mohsenian-Rad、V. W. Wong、J. Jatskevich 和 R. Schober，《智能电网中基于激励的最优自主能源消耗调度算法》，载于《2010年创新智能电网技术（ISGT）》，第 1–6 页，IEEE，2010年。[5] X. Chen、L. Jiao、W. Li 和 X. Fu，《面向移动边缘云计算的高效多用户计算卸载》，《IEEE/ACM网络汇刊》，第 24 卷，第 5 期，第 2795–2808 页，2015年。[6] J. Apostolopoulos、T. Wong、W.-t. Tan 和 S. Wee，《内容分发网络中的多描述流媒体》，载于《IEEE计算机与通信学会第二十一届年度联合会议论文集》，第 3 卷，第 1736–1745 页，IEEE，2002年。[7] A. Nedi´c 和 A. Ozdaglar，《多智能体优化的分布式次梯度方法》，IEEE《自动控制汇刊》，第 54 卷，第 1 期，第 48–61 页，2009年。[8] K. Yuan、Q. Ling 和 W. Yin，《去中心化梯度下降的收敛性》，SIAM《优化杂志》，第 26 卷，第 3 期，第 1835–1854 页，2016年。[9] J. Chen 和 A. H. Sayed，《网络分布式优化与学习的扩散自适应策略》，IEEE《信号处理汇刊》，第 60 卷，第 8 期，第 4289–4305 页，2012年。[10] R. You 和 S. Pu，《用于去中心化非凸优化的随机推拉》，arXiv 预印本 arXiv:2506.07021，2025年。[11] Y. Bo 和 Y. Wang，《量化可避免分布式优化中的鞍点》，《美国国家科学院院刊》，第 121 卷，第 17 期，第 e2319625121 页，2024年。[12] A. H. Sayed，《自适应网络》，《IEEE会刊》，第 102 卷，第 4 期，第 460–497 页，2014年。[13] J. Zhang、T. He、S. Sra 和 A. Jadbabaie，《梯度裁剪为何能加速训练：关于自适应性的理论解释》，arXiv 预印本 arXiv:1905.11881，2019年。[14] H. Li、J. Qian、Y. Tian、A. Rakhlin 和 A. Jadbabaie，《广义光滑性下的凸与非凸优化》，《神经信息处理系统进展》，第 36 卷，第 40238–40271 页，2023年。[15] M. Crawshaw、M. Liu、F. Orabona、W. Zhang 和 Z. Zhuang，《广义 SignSGD 对无界光滑性的鲁棒性》，《神经信息处理系统进展》，第 35 卷，第 9955–9968 页，2022年。[16] B. Zhang、J. Jin、C. Fang 和 L. Wang，《非凸优化裁剪算法的改进分析》，《神经信息处理系统进展》，第 33 卷，第 15511–15521 页，2020年。[17] E. Gorbunov、N. Tupitsa、S. Choudhury、A. Aliev、P. Richt´arik、S. Horv´ath 和 M. Tak´aˇc，《凸 (L_0, L_1)-光滑优化方法：裁剪、加速与自适应性》，载于《第十三届国际学习表征会议》，第 1309–1353 页，2025年。[18] D. Vankov、A. Rodomanov、A. Nedich、L. Sankar 和 S. U. Stich，《利用梯度方法优化 (L_0, L_1)-光滑函数》，arXiv 预印本 arXiv:2410.10800，2024年。[19] F. H¨ubler、J. Yang、X. Li 和 N. He，《宽松光滑性下的参数无关优化》，载于《国际人工智能与统计会议》，第 4861–4869 页，PMLR，2024年。[20] Z. Chen、Y. Zhou、Y. Liang 和 Z. Lu，《广义光滑非凸优化与光滑非凸优化同样高效》，载于《国际机器学习会议》，第 5396–5427 页，PMLR，2023年。[21] X. Yang、H. Zhang、W. Chen 和 T.-Y. Liu，《用于差分隐私非凸优化的带扰动归一化/裁剪随机梯度下降》，arXiv 预印本 arXiv:2206.13033，2022年。[22] M. Faw、L. Rout、C. Caramanis 和 S. Shakkottai，《超越一致光滑性：自适应随机梯度下降的停止分析》，载于《第三十六届年度学习理论会议》，第 89–160 页，PMLR，2023年。[23] B. Wang、H. Zhang、Z. Ma 和 W. Chen，《非凸目标函数上 AdaGrad 的收敛性：简明证明与放宽的假设》，载于《第三十六届年度学习理论会议》，第 161–190 页，PMLR

，2023年。

[24] 放宽假设下 Adam 的收敛性，神经信息处理系统进展，第 36 卷，第 52166–52196 页，2023 年。[25] A. Reisizadeh 、 H. Li 、 S. Das 和 A. Jadbabaie，“非凸优化的方差缩减裁剪”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP），第 1–5 页，2025年。[26] Q. Zhang、Y. Zhou、S. Khan、A. Prater-Bennette、L. Shen 和 S. Zou，“重新审视大规模非凸分布鲁棒优化”，载于国际学习表征会议，2025年。[27] Z. Jiang、A. Balu 和 S. Sarkar，“采用梯度下降方法的去中心化宽松光滑优化”，arXiv 预印本 arXiv:2508.08413，2025年。[28] L. Luo、X. Cui、T. Jia 和 C. Chen，“宽松光滑性下的去中心化随机非凸优化”，arXiv 预印本 arXiv:2509.08726，2025年。[29] T. Sun、Q. Wang、D. Li 和 B. Wang，“弱光滑性假设下非凸分布式随机梯度下降的裁剪”，2023年。[30] M. Assran、N. Loizou、N. Ballas 和 M. Rabbat，“用于分布式深度学习的随机梯度推送”，载于第 36 届国际机器学习会议（ICML）论文集，《机器学习研究论文集》，第 344–353 页，PMLR，2019年。[31] E. Gorbunov、F. Hanzely 和 P. Richt´arik，“本地随机梯度下降：统一理论与新的高效方法”，载于国际人工智能与统计会议，第 3556–3564 页，PMLR，2021年。[32] 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”，神经信息处理系统进展，第 30 卷，2017年。[33] B. E. Woodworth、K. K. Patel 和 N. Srebro，“小批量与本地随机梯度下降：异构分布式学习”，神经信息处理系统进展，第 33 卷，第 6281–6292 页，2020年。[34] B. Gharesifard 和 J. Cort´es，“有向图何时存在双随机邻接矩阵？”，载于 2010年美国控制会议论文集，第 2440–2445 页，IEEE，2010年。[35] L. Sabattini、C. Secchi 和 N. Chopra，“保持有向图强连通性的去中心化估计与控制”，IEEE《控制论汇刊》，第 45 卷，第 10 期，第 2273–2286 页，2014 年。[36] S. Sundhar Ram 、 A. Nedi´c 和 V. V. Veeravalli，“凸优化的分布式随机次梯度投影算法”，《优化理论与应用杂志》，第 147 卷，545，2010 3，516 no. pp. – 。[37] R. Xin、U. A. Khan 和 S. Kar，“具有加速收敛的方差缩减去中心化随机优化”，IEEE《信号处理汇刊》，第 68 卷，第 6255–6271 页，2020年。[38] S. Pu 和 A. Nedi´c，“分布式随机梯度跟踪方法”，《数学规划》，第 187 卷，第 1 期，第 409–457 页，2021年。[39] C. Xi、Q. Wu 和 U. A. Khan，“有向网络上的分布式优化”，《神经计算》，第 267 卷，第 508–515 页，2015年。[40] A. Nedi´c 和 A. Olshevsky，“时变有向图上的分布式优化”，IEEE《自动控制汇刊》，第 60 卷，第 3 期，第 601–615 页，2014年。[41] G. Scutari 和 Y. Sun，“时变有向图上的分布式非凸约束优化”，《数学规划》，第 176 卷，第 497–544 页，2019年。[42] A. Nedi´c、A. Olshevsky 和 W. Shi，“实现时变图上分布式优化的几何收敛”，SIAM《优化杂志》，第 27 卷，第 4 期，第 2597–2633 页，2017年。[43] S. Pu、W. Shi、J. Xu 和 A. Nedi´c，“用于网络分布式优化的推拉梯度方法”，IEEE《自动控制汇刊》，第 66 卷，第 1 期，第 1–16 页，2020年。[44] H. Li 和 Z. Lin，“重新审视光滑分布式优化的 EXTRA”，SIAM《优化杂志》，第 30 卷，第 3 期，第 1795–1821 页，2020年。[45] K. Yuan、X. Huang、Y. Chen、X. Zhang、Y. Zhang 和 P. Pan，“重新审视光滑非凸随机去中心化优化的最优收敛速度”，神经信息处理系统进展，第 35 卷，第 36382–36395 页，2022年。[46] Z. Song、L. Shi、S. Pu 和 M. Yan，“去中心化优化的最优梯度跟踪”，《数学规划》，第 207 卷，第 1 期，第 1–53 页，2024年。[47] J. Wang、R. Das、G. Joshi、S. Kale、Z. Xu 和 T. Zhang，“异构数据下联邦平均的惊人有效性”，arXiv 预印本 arXiv:2206.04723，2022年。[48] T. Li、A. K. Sahu、M. Zaheer、M. Sanjabi、A. Talwalkar 和 V. Smith，“异质网络中的联邦优化”，《机器学习与系统论文集

，第 2 卷，第 429–450 页，2020年。

[49] W. J. Fu, 《惩罚回归：桥回归与套索回归》, 《计算与图形统计学杂志》, 第7卷, 1998年3, 397no. pp. – 。[50] R. Yang、J. Tian 和 Y. Zhang, 《面向个性化联邦学习的正则化互学》, 载于《亚洲机器学习会议》, 第1521–1536页, PMLR, 2021年。[51] V. Smith、C.-K. Chiang、M. Sanjabi 和 A. S. Talwalkar, 《联邦多任务学习》, 《神经信息处理系统进展》, 第30卷, 2017年。[52] S. P. Karimireddy、S. Kale、M. Mohri、S. Reddi、S. Stich 和 A. T. Suresh, 《Scaffold：联邦学习的随机受控平均》, 载于《国际机器学习会议》, 第5132–5143页, PMLR, 2020年。[53] C. Chang 和 C.-J. Lin, 《Libsvm：支持向量机库》, 《ACM智能系统与技术汇刊》（TIST）, 第2卷, 第3期, 第1–27页, 2011年。[54] R. Xin 和 U. A. Khan, 《一种在有向图上进行优化且具有几何收敛性的线性算法》, 《IEEE控制系统快报》, 第2卷, 第3期, 第315–320页, 2018年。[55] A. Krizhevsky 和 G. Hinton, 《从微型图像中学习多层特征》, 技术报告, 多伦多大学, 2009年。[56] Z. Jiang、A. Balu、C. Hegde 和 S. Sarkar, 《固定拓扑网络中的协作式深度学习》, 《神经信息处理系统进展》, 第30卷, 2017年。[57] Z. Jiang、A. Balu、C. Hegde 和 S. Sarkar, 《协作式深度学习中的共识与最优性权衡》, 《人工智能前沿》, 第4卷, 文章编号573731, 2021年。[58] X. Chen、B. Karimi、W. Zhao 和 P. Li, 《关于去中心化自适应梯度方法的收敛性》, 载于《亚洲机器学习会议》, 第217–232页, PMLR, 2023年。