arXiv:2301.01313v1 [math.OC] 3 Jan 2023

带有本地步长的去中心化梯度跟踪

刘悦
多伦多大学
加拿大

林涛
西湖大学
中国

Anastasia Koloskova
EPFL
瑞士

Sebastian U. StichCISPA 亥姆霍兹信息安全中心德国

摘要

梯度跟踪（GT）是一种用于解决网络上的去中心化优化问题（例如训练机器学习模型）的算法。GT 的一个关键特性是采用跟踪机制，以克服节点之间的数据异质性。

我们提出了一种新颖的去中心化跟踪机制 K-梯度跟踪，使梯度跟踪能够进行通信高效的本地更新，同时继承梯度跟踪的数据独立性。我们证明了 K-梯度跟踪在光滑非凸函数上的收敛速率，并证明其可渐近地将通信开销降低线性因子 K，其中 K 表示本地步数。我们在凸和非凸基准问题，以及使用 MNIST 数据集进行的非凸神经网络训练任务中，展示了这种异质性校正的稳健性和有效性。

1 引言

我们考虑分布式优化问题，其中模型 $\mathbf{x}$ 上的目标函数 $f(\mathbf{x}) \in \mathbb{R}^d$ 定义为 n 个不同分量 $\{f_1(\mathbf{x}) \ldots, f_n(\mathbf{x})\}$的平均值，即：

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}).$$

在分布式应用中，不同的参与者（或“客户端”）参与训练。这些客户端可以是移动边缘设备或计算节点。通常，每个分量 $f_i(\mathbf{x})$仅由单个客户端持有（例如，当 $f_i(\mathbf{x})$基于仅在客户端本地可用的训练数据定义时）。这使得分布式优化问题比集中式问题更难求解。

除了迭代次数方面的收敛速率外，通信效率也是分布式算法设计中最重要的指标之一。为便于说明，我们考虑全局函数梯度的计算：$\nabla f(\mathbf{x}) = \frac{1}{n}\sum_i \nabla f_i(\mathbf{x})$，它构成了一般一阶方法的基础。由于每个客户端只能计算局部梯度 $\nabla f_i(\mathbf{x})$，因此还需要计算这些局部梯度的平均值。集中式算法 [14, 23] 通过中央控制器实现这种全局聚合，例如使用参数服务器 [16]。然而，这种方法要求所有客户端同时与中央服务器通信，导致该热点出现通信瓶颈，并拖慢运行时间。去中心化算法 [10, 18, 31] 不进行精确平均，而是仅通过 Gossip 平均进行部分通信，并允许节点仅与较少的节点（例如，仅与其邻居）通信，从而减少通信开销，避免形成最繁忙的节点。节点之间的连接方式构成了网络拓扑。

去中心化优化中最具挑战性的方面之一是数据异质性，即训练数据在各节点之间并非独立同分布。此类非独立同分布通常出现在实际应用中，例如，来自手机、传感器或医院的训练数据可能存在区域差异 [8]。在这种情况下，每个客户端的局部经验损失各不相同。这会减缓

收敛 [12, 25]，甚至导致局部过拟合（通常称为 客户端漂移），因为客户端在优化过程中可能会偏离全局最优解 [6, 9, 12, 17]。
已有多种去中心化算法被证明能够缓解异质性。然而，大多数算法仅被证明能在强凸函数下收敛 [1, 15, 24, 28, 32], ，或者虽被证明适用于光滑非凸函数，但对网络拓扑有严格限制 [，例如 30]。为解决任意网络中光滑非凸函数下的数据异质性问题，研究者提出了随机梯度跟踪（GT） [11, 22, 24, 27, 35]算法。其收敛速率仅取决于初始点处的数据异质性，通过适当初始化可以将其完全消除。然而，客户端每次模型更新后都必须与网络中的所有邻居进行通信。因此，这些方法仍然会产生较高的通信开销。

为进一步降低分布式训练中的通信开销，研究者提出了多种工程技术，例如使用大批量 [4, 20, 33], 、模型/梯度压缩 [13] 或异步通信 [19]。在本研究中，我们重点关注通过局部更新降低通信频率。这种方法在实践中通常很有效，但在理论分析方面仍然颇具挑战 [5, 14, 23, 29]。然而，大量执行本地步可能会加剧客户端漂移。由此产生的优化困难可能抵消通信方面的节省 [9, 12]。在去中心化优化中，如何在引入本地步的同时实现对异质性的独立性，仍鲜有研究。

将本地更新整合到梯度跟踪中并非易事。例如，在梯度跟踪中简单地跳过通信轮次（从而在间隔期间执行若干本地更新），在实践中效果并不好[1]。一项同期研究 LU-GT[26]分析了梯度跟踪周期性跳过通信时的性能，但仅限于确定性场景。[2]

为此，我们精心设计了一种新颖的跟踪机制，使梯度跟踪能够与本地步相结合。[3] 由此得到的算法——K-GT，其中 K 表示本地步的数量——是一种新颖的去中心化方法，可通过局部更新实现通信高效跟踪。我们证明， K-GT 的收敛性仅取决于起始点处的数据异质性，并且通过额外一轮全局通信，可以完全消除这种较弱的数据依赖性。只要 K-GT 使用与梯度跟踪相同的初始化， K-GT 就能继承梯度跟踪不受异质性影响的性质。我们证明， K-GT（算法1）在通信轮次方面，随着本地步数 K 和客户端数量 n的增加，可实现渐近线性加速；并且它能在$\mathcal{O}\left(\frac{\sigma^2}{nK\epsilon^2}\right)$ 轮内收敛到一个 ϵ-近似驻点。与梯度跟踪相比，通信轮次渐近减少了 K 倍。我们进一步表明，若采用适当的初始化，收敛速率（包括高阶项）不依赖于数据异质性；而对于不采用跟踪机制的去中心化随机梯度下降（D-SGD）等方法，情况则并非如此。

本文的结构安排如下：第 2 节精确描述分布式优化问题的设定。第 3 节介绍 K-GT 的算法设计，并说明其如何帮助纠正异质性。本文的主要结果给出了该算法的收敛速率，见定理 3.2。第 4 节推广梯度跟踪框架，并讨论其他 GT 替代算法的缺陷，这些缺陷也可能源于同一框架。此外，我们给出这些算法的收敛结果，并通过比较表明， K-GT 在理论上具有最高的通信效率。第 5 节通过数值示例，详细比较 GT 变体与基线 D-SGD。

主要贡献。 我们在下文总结主要结果。

- 我们为分布式优化提出了一种新颖的梯度跟踪算法，并分析其收敛性质。我们证明，在适当初始化的条件下，K-GT 具有与异质性无关的复杂度估计，并证明它在 $\mathcal{O}\left(\frac{\sigma^2}{nK\epsilon^2}\right)$ 轮内渐近收敛，其中 n 表示客户端数量， K 表示本地步数，σ^2 表示随机噪声水平，ϵ 表示准确率。与 GT 基线相比，这将复杂度提升了 K 倍。

- 我们通过研究以下内容提供了更多理论见解：(i) GT 的朴素局部扩展——周期性 *GT*——的收敛性，说明当随机噪声较大时，其表现不如 K-GT；以及 (ii) 一种计算效率较低的变体，大批量梯度跟踪，其迭代次数与 K-GT相当，但计算复杂度并不相当。

- 我们通过实验验证了针对强凸函数和非凸函数的理论结果，并阐释了噪声、本地步和数据异质性对收敛的影响。 K-GT 能够抵御数据异质性，同时提升通信效率，并且相比基线算法具有更好的泛化性能。

[1]我们在下文第 *5* 节评估这一变体（称为周期性梯度跟踪），例如参见图 2。 [2]这项同期工作是在我们完成本手稿的过程中独立开展的。我们将在本手稿的下一版本中加入更详细的比较。[3]本文的部分结果此前已发表于 YL 的硕士论文 [21]

表 1：不同工作条件下的比较。 $\Delta \leq n$ 表示通信图的最大度。K-梯度跟踪是首个采用本地步的完全去中心化跟踪算法。

算法	设置		
	通信最繁忙节点的通信开销	本地步	异质性鲁棒性[a]
SCAFFOLD [9]	$\mathcal{O}(n)$	✓	✓
GOSSIP-PGA [2]	$\mathcal{O}(\Delta)$	✓	✗
D-SGD [12]		✓	✗
梯度跟踪 [24]		✗	✓
D^2 [30]		✗	✓
K-梯度跟踪 **[我们的方法]**		✓	✓

[a]The data 异构 异质性不会影响最坏情况收敛速率（但可能需要特殊初始化).

2 问题设定

本节将介绍相关符号和问题设定。

2.1 去中心化优化问题

我们考虑以下由 n个客户端损失函数求和构成的优化问题：

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

其中， n 表示系统中的客户端数量， ξ_i 是从 $\mathcal{D}_i$ 中抽取的随机样本，而 $\mathcal{D}_i$ 表示仅在节点 $i \in [n]$上可用的局部分布。根据应用场景， $\mathcal{D}_i$ 可以是任意的，且在不同客户端之间有所差异。这一设定既适用于经验风险最小化，也适用于在线优化。

在本研究中，我们考虑一般的光滑非凸函数和有界随机噪声。

假设 1 (光滑性)。 每个函数 $f_i(x): \mathbb{R}^d \to \mathbb{R}$, $\forall i \in [n]$ 均可微，且存在一个常数$L > 0$ ，使得对于每个$\mathbf{x}$、$\mathbf{y} \in \mathbb{R}^d$,

$$f_i(\mathbf{y}) \leq f_i(\mathbf{x}) + \nabla f_i(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}||\mathbf{x}-\mathbf{y}||_2^2 .$$

假设 2(有界方差)。 每个客户端的方差均有界，

$$\forall i \in [n],\ \forall \mathbf{x} \in \mathbb{R}^d,\ \mathbb{E}_{\xi\sim\mathcal{D}_i}||\nabla F_i(\mathbf{x};\xi) - \nabla f_i(\mathbf{x})||_2^2 \leq \sigma^2 .$$

2.2 通信图

训练在去中心化网络上进行，其拓扑建模为一个无向图：$(\mathcal{V},\mathcal{E})$，其中 $\mathcal{V}:=\{1,2,\ldots,n\}$ 是节点集，$\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$是边集。节点（或客户端）代表一个计算节点，客户端仅沿着边 $e \in \mathcal{E}$进行通信。我们将邻接矩阵记为 $\mathbf{W} \in \mathbb{R}^{n\times n}$，其中 $w_{ij} = 0$表示节点 i 和 j 不相连，即 $e_{ij} = (i,j) \notin \mathcal{E}$。

假设 3 (混合率)。 给定由非负实数构成的对称双随机混合矩阵 $\mathbf{W} \in \mathbb{R}^{n\times n}$，即 $\forall i,j \in [n]$, $w_{ij} \geq 0$, $\sum_{i=1}^n w_{ij} = \sum_{j=1}^n w_{ij} = 1$，共识距离在平均步骤后呈线性递减，即存在一个 $1 \geq p > 0$ 使得

$$||\mathbf{X}\mathbf{W} - \bar{\mathbf{X}}||_F^2 \leq (1-p)||\mathbf{X}-\bar{\mathbf{X}}||_F^2,\ \forall \mathbf{X} \in \mathbb{R}^{d\times n} .$$

请注意，如果常用的网络参数 $\rho := ||\mathbf{W} - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}||$ [2] 严格小于 1，则 $1 \geq p > 0$ [例如，参见 11, 12]。混合率描述了网络的连通性。 p的值越大，通信图的连通性越好。 $p = 1$ 对于完全图 $\mathbf{W} = \frac{1}{n}\mathbf{1}\mathbf{1}^T$，而 $p = 0$ 对于非连通图 $\mathbf{W} = \mathbf{I}_n$。

2.3 数据异质性与校正项

当每个客户端上的局部分布 $\{\mathcal{D}_i\}$ 都相同时，局部函数 $\{f_i(\mathbf{x})\}$ 也彼此相同，即$f_i(\mathbf{x}) \equiv f(\mathbf{x})$。否则，异质局部分布会导致异质局部函数。异质性通常通过局部梯度 $\{\nabla f_i(\mathbf{x})\}$ 与全局梯度 $\nabla f(\mathbf{x})$ [9, 12] 之间的差异来衡量，如下所示。

假设4 (数据异质性)。 存在常数 $\zeta^2 > 0$ 和 $B \geq 1$， 使得

$$\forall \mathbf{x} \in \mathbb{R}^d,\ \frac{1}{n}\sum_{i=1}^{n} ||\nabla f_i(\mathbf{x})||^2 \leq \bar{\zeta}^2 + B^2||\nabla f(\mathbf{x})||^2\,,$$

其中， $\bar{\zeta}^2$ 和 B^2 均表示系统内异质性的程度。

基线去中心化随机梯度下降（D-SGD）使用相对于本地模型的朴素梯度，其收敛不可避免地会受到 $\bar{\zeta}^2$ 和 B [25]的影响。

2.3.1 符号说明

梯度跟踪算法主要操作两个变量：模型迭代量 $\mathbf{x} \in \mathbb{R}^d$ 和跟踪变量$\mathbf{z} \in \mathbb{R}^d$。具体而言，我们将节点 i 在通信轮次 t 的本地步 k 上的向量$\mathbf{y} \in \{\mathbf{x}, \mathbf{z}\}$ 记为$\mathbf{y}_i^{(t)+k}$， 并将其平均值记为 $\bar{\mathbf{y}} = \frac{1}{n}\sum_i \mathbf{y}_i$。

所有 $i \in [n]$ 对应的向量 $\mathbf{y}_i$ 组成的集合，以矩阵形式表示时记为一个大写字母，其列为$\mathbf{y}_i$，即:

$$\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}, \quad \bar{\mathbf{Y}} = [\bar{\mathbf{y}}, \ldots, \bar{\mathbf{y}}] = \tfrac{1}{n}\mathbf{Y}\mathbf{1}_n\mathbf{1}_n^T \in \mathbb{R}^{d\times n}\,.$$

此外，我们将此矩阵定义扩展至式（1）中相对于模型 $\mathbf{X}$ 在样本上的$\xi = [\xi_1, \ldots, \xi_n]$梯度和随机梯度，其中 $\xi_i \sim \mathcal{D}_i$,

$$\begin{aligned}\nabla F(\mathbf{X};\xi) &= [\nabla F_1(\mathbf{x}_1;\xi_1), \ldots, \nabla F_n(\mathbf{x}_n;\xi_n)] \in \mathbb{R}^{d\times n},\\ \nabla f(\mathbf{X}) &= \mathbb{E}_{(\xi_1,\ldots,\xi_n)}\nabla F(\mathbf{X};\xi) = [\nabla f_1(\mathbf{x}_1), \ldots, \nabla f_n(\mathbf{x}_n)] \in \mathbb{R}^{d\times n}\,.\end{aligned}$$

2.3.2 梯度跟踪

梯度跟踪算法（GT） [27]由以下更新方程定义:

$$\begin{aligned}\mathbf{X}^{(t+1)} &= (\mathbf{X}^{(t)} - \eta\mathbf{Z}^{(t)})\mathbf{W}\\ \mathbf{Z}^{(t+1)} &= \mathbf{Z}^{(t)}\mathbf{W} + \mathbf{G}^{(t+1)} - \mathbf{G}^{(t)},\end{aligned} \tag{2}$$

的矩阵形式。这里，$\mathbf{G}^{(t)} = \nabla F(\mathbf{X}^{(t)};\ \xi^{(t)})$ 和 $\eta > 0$ 表示步长。

当不同节点之间的数据存在异质性时， $\{\nabla F_i(\mathbf{x};\xi_i)\ \ \forall i\}$ 各不相同。但 GT 使用偏差校正来补偿每个节点的异质梯度。这一校正由跟踪变量$\mathbf{Z}$ 控制，它取代了朴素梯度:

$$\mathbf{Z}^{(t+1)} = \nabla F(\mathbf{X}^{(t+1)};\xi^{(t+1)}) + \underbrace{\mathbf{Z}^{(t)}\mathbf{W} - \mathbf{G}^{(t)}}_{\text{correction}} \tag{3}$$

由于更新式（2）同时更新模型$\mathbf{X}$ 和跟踪变量$\mathbf{Z}$，因此无需额外考虑局部梯度的异质性。已证明，无论数据异质性如何，GT 都能收敛 [27]。

3 K-GT：梯度求和跟踪算法

在本节中，我们将介绍新的去中心化随机算法 K-梯度跟踪，并分析其在一般非凸函数下的收敛性。

3.1 算法

在 K-梯度跟踪算法中，我们允许每个客户端在每个通信轮次之间执行 $K \geq 1$ 本地步。为补偿数据异质性，我们在随机梯度之上采用类似于式(3)的校正项。我们将节点 $\mathbf{c}_i$ 上的校正项记为 i。然后，每个节点重复以下更新规则， $i \in [n]$:

1. 通过从分布 $\mathcal{D}_i$ 中采样 ξ_i， 计算本地随机梯度 $\nabla F_i(\mathbf{x}_i;\xi_i)$;
2. 在第 t 次通信中，使用第 $t+k$ 次迭代的随机梯度和校正项 $\mathbf{c}_i^{(t)}$ 更新本地模型 $\mathbf{x}_i^{(t)+k+1} = \mathbf{x}_i^{(t)+k} - \eta_c\big(\nabla F_i(\mathbf{x}_i^{(t)+k};\xi_i^{(t)+k}) + \mathbf{c}_i^{(t)}\big)$;

3. 重复步骤 (1)-(2) K 次，然后获得本地步中的跟踪量，$\mathbf{z}_i^{(t)} = \frac{1}{K\eta_c}\left(\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t)+K}\right)$。与邻居交换 $\{\mathbf{x}_i, \mathbf{c}_i\}$（矩阵形式）：

$$\begin{aligned}\mathbf{X}^{(t+1)} &= \left(\mathbf{X}^{(t)} - \eta_s(\mathbf{X}^{(t)} - \mathbf{X}^{(t)+K})\right)\mathbf{W}\,,\\ \mathbf{C}^{(t+1)} &= \mathbf{C}^{(t)} + \mathbf{Z}^{(t)}(\mathbf{W} - \mathbf{I})\,.\end{aligned} \tag{4}$$

完整算法总结如下，见算法1。

命题 3.1 (梯度和跟踪)。将$\mathbf{Z}^{(t)} = \frac{1}{K\eta_c}\left(\mathbf{X}^{(t)} - \mathbf{X}^{(t)+K}\right)$定义为通信轮次 t 期间的跟踪变量。在 K-*GT* 中，通信时模型 $\mathbf{X}^{(t)}$ 和跟踪变量 $\mathbf{Z}^{(t)}$ 的更新规则可改写为（$\eta = \eta_s\eta_c$）：

$$\begin{aligned}\mathbf{X}^{(t+1)} &= \left(\mathbf{X}^{(t)} - K\eta\mathbf{Z}^{(t)}\right)\mathbf{W}\,,\\ \mathbf{Z}^{(t+1)} &= \mathbf{Z}^{(t+1)}\mathbf{W} + \mathbf{G}^{(t+1)} - \mathbf{G}^{(t)}\,,\end{aligned} \tag{5}$$

其中，$\mathbf{G}^{(t)} = \frac{1}{K}\sum_k \nabla F(\mathbf{X}^{(t)+k};\ \xi^{(t)+k})$ 表示本地步上的平均更新。

详细证明见附录B.1。

注 1。若 $K = 1$在（5）中，K-GT 等价于带有梯度跟踪的梯度跟踪 [27]，其中 $\eta = \eta_s\eta_c$。

K-GT 在通信最充分时本质上就是运行随机梯度下降。为了理解 K-GT 背后的直觉，我们考虑每次迭代时的全局平均值 $\mathbf{X}$，它的更新方式与标准随机梯度下降相同：

$$\begin{aligned}\bar{\mathbf{X}}^{(t)+k+1} &= \left(\mathbf{X}^{(t)+k} - \eta_c(\nabla F(\mathbf{X}^{(t)+k};\xi^{(t)+k}) + \mathbf{C}^{(t)})\right)\tfrac{\mathbf{1}\mathbf{1}^T}{n}\\ &= \bar{\mathbf{X}}^{(t)} - \eta_c\left(\overline{\nabla F(\mathbf{X}^{(t)+k};\xi^{(t)+k})} + \bar{\mathbf{C}}^{(t)}\right).\end{aligned}$$

若初始化为 $\mathbf{C}^{(0)} = \nabla F(\mathbf{X}^{(0)};\ \xi^{(0)})\left(\frac{\mathbf{1}\mathbf{1}^T}{n} - \mathbf{I}\right)$，则校正项的平均值满足

$$\begin{aligned}\bar{\mathbf{C}}^{(t+1)} &= \bar{\mathbf{C}}^{(t)} + \mathbf{Z}^{(t)}(\mathbf{W} - \mathbf{I})\tfrac{\mathbf{1}\mathbf{1}^T}{n} = \bar{\mathbf{C}}^{(t)}\,,\\ \bar{\mathbf{C}}^{(0)} &= \nabla F(\mathbf{X}^{(0)};\xi^{(0)})\left(\tfrac{\mathbf{1}\mathbf{1}^T}{n} - \mathbf{I}\right)\frac{\mathbf{1}\mathbf{1}^T}{n} \equiv \mathbf{0}\,.\end{aligned}$$

那么模型迭代量的平均值满足

$$\bar{\mathbf{X}}^{(t)+k+1} = \bar{\mathbf{X}}^{(t)} - \eta_c\overline{\nabla F(\mathbf{X}^{(t)+k};\xi^{(t)+k})},$$

它使用平均随机梯度更新模型。

这种校正如何改进 D-SGD? 我们考虑应用 [30] 中类似的分析，以说明 K-GT 的有效性。假设 $\mathbf{X}^{(t)}$ 已达到最优解 $\mathbf{X}^\star := \mathbf{x}^\star\mathbf{1}^T$，且所有本地模型都等于最优解$\mathbf{x}^\star$。根据附录中的分析（引理 C.8），校正项将等于

$$\mathbf{c}_i^\star := -\nabla F_i(\mathbf{x}^\star;\xi) + \frac{1}{n}\sum_j \nabla F_j(\mathbf{x}_j^\star;\xi)\,.$$

那么 K-GT 的下一次本地更新将为

$$\begin{aligned}\mathbf{X}^{(t)+1} &= \mathbf{X}^\star - \eta_c(\nabla F(\mathbf{X}^\star;\xi) + \mathbf{C}^\star)\\ &= \mathbf{X}^\star - \eta_c\nabla F(\mathbf{X}^\star;\xi)\tfrac{\mathbf{1}\mathbf{1}^T}{n}\,.\end{aligned}$$

这个例子表明，对于 K-GT，当我们仅通过本地更新逼近解时，收敛情况取决于 $\mathbf{E}||\nabla F(\mathbf{X}^\star;\ \xi)\frac{\mathbf{1}\mathbf{1}^T}{n}||_F^2$ 的大小，而该大小有界于 $\mathcal{O}(\sigma^2)$。

然而，考虑 D-SGD 的相同情况，

$$\mathbf{X}^{(t)+1} = \mathbf{X}^\star - \eta\nabla F(\mathbf{X}^\star;\xi).$$

在不同节点上，$\nabla F_i(\mathbf{X}^\star;\ \xi)$ 由于数据异质性而彼此不同，且这种偏差只能用 ζ^2 来刻画，正如假设4所示。那么，在解的邻域内达到相同收敛量级时，D-SGD 的上界为 $\mathcal{O}(\sigma^2 + \bar{\zeta}^2)$ [30], 这显然比 K-梯度跟踪更差。D-SGD 中由数据异质性导致的额外 $\mathcal{O}(\bar{\zeta}^2)$ 项，如果始终只使用单一随机梯度 [25]，就永远无法得到改善。

算法 1 K-梯度跟踪：梯度和跟踪

1：参数： T：通信次数； K：本地步数； η_c, η_s：本地步长、通信步长； $\mathbf{W}$：给定拓扑。**2：初始化：** $\forall i, j \in [n]$, $\mathbf{x}_i^{(0)} = \mathbf{x}_j^{(0)}$； $\mathbf{c}_i^{(0)} = -\nabla F_i(\mathbf{x}^{(0)}; \xi_i) + \frac{1}{n}\sum_j \nabla F_j(\mathbf{x}^{(0)}; \xi_j)$[a]。**3：对客户端** $i \in \{1, \dots, n\}$ **并行执行 4：对通信轮次：从** $t \leftarrow 0$ **到** $T-1$ **执行 5：对本地步：从** $k \leftarrow \overline{0}$ **到** $K-1$ **执行 6：** $\mathbf{x}_i^{(t)+k+1} = \mathbf{x}_i^{(t)+k} - \eta_c(\nabla F_i(\mathbf{x}_i^{(t)+k}; \xi_i^{(t)+k}) + \mathbf{c}_i^{(t)})$ **7：循环结束 8：** $\mathbf{z}_i^{(t)} = \frac{1}{K\eta_c}(\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t)+K})$ **9：** $\mathbf{c}_i^{(t+1)} = \mathbf{c}_i^{(t)} - \mathbf{z}_i^{(t)} + \sum_j w_{ij}\mathbf{z}_j^{(t)}$

10: $\mathbf{x}_i^{(t+1)} = \sum_j w_{ij}(\mathbf{x}_j^{(t)} - K\eta_s\eta_c\mathbf{z}_j^{(t)})$ ▷ 更新跟踪变量

▷ 更新模型参数

11: **结束循环**

12: **结束循环**

[a]理论上，要进行与异质性无关的分析，校正项中的这一初始化是必需的。我们稍后通过实验表明，在实践中只需选择$\mathbf{c}_i^{(0)}$即可取得良好效果。

3.2 主定理：非凸函数上的与数据无关的收敛

在本节中，我们给出 梯度跟踪的收敛速率 K。注意， p 是假设3中定义的网络参数 .

定理 3.2 (K-梯度跟踪收敛)。 对于算法 *1* 中的方案，若混合矩阵满足假设*3* 且误差任意 $\epsilon > 0$，则存在常数步长 $\eta_c = O(\frac{p}{KL})$ 和 $\eta_s = O(p)$，使得在假设 *1* 和 *2*下，对于 L-光滑（可能非凸）函数，有 $\frac{1}{T+1}\sum_t \mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 \leq \epsilon$经过

$$\mathcal{O}\left(\frac{\sigma^2}{Kn\epsilon^2} + \frac{\sigma}{p^2\sqrt{K}\epsilon^{\frac{3}{2}}} + \frac{1}{p^2\epsilon}\right) \cdot L$$

通信轮次。

4 讨论

本节将介绍并比较其他可能的方式，以便为 GT 引入与 K-GT 类似通信模式的本地步。

4.1 其他 GT 替代算法

4.1.1 周期通信梯度跟踪（周期性梯度跟踪）。

还有另一种方法可以将本地步纳入上述框架（2）。不同于通过固定拓扑$\mathbf{W}$进行通信，通信图会随时间变化，记为 $\mathbf{W}^{(t)}$。请注意，$\mathbf{W}^{(t)} = \mathbf{I}$，这意味着实际上没有通信。如果 $\mathbf{W}^{(t)}$ 在 $\{\mathbf{W}、\mathbf{I}\}$之间周期性切换，也会降低通信频率。具体细节见算法2（附录 A.1）。

K-GT 的噪声低于周期性梯度跟踪。周期性梯度跟踪的本地步可以改写为校正随机梯度下降，与 K-GT 的处理方式相同。但周期性梯度跟踪在通信时采用了不同的校正项更新方式，

$$\mathbf{C}^{(t+1)} = \mathbf{C}^{(t)}\mathbf{W} + \nabla F(\mathbf{X}^{(t)+K-1}; \xi^{(t)+K-1})(\mathbf{W} - \mathbf{I}). \tag{6}$$

这种改写的等价性已在附录 B.2.1 中证明。

然而，如果我们直接 重新改写方程（4），可得 K-梯度跟踪使用 K 随机梯度的平均值，即 ，

$$\mathbf{C}^{(t+1)} = \mathbf{C}^{(t)}\mathbf{W} + \frac{1}{K}\sum_k \nabla F(\mathbf{X}^{(t)+k}; \xi^{(t)+k})(\mathbf{W} - \mathbf{I}),$$

这可以将随机噪声降低 K。周期性梯度跟踪只使用一个随机梯度，因此会受到更大的随机噪声影响。

在随机梯度中使用更多随机样本可以降低周期性梯度跟踪中的噪声。 如果使用更多随机样本 6) $\{\xi^{(t),s} | s = 0, \dots, K-1\}$ 替换 $\nabla F(\mathbf{X}^{(t)+K-1};\ \xi^{(t)+K-1})$

表 2：达到目标精度所需通信轮次的比较 ϵ（针对非凸函数）。我们关于 K-梯度跟踪和周期性梯度跟踪的结果，在采用本地步时，按照异质性参数 $\bar{\zeta}^2$（定义见假设 4）改善了 D-SGD 的收敛速率，并加快了梯度跟踪的收敛速率。

本地步	算法	通信轮次
$K=1$	梯度跟踪 [11]	$\mathcal{O}\left(\frac{\sigma^2}{n\epsilon^2}+\frac{\sigma}{p^{\frac{3}{2}}\epsilon^{\frac{3}{2}}}+\frac{1}{p^2\epsilon}\right)$
$K>1$	D-SGD [12]	$\mathcal{O}\left(\frac{\sigma^2}{Kn\epsilon^2}+\left(\frac{\zeta}{p}+\frac{\sigma}{\sqrt{pK}}\right)\frac{1}{\epsilon^{\frac{3}{2}}}+\frac{1}{p\epsilon}\right)$
	K-梯度跟踪 **[我们的方法]**	$\mathcal{O}\left(\frac{\sigma^2}{Kn\epsilon^2}+\frac{\sigma}{p^2\sqrt{K}\epsilon^{\frac{3}{2}}}+\frac{1}{p^2\epsilon}\right)$
	周期性梯度跟踪 **[我们的方法]**	$\mathcal{O}\left(\frac{\sigma^2}{Kn\epsilon^2}+\frac{\sigma}{p^2\epsilon^{\frac{3}{2}}}+\frac{1}{p^2\epsilon}\right)$
	使用全梯度的周期性梯度跟踪 **[我们的方法]**	$\mathcal{O}\left(\frac{\sigma^2}{Kn\epsilon^2}+\frac{\sigma}{p^2\sqrt{K}\epsilon^{\frac{3}{2}}}+\frac{1}{p^2\epsilon}\right)$

with

$$\nabla F(\mathbf{X}^{(t)+K-1};\xi^{(t)+K-1})=\frac{1}{K}\sum_s \nabla F(\mathbf{X}^{(t)+K-1};\xi^{(t)+K-1,s}),$$

那么，校正项 $\mathbf{C}^{(t)}$ 的随机噪声水平与 K-梯度跟踪相同。然而，使用更多样本来计算随机梯度下降，比 K-梯度跟踪需要多得多的额外计算。

定理 4.1 (周期性梯度跟踪收敛)。 对于算法中的方案*2* (附录*A.1*)，采用假设 *3* 中的混合矩阵并考虑任意误差 $\epsilon>0$，存在一个常数步长 $\eta=O(\frac{p^2}{KL})$ 使得在假设*1*和*2*下，对于 L-光滑（可能非凸）函数，有 $\frac{1}{T+1}\sum_t \mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2\le\epsilon$ 经过

$$\mathcal{O}\left(\frac{\sigma^2}{Kn\epsilon^2}+\frac{\sigma}{p^2\epsilon^{\frac{3}{2}}}+\frac{1}{p^2\epsilon}\right)\cdot L$$

通信轮次。反之，如果考虑使用全批量跟踪算法*3*(附录 *A.1*)，则收敛速率可提升为

$$\mathcal{O}\left(\frac{\sigma^2}{Kn\epsilon^2}+\frac{\sigma}{p^2\sqrt{K}\epsilon^{\frac{3}{2}}}+\frac{1}{p^2\epsilon}\right)\cdot L.$$

请注意，后一个结果针对的是全批量跟踪，每个通信轮次都需要额外的计算开销（不同于 K-梯度跟踪）。

4.1.2 大批量梯度跟踪（大批量梯度跟踪）

除了本地训练，大批量训练在分布式环境中也很流行，可用于实现加速。与大批量随机梯度下降类似，我们使用 $\mathbf{G}^{(t)}=\sum_k \nabla F(\mathbf{X}^{(t)};\xi^{(t),k})$，在（2）中利用 K 独立同分布随机样本 $\{\xi^{(t),k}|k=0,\ldots,K-1\}$，并令 $\mathbf{W}^{(t)}=\mathbf{W}$。从理论上讲，可以改善达到所需准确率所需的渐近通信轮次 ϵ，从 $\mathcal{O}(\frac{\sigma^2}{n\epsilon^2})$ [11] 降至 $\mathcal{O}(\frac{\sigma^2}{nK\epsilon^2})$，同时保持与异质性无关。

我们通过实验表明，大批量梯度跟踪与异质性无关，且其通信性能与 K-梯度跟踪相同（图1、3）。

4.2 收敛性比较

我们在表2中总结了相关去中心化算法的收敛速率。为了分析其他依赖数据异质性的方法的收敛性，还需额外假设来衡量数据异质性 [9, 12]。

K-梯度跟踪在高噪声区域通过本地步实现加速。 当 ϵ 足够小时，噪声主导收敛速率（$\sigma>0$），且对于梯度跟踪、周期性梯度跟踪和 K-梯度跟踪，收敛速率不受图参数 p 影响。经过足够长的瞬态时间后，周期性梯度跟踪和 K-梯度跟踪在 $\mathcal{O}(\frac{\sigma^2}{nK\epsilon^2})$ 下，相较于收敛速率为 $\mathcal{O}(\frac{\sigma^2}{n\epsilon^2})$ 的梯度跟踪，通过 K 实现线性加速。此外，与 GT 基线相比，K-梯度跟踪的瞬态时间也会随 $\mathcal{O}(\frac{1}{\sqrt{K}})$ 而缩短。

在非渐近情形下，梯度跟踪方法通常比扩散方法（例如 D-SGD）对网络参数更为敏感。我们对 K-GT 的分析中，对网络参数 p 的依赖性比朴素

梯度跟踪更为严重。将我们的分析与同期工作中对梯度跟踪所做的更精细分析结合起来 [11]将是一个有趣的后续研究方向——不过，在这项工作中，我们专注于为梯度跟踪引入本地步。

对于 K-GT，数据异质性的影响可以消除。 K-GT 通常不能完全解决数据异质性问题，且其表现取决于初始点处的数据异质性；GT 也存在同样的情况。已有研究证明，在非渐近区间，GT 对初始点处数据异质性的依赖实际上会有所减弱 [11]。然而，只需针对初始迭代进行一轮全局通信（例如算法 1），即可从 GT、K-GT 和周期性梯度跟踪的复杂度估计中消除异质性。表 2 为简化呈现，省略了收敛速率中的初始化项。相反，异质性 $\bar{\zeta}^2$ 在任何情况下都无法从 D-SGD 中消除，并且会减缓其收敛速度。

与 K-GT 相比，周期性梯度跟踪受噪声的影响更大。 在渐近区间，K-GT$\mathcal{O}(\frac{\sigma}{\sqrt{K}})$ 的暂态时间会随本地步数增加而缩短，而周期性梯度跟踪 $\mathcal{O}(\sigma)$ 则不会。不过，如第 4.1.1 节所述，这一噪声项可以得到改善。如果在方程（6）中考虑全梯度，周期性梯度跟踪的表现与 K-GT 相近，但需要付出额外计算的代价。

5 实验结果

我们通过与 D-SGD 和周期性梯度跟踪进行比较，评估 K-梯度跟踪的有效性。

5.1 设置

我们在两种设置下进行实验。

1. 合成数据集：我们首先构建分布式最小二乘目标函数 $f_i(\mathbf{x}) = \frac{1}{2}||\mathbf{A}_i\mathbf{x} - \mathbf{b}_i||^2$，其中海森矩阵固定为 $\mathbf{A}_i^2 = \frac{i^2}{n} \cdot \mathbf{I}_d$，并为每个客户端 $i \in [n]$ 采样一个 $\mathbf{b}_i \sim \mathcal{N}(0, \frac{\bar{\zeta}^2}{i^2} \mathbf{I}_d)$，其中 $\bar{\zeta}^2$ 可以控制局部目标函数之间的偏差 [12]。随机噪声通过添加高斯噪声来控制，噪声为 $\sigma^2 = 1$。2. 真实-世界数据集，MNIST [3]：我们测试所有客户端协作训练卷积神经网络（CNN）的情形[4]，使用真实世界数据集 MNIST。该数据集总共包含 60,000 张图像，尺寸为28×28，并有 10 个标签。

两组实验均采用环形拓扑。为简便起见，所有实验中我们都不使用算法 1 中的初始化，而是初始化$\mathbf{c}_i = \mathbf{0}$。对于 MNIST 的数据划分，我们同时考虑同质和异质两种情况。同质数据集会先进行随机打乱，然后均匀划分给所有客户端。我们称之为“随机”设置。异质数据集的构造方式是，每个客户端只能访问某个类别子集。我们称之为“排序”情形，此时客户端之间的数据差异最大。我们使用 $n = 5$ 和 $n = 10$ 个客户端，分别对应每个客户端可访问一个类别和两个类别的情形；其中，$n = 10$ 这一情形的异质性比 $n = 5$ 更严重。

参数调优。 对于合成数据集，我们采用相同的学习率 η_s=1 和 η=1e-3。对于 MNIST，我们从 {0.5、0.1、0.05、0.01、0.005、0.001} 中为算法和每个客户端批量大小为 128 的设置调优，选取最佳的固定学习率。需要注意的是，尽管我们的算法采用固定学习率，但使用更复杂且随时间变化的学习率调度器肯定能带来更好的性能。

比较。 我们主要展示 K-GT 相较于基线 D-SGD 在收敛速率上的加速效果和鲁棒性。我们还考察了第 4 节中讨论的几种支持本地步的 GT 变体的性能。

5.2 数值结果

K-GT 对异质性的鲁棒性最强。 在凸情形下，如图 1 所示，只有 D-SGD 会出现客户端漂移；随着 $\bar{\zeta}^2 \neq 0$ 增大，D-SGD 最终得到的模型质量会变差。然而，K-GT、周期性梯度跟踪（使用或不使用全梯度）和大批量梯度跟踪都不会受到“客户端漂移”的影响，最终无论 $\bar{}$以及 ζ 和 K（本地步数或随机样本数）如何增加，模型质量都能保持在 一致的水平。在非凸情形下，由于最优性条件和优化轨迹比凸情形更复杂，因此当数据划分是随机的时，图 2 中所有方法的泛化性能都无法完全恢复到基线水平。然而，K-GT 在数据划分为非独立同分布时始终表现更优，而且如图 2b 所示，异质性程度越高，改进效果越显著。

[4]这里为了简化起见，我们只考虑一个非常简单的、不含批归一化层的网络 [7]，因为批归一化本身假设不同批次的数据分布均匀，而这并非我们关注的情况。详细的网络结构见附录D。

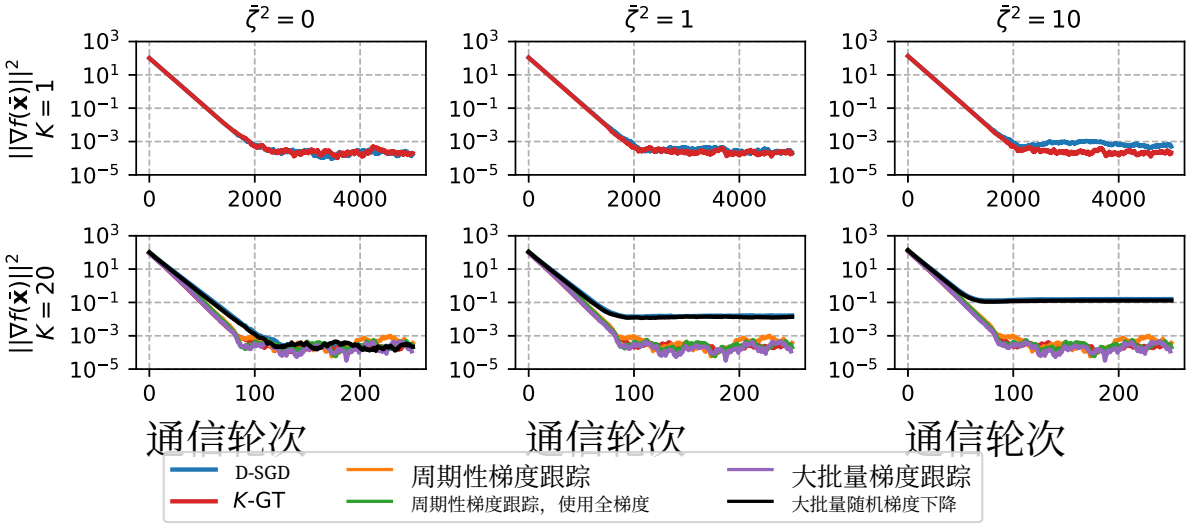
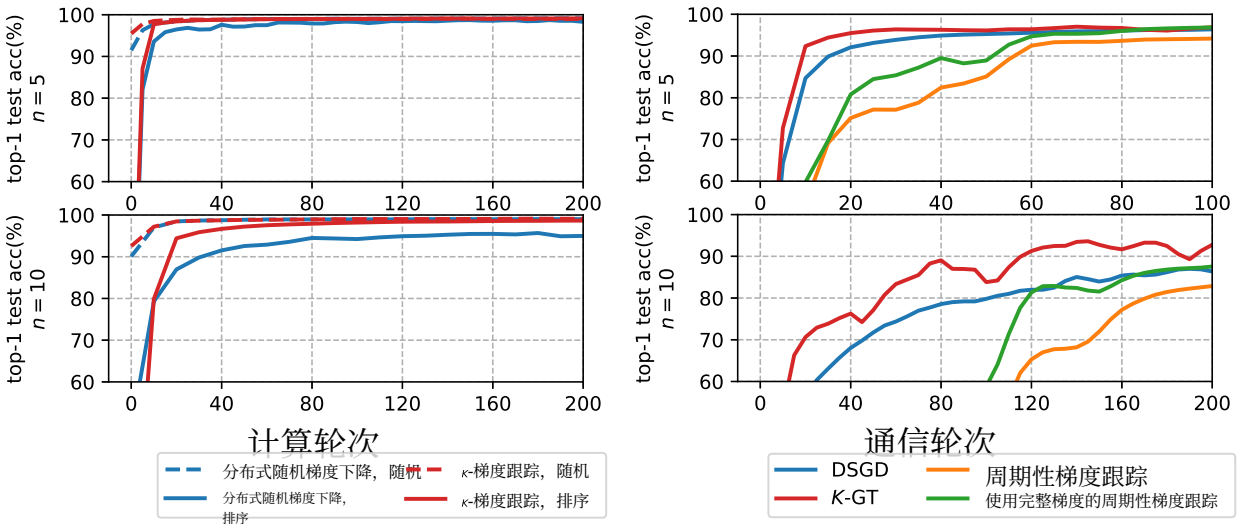


图 1：10 个客户端在环形网络上训练 **合成** 凸函数，并加入噪声 $\sigma^2 = 1$。对于 $K = 1$（上排），总共进行了 5000 个通信轮次；而对于 $K = 20$（下排），仅进行了 250 轮。所有方法使用相同的学习率，结果取三次重复实验的平均值。随着异质性增加（$\bar{\zeta}$ 变大），D-SGD 的客户端漂移更加严重，K 也同样如此。K-梯度跟踪、梯度跟踪以及使用完整梯度的梯度跟踪在不同的 $\bar{\zeta}^2$ 下表现一致，同时在 $K > 1$ 时减少了通信量。



(a) $K = 1$，数据采用随机或排序方式进行划分。

(b) $K = \frac{1}{n}$ 轮次，并且数据经过排序后进行划分。

图 2：在 **MNIST** 数据集上，D-SGD（蓝色）、K-GT（红色）、不使用（橙色）和使用（绿色）全梯度的 GT 在 $n = 5$（上排）和 $n = 10$（下排）个客户端上的泛化性能。横 x轴对应于 (a) 遍历整个数据集的次数（轮次）和 (b) 通信轮次。在 (a) 中，当 $K = 1$ 时，K-GT 和周期性梯度跟踪与 GT 基线相同（备注 1）。学习率经过调优以达到最佳效果。请注意，以批量大小为 128 遍历一轮全局数据集，相当于使用随机梯度下降进行 470 次计算。对于 (b)，由于 $nK = 470$ 在 $n = 5$（右上）和 $n = 10$（左上）中均固定，因此通信轮次之间的本地步数为 $K_{n=5} = 2K_{n=10}$。

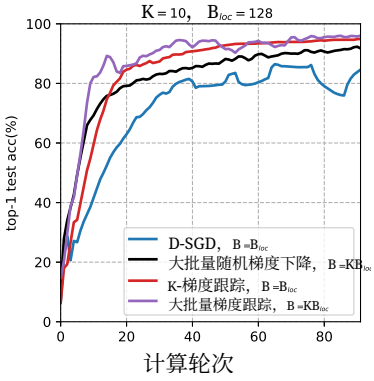


图 3：MNIST 上大批量训练与采用本地步的训练之间的泛化性能比较。大批量训练使用 KB_{loc} 本地批量大小，并在每次更新时进行通信；采用本地步的训练使用 B_{loc} 本地批量大小，并每隔 K 次本地更新定期通信。

本地步可减少通信。 在图 $K = 1$ 到 $K = 20$ （见图 1） K，-GT 和其他 GT 替代算法达到相同目标所需的轮次从 2000 轮减少到仅 100 轮，借助本地步实现了通信量的线性减少。然而，增加本地步数会使 D-SGD 的模型质量进一步下降。同时，将本地步引入非凸函数的训练仍可减少通信，但减少幅度并不像凸情形那样达到线性因子 K 。在图 2b中，我们固定每轮数据遍历的 $nK = 1$ ，使得无论是哪个 $n = 5$ 还是 $n = 10$ 客户端，都在计算 1 轮次后通信一次。与图 $K = 1$ （见图 1）相比，当 $K > 1$ 时，加速效果仍然非常显著。不过请注意，当数据划分具有异质性时，增加本地步数会导致更严重的质量损失，但 K-GT 仍然表现更优。

大批量梯度跟踪的性能与采用本地步的跟踪相近。 无论是凸函数（图1）还是非凸函数（图3），无论采用 K-梯度跟踪还是大批量梯度跟踪，在相同的通信轮次和相同的计算轮次后，都能达到相近的准确率水平。然而，对于 D-SGD，采用本地步进行训练甚至可能更加稳定，且比大批量训练具有更好的泛化能力，这一点已在实证研究中得到探讨 [20]。

6 结论

去中心化学习是推动深度学习普及的一项有前景的基础技术。尤其是在边缘人工智能应用中，用户的数据并不服从均匀分布。因此，去中心化学习算法需要具备应对数据异质性的鲁棒性。

我们提出了一种新颖的去中心化优化算法（K-GT），支持通信高效的本地更新步，并能克服数据差异。该跟踪机制使用累积梯度和，类似于动量，因此无需采用大批量大小，便可降低局部更新间的方差。我们通过收敛保证和实证评估，证明了 K-GT 的优越性。

参考文献

[1] S. Alghunaim 和 A. Sayed，原始-对偶梯度方法的线性收敛及其在分布式优化中的性能，Automatica 117 (2020)，第 109003 页。[2] Chen 等人，通过周期性全局平均加速谣传随机梯度下降，载于 ICML，2021 年。[3] L. Deng，用于机器学习研究的手写数字图像 MNIST 数据库，IEEE信号处理杂志 29 (2012)，第 141–142 页。[4] P. Goyal、P. Dollár、R.B. Girshick、P. Noordhuis、L. Wesolowski、A. Kyrola、A. Tulloch、Y. Jia 和 K. He，精确的大批量小批量随机梯度下降：1 小时内训练 ImageNet，arXiv 预印本 arXiv:1706.02677 (2017)。[5] F. Haddadpour、M.M. Kamani、M. Mahdavi 和 V. Cadambe，具有周期平均的本地 SGD：更紧的分析与自适应同步，载于神经信息处理系统大会，2019 年。[6] K. Hsieh、A. Phanishayee、O. Mutlu 和 P. Gibbons，去中心化机器学习中的非独立同分布数据困境，载于 ICML，2020 年。[7] S. I. 和 C. S.，批量归一化：通过减少内部协变量偏移加速深度网络训练，载于 ICML，2015 年。[8] P. Kairouz、H.B. McMahan、B. Avent、A. Bellet、M. Bennis、A.N. Bhagoji、K. Bonawitz、Z. Charles、G. Cormode、R. Cummings、R.G.L. D'Oliveira、H. Eichner、S.E. Rouayheb、D. Evans、J. Gardner、Z. Garrett、A. Gascón、B. Ghazi、P.B. Gibbons、M. Gruteser、Z. Harchaoui、C. He、L. He、Z. Huo、B. Hutchinson、J. Hsu、M. Jaggi、T. Javidi、G. Joshi、M. Khodak、J. Konečný、A. Korolova、F. Koushanfar、S. Koyejo、T. Lepoint、Y. Liu、P. Mittal、M. Mohri、R. Nock、A. Özgür、R. Pagh、M. Raykova、H. Qi、D. Ramage、R. Raskar、D. Song、W. Song、S.U. Stich、Z. Sun、A.T. Suresh、F. Tramèr、P. Vepakomma、J. Wang、L. Xiong、Z. Xu、Q. Yang、F.X. Yu、H. Yu 和 S. Zhao，联邦学习的进展与开放问题，Foundations and Trends® in Machine Learning 14 (2021)，第 1–210 页。[9] S.P. Karimireddy、S. Kale、M. Mohri、S. Reddi、S.U. Stich 和 A.T. Suresh，SCAFFOLD：用于联邦学习的随机控制平均，载于 ICML，2020 年。[10] A. Koloskova、T. Lin、S.U. Stich 和 M. Jaggi，采用任意通信压缩的去中心化深度学习，ICLR (2020)。[11] A. Koloskova、T. Lin 和 S. Stich，去中心化机器学习中梯度跟踪的改进分析，神经信息处理系统大会 (2021)。[12] A. Koloskova、N. Loizou、S. Boreiri、M. Jaggi 和 S. Stich，具有变化拓扑和本地更新的去中心化 SGD 统一理论，载于 ICML，2020 年。[13] A. Koloskova、S. Stich 和 M. Jaggi，采用压缩通信的去中心化随机优化与谣传算法，载于 ICML，2019 年。[14] J. Konečnỳ、H.B. McMahan、F.X. Yu、P. Richtárik、A.T. Suresh 和 D. Bacon，联邦学习：提高通信效率的策略，arXiv 预印本 arXiv:1610.05492 (2016)。[15] B. Li、S. Cen、Y. Chen 和 Y. Chi，采用梯度跟踪与方差缩减的网络通信高效分布式优化，载于 AISTATS，2020 年。[16] M. Li、D.G. Andersen、A.J. Smola 和 K. Yu，采用参数服务器的通信高效分布式机器学习，载于神经信息处理系统大会，2014 年。[17] T. Li、A.K. Sahu、M. Zaheer、M. Sanjabi、A. Talwalkar 和 V. Smith，异构网络中的联邦优化，载于机器学习与系统会议，2020 年。[18] 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案例研究，载于神经信息处理系统大会，2017 年。[19] X. Lian、W. Zhang、C. Zhang 和 J. Liu，异步去中心化并行随机梯度下降，载于 ICML，2018 年。[20] T. Lin、S.U. Stich、K.K. Patel 和 M. Jaggi，别用大批量小批量，改用本地 SGD，载于 ICLR，2020 年。[21] Y. Liu，异构数据上的去中心化训练方差缩减，硕士论文，苏黎世联邦理工学院，2021 年。可于 https://pub.tik.ee.ethz.ch/students/2020-HS/MA-2020-32.pdf 获取。

[22] P.D. Lorenzo 和 G. Scutari，NEXT：网络内非凸优化，《IEEE 网络信号与信息处理汇刊》2（2016），第 120–136 页。[23] B. McMahan、E. Moore、D. Ramage、S. Hampson 和 B.A.y Arcas，基于去中心化数据的深度网络通信高效学习，载于 AISTATS。PMLR，2017，第 1273–1282 页。[24] A. Nedi´c、A. Olshevsky 和 W. Shi，在时变图上实现分布式优化的几何收敛，《SIAM 优化期刊》27（2017），第 2597–2633 页。[25] A. Nemirovski、A. Juditsky、G. Lan 和 A. Shapiro，随机规划的稳健随机逼近方法，《SIAM 优化期刊》19（2009），第 1574–1609 页。[26] E.D.H. Nguyen、S.A. Alghunaim、K. Yuan 和 C.A. Uribe，关于带本地更新的梯度跟踪性能（2022）。参见 https://arxiv.org/abs/2210.04757。[27] S. Pu 和 A. Nedi´c，一种分布式随机梯度跟踪方法，载于 ICDC。2018。[28] W. Shi、Q. Ling、G. Wu 和 W. Yin，EXTRA：用于去中心化共识优化的精确一阶算法，《SIAM 优化期刊》25（2015），第 944–966 页。[29] S.U. Stich，本地 SGD 收敛快速且通信开销小，载于 ICLR。2019。[30] H. Tang、X. Lian、M. Yan、C. Zhang 和 J. Liu，D^2：基于去中心化数据的去中心化训练，载于 ICML。2018。[31] J.N. Tsitsiklis，去中心化决策与计算中的问题。技术报告，麻省理工学院剑桥信息与决策系统实验室，1984。[32] 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》68（2020），第 6255–6271 页。[33] Y. You、Z. Zhang、C.J. Hsieh、J. Demmel 和 K. Keutzer，数分钟内训练 ImageNet，载于 ICPP。2018。[34] K. Yuan、W. Xu 和 Q. Ling，原始方法能否在去中心化动态优化中优于原始-对偶方法？《IEEE 信号处理汇刊》68（2020），第 4466–4480 页。[35] J. Zhang 和 K. You，用于非凸经验风险最小化的去中心化随机梯度跟踪，arXiv 预印本 arXiv:1909.02712（2019）。

A 算法

A.1 周期性算法

算法 2周期性梯度跟踪：采用周期性通信的梯度跟踪

参数：
T：通信次数；K：本地步数；η_s, η_c：通信、本地步长；$\mathbf{W}$：给定拓扑。
初始化： $\mathbf{x}_i^0 = \mathbf{x}_j^0$, $\mathbf{z}_i^0 = \frac{1}{n}\sum_i F_i(\mathbf{x}^0; \xi_i) = \mathbf{z}_j^0, \quad \forall\, i,\ \ j \in [n]$[a]
对于节点 $i \in \{1, ..., n\}$ 并行 **执行**
  对于通信： $t \leftarrow 0$ 至 $T-1$ **执行**
    对于本地步： $k \leftarrow 0$ 至 $K-2$ **执行**
      $\mathbf{x}_i^{(t)+k+1} = \mathbf{x}_i^{(t)+k} - \eta_c \mathbf{z}_i^{(t)+k}$
      $\mathbf{z}_i^{(t)+k+1} = \mathbf{z}_i^{(t)+k} + \nabla F_i(\mathbf{x}_i^{(t)+k+1}; \xi_i^{(t)+k+1}) - \nabla F_i(\mathbf{x}_i^{(t)+k}; \xi_i^{(t)+k})$
    结束循环
    $\mathbf{x}_j^{(t)+K} = \mathbf{x}_i^{(t)+K-1} - \eta_c \mathbf{z}_i^{(t)+K-1}$
    $\mathbf{x}_i^{(t+1)} = \sum_j w_{ij}\left(\mathbf{x}_j^{(t)} - \eta_s(\mathbf{x}_j^{(t)} - \mathbf{x}_j^{(t)+K})\right)$
    $\mathbf{z}_i^{(t+1)} = \sum_j w_{ij}\mathbf{z}_j^{(t)+K-1} + \nabla F_i(\mathbf{x}_i^{(t+1)}; \xi_i^{(t+1)}) - \nabla F_i(\mathbf{x}_i^{(t)+K-1}; \xi_i^{(t)+K-1})$
  结束循环
结束循环

[a]跟踪变量$\mathbf{z}_i^{(0)}$ 的这一初始化是理论上进行与异质性无关分析所必需的。事实上，我们稍后通过实验表明，$\mathbf{z}_i^{(0)} = \nabla F_i(\mathbf{x}^0; \xi)$在实践中效果良好。

算法3采用全批量梯度的周期性梯度跟踪

参数： T：**通信次数；** K：**本地步数；** η_s, η_c：**通信、本地步长；** $\mathbf{W}$：**给定的拓扑。** **初始化：** $\mathbf{x}_i^0 = \mathbf{x}_j^0$, $\mathbf{c}_i^0 = -\nabla F_i(\mathbf{x}^0; \xi_i) + \frac{1}{n}\sum_j \nabla F_j(\mathbf{x}^0; \xi_j)$[a] **对节点** $i \in \{1, ..., n\}$ **并行执行** **对通信轮次：** **从** $t \leftarrow 0$ **到** $T-1$ **执行** **对本地步：** **从** $k \leftarrow 0$ **到** $K-2$ **执行** $\mathbf{z}_i^{(t)+k} = \nabla F_i(\mathbf{x}_i^{(t)+k}; \xi_i^{(t)+k}) + \mathbf{c}_i^{(t)}$ $\mathbf{x}_i^{(t)+k+1} = \mathbf{x}_i^{(t)+k} - \eta_c \mathbf{z}_i^{(t)+k}$ **结束循环** **计算** $\mathbf{x}_i^{(t)+K-1}$ **上的全梯度，** $g_i = \nabla f_i(\mathbf{x}_i^{(t)+K-1})$。 $\mathbf{x}_j^{(t)+K} = \mathbf{x}_i^{(t)+K-1} - \eta_c(g_i + \mathbf{c}_i^{(t)})$ $\mathbf{x}_i^{(t+1)} = \sum_j w_{ij}\left(\mathbf{x}_j^{(t)+K-1} - \eta_s(\mathbf{x}_j^{(t)} - \mathbf{x}_j^{(t)+K})\right)$ $\mathbf{c}_i^{(t+1)} = \sum_j w_{ij}\mathbf{c}_j^{(t)} + \sum_j w_{ij}g_j - g_i$ **结束循环** **结束循环**

[a]这种校正项$\mathbf{c}_i^{(0)}$ 初始化在理论上是进行与异质性无关的分析所必需的。事实上，我们将在后文通过实验表明，$\mathbf{c}_i^{(0)} = \mathbf{0}$ 在实践中效果良好。

B 命题证明

本节将证明前文讨论的命题。

B.1 K-梯度跟踪的跟踪性质

命题 B.1 (梯度和跟踪)。 定义$\mathbf{Z}^{(t)} = \frac{1}{K\eta_c}\left(\mathbf{X}^{(t)} - \mathbf{X}^{(t)+K}\right)$ 为通信轮次 t期间的跟踪变量。在通信轮次 K 的梯度跟踪中，模型 $\mathbf{X}^{(t)}$ 和跟踪变量 $\mathbf{Z}^{(t)}$ 的更新规则

可改写为 （$\eta = \eta_s \eta_c$）：

$$\begin{aligned} \mathbf{X}^{(t+1)} &= \left(\mathbf{X}^{(t)} - K\eta \mathbf{Z}^{(t)}\right)\mathbf{W}, \\ \mathbf{Z}^{(t+1)} &= \mathbf{Z}^{(t+1)}\mathbf{W} + \mathbf{G}^{(t+1)} - \mathbf{G}^{(t)}, \end{aligned} \tag{5}$$

其中，$\mathbf{G}^{(t)} = \frac{1}{K}\sum_k \nabla F(\mathbf{X}^{(t)+k}; \xi^{(t)+k})$ 表示本地步上的平均更新。

证明。 模型的更新方案 K-梯度跟踪已在方程（4）中给出。接下来，若定义$\mathbf{Z}^{(t)} = \frac{1}{K\eta_c}\left(\mathbf{X}^{(t)} - \mathbf{X}^{(t)+K}\right)$，以及 $\eta = \eta_s \eta_c$，经过简单的重新表述，即可推导出上面所示的方程组。

□

B.2 周期性梯度跟踪的重新表述

周期性梯度跟踪实际上是一种跳过通信的时变梯度跟踪。也就是说，当$\mathbf{W}^{(t)} = \mathbf{W}$ 在方程（2）中且 mod(t, K)=0 时；否则$\mathbf{W}^{(t)} = \mathbf{I}$ 不进行通信并执行本地步。

接下来，我们采用 K-梯度跟踪的记号，将第 t个通信轮次之后的第 k个本地步上的模型记为$\mathbf{X}^{(t)+k}$。跟踪变量$\mathbf{Z}^{(t)+k}$也采用相同的记号。以下各节将首先说明，周期性梯度跟踪可以等价地重新表述为校正随机梯度下降，并且在整个本地步期间校正项保持不变；然后给出校正项和模型的更新方案。

B.2.1 校正随机梯度下降

命题 B.2。 在本地步期间，本地跟踪变量可等价地改写为带校正项的随机梯度下降，即

$$\mathbf{Z}^{(t)+k+1} = \nabla F(\mathbf{X}^{(t)+k+1}; \xi^{(t)+k+1}) + \underbrace{\mathbf{Z}^{(t)+k} - \nabla F(\mathbf{X}^{(t)+k}; \xi^{(t)+k})}_{\mathbf{C}^{(t)+k}}.$$

并且在本地步期间，校正项 $\mathbf{C}$ 保持不变，即 $\mathbf{C}^{(t)+k+1} = \mathbf{C}^{(t)+k}$，　$\forall k \in \{0, ..., K-1\}$，且仅在每次通信时更新。

证明。 我们知道，本地模型使用 $\mathbf{Z}$ 而非 $\nabla F(\mathbf{X}; \xi)$ 进行更新。我们将 $\mathbf{Z}$ 相对于随机梯度下降的偏差定义为 $\mathbf{C}$。反证法假设每个本地迭代 $(t)+k$ 的偏差都不同。也就是说$\mathbf{C}^{(t)+k+1} \neq \mathbf{C}^{(t)+k}$.于是，对于每次本地更新，我们有

$$\begin{aligned} \mathbf{Z}^{(t)+k+1} &= \mathbf{Z}^{(t)+k} + \nabla F(\mathbf{X}^{(t)+k+1}; \xi^{(t)+k+1}) - \nabla F(\mathbf{X}^{(t)+k}; \xi^{(t)+k}) \\ \mathbf{Z}^{(t)+k+1} - \nabla F(\mathbf{X}^{(t)+k+1}; \xi^{(t)+k+1}) &= \mathbf{Z}^{(t)+k} - \nabla F(\mathbf{X}^{(t)+k}; \xi^{(t)+k}) \\ \mathbf{C}^{(t)+k+1} &= \mathbf{C}^{(t)+k}, \quad \forall k \in \{0, ..., K-1\} \end{aligned}$$

这与假设的事实$\mathbf{C}^{(t)+k+1} \neq \mathbf{C}^{(t)+k}$.矛盾

□

B.2.2 更新方案的重述

命题 B.3。 如果我们进一步考虑本地步和通信各自采用不同的步长，就可以将周期性梯度跟踪等价地改写如下，

- 本地步。 我们将本地步视为校正随机梯度下降。校正项 $\mathbf{C} \in \mathbb{R}^{d \times n}$ 表示本地更新与通信更新之间的差异。对于本地步，即 $\forall k \in \{0, ..., K-1\}$，$\mathbf{X}^{(t)+0} \equiv \mathbf{X}^{(t)}$，

$$\mathbf{X}^{(t)+k+1} = \mathbf{X}^{(t)+k} - \eta_c(\nabla F(\mathbf{X}^{(t)+k}; \xi^{(t)+k}) + \mathbf{C}^{(t)}), \tag{7}$$

  其中$\mathbf{C}^{(t)}$ 在所有本地步中均为常数。

- 通信。 随后，它会同步$\mathbf{X}$ 和 $\mathbf{C}$，

$$\begin{aligned} \mathbf{X}^{(t+1)} &= \left(\mathbf{X}^{(t)} - \eta_s(\mathbf{X}^{(t)} - \mathbf{X}^{(t)+K})\right)\mathbf{W} \\ \mathbf{C}^{(t+1)} &= \mathbf{C}^{(t)}\mathbf{W} + \nabla F(\mathbf{X}^{(t)+K-1}; \xi^{(t)+K-1})(\mathbf{W} - \mathbf{I}) \end{aligned} \tag{8}$$

证明。 根据命题 B.2，本地更新等价于校正随机梯度下降。注意，模型更新和通信使用了不同的步长 η_c 和 η_s。

修正 nC 根据命题 B.2，在本地步期间保持不变，因此考虑其在通信期间的更新 n.

$$\Leftrightarrow (\nabla F(\mathbf{X}^{(t+1)};\ \xi^{(t+1)})\ + \mathbf{C}^{(t+1)}) = \mathbf{Z}^{(t+K-1)}\mathbf{W} + \nabla F(\mathbf{X}^{(t+1)};\ \xi^{(t+1)}) - \nabla F(\mathbf{X}^{(t+K-1)};\ \xi^{(t+K-1)})$$
$$\Leftrightarrow \mathbf{C}^{(t+1)} = \mathbf{C}^{(t)}\mathbf{W} + \nabla F(\mathbf{X}^{(t+K-1)};\ \xi^{(t+K-1)})(\mathbf{W} - \mathbf{I})$$

□

C 定理证明

C.1 技术工具

本节主要介绍一些有助于收敛分析的分析工具 .

命题 C.1 (光滑性假设的推论 1)。 假设 *1* 蕴含 $\forall i$且 $\forall \mathbf{x}$、 $\mathbf{y} \in \mathbb{R}^d$,

$$||\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})|| \leq L||\mathbf{x} - \mathbf{y}||.$$

引理 C.2. 对于任意一组 n 向量 $\{a_i\}_{i=1}^n$, $a_i \in \mathbb{R}^d$, $||\frac{1}{n}\sum_i^n a_i||^2 \leq \frac{1}{n}\sum_i^n ||a_i||^2$.

引理 C.3. 对于给定的两个向量 $a, b \in \mathbb{R}^d$, $2\langle a, b\rangle \leq \alpha||a||^2 + \frac{1}{\alpha}||b||^2$, $\alpha > 0$, 这等价于 $||a+b||^2 \leq (1+\alpha)||a||^2 + (1+\frac{1}{\alpha})||b||^2$。

备注 2. 上述不等式对弗罗贝尼乌斯范数下的矩阵也成立。对于 $\mathbf{A}$, $\mathbf{B} \in \mathbb{R}^{d\times n}$, $||\mathbf{AB}||_F \leq ||\mathbf{A}||_F||\mathbf{B}||_2$.

引理 C.4 （方差上界）**.** 若存在 n 均值为零的随机变量 $\{\xi_i\}_{i=1}^n$，它们之间可能并不独立，但每个变量的方差都小于 σ^2，则其和的方差上界为 $\mathbf{E}||\sum_i \xi_i||^2 \leq n\sigma^2$.

证明。$\mathbf{E}||\sum_i \xi_i||^2 \leq \mathbf{E}\left(n\sum_i ||\xi_i||^2\right) \leq n\sigma^2$.

□

引理 C.5 (递归展开 [12])。 对于任意参数 $r_0 \geq 0$, $b \geq 0$, $e \geq 0$, $u \geq 0$ 存在常数步长$\eta \leq \frac{1}{u}$ 使得

$$\Psi_T := \frac{r_0}{T+1}\frac{1}{\eta} + b\eta + e\eta^2 \leq 2(\frac{br_0}{T+1})^{\frac{1}{2}} + 2e^{\frac{1}{3}}(\frac{r_0}{T+1})^{\frac{2}{3}} + \frac{ur_0}{T+1}$$

补充定义

在继续证明收敛定理之前，我们需要补充一组定义，用来描述所跟踪的各种误差。为简单起见，我们定义特殊矩阵$\mathbf{J} = \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}$，它可用于计算**平均矩阵**，$\mathbf{XJ} = \mathbf{X} = [\bar{\mathbf{x}}\ \bar{\mathbf{x}}\ \ldots\ \bar{\mathbf{x}}]$.

我们将**客户端方差**（或共识距离）定义为每个节点偏离其平均模型的程度：$\Xi_t = \frac{1}{n}\sum_i^n \mathbf{E}||\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}||^2$.

由于我们在 通信间隔内执行本地步，因此我们将本地进展定义为每个节点相对于 s
全局平均 起点的移动量，即**客户端漂移**:

- at k-th local step: $e_{k,t} := \frac{1}{n}\sum_i^n \mathbf{E}||\mathbf{x}_i^{(t)+k} - \bar{\mathbf{x}}^{(t)}||^2$
- accumulation of local steps: $\mathcal{E}_t := \sum_{k=0}^{K-1} e_{k,t} = \sum_{k=0}^{K-1}\frac{1}{n}\sum_i^n \mathbf{E}||\mathbf{x}_i^{(t)+k} - \bar{\mathbf{x}}^{(t)}||^2$

由于我们使用校正项更新模型，校正梯度将与全局更新的方向一致，而非本地更新的方向。校正项在每次通信时更新，并在本地步期间保持不变。我们将**该校正项的质量**定义为其对全局更新与本地更新之间真实偏差的近似程度，$\gamma_t = \frac{1}{\eta L^2}\mathbf{E}||\mathbf{C}^{(t)} + \nabla f(\bar{\mathbf{X}}^{(t)}) - \nabla f(\bar{\mathbf{X}}^{(t)})\mathbf{J}||_F^2$，其中$\mathbf{J} = \frac{1}{n}\mathbf{1}\mathbf{1}^T$。

C.2 收敛性分析

本节将给出定理 3.2 和定理4.1的证明。根据前面的分析， K-梯度跟踪和周期性梯度跟踪在本地步中等价于校正随机梯度下降，并且在通信过程中具有相似的模式。因此，我们可以在同一证明框架下对它们进行分析。

为了证明这些定理，我们首先在以下各节中给出客户端漂移、共识距离和校正项质量的递推关系。

界定客户端漂移

接下来，我们将考察本地步内取得的进展，即下一次通信前累积的模型更新。

引理 C.6. 假设节点 $\eta_c \leq \frac{1}{8KL}$ 的本地步长为 $\eta_s \geq 0$, 对于任意通信步长，我们可以将漂移界定为

$$\mathcal{E}_t \leq 3(K\Xi_t) + 12K^2\eta_c^2L^2(K\gamma_t) + 6K^2{\eta_c}^2(K\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2) + 3K^2{\eta_c}^2\sigma^2$$

证明. 首先，注意到 $K = 1, \mathcal{E}_t = \frac{1}{n}\mathbf{E}||\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}||_F^2$,

$$0 \leq \frac{2}{n}\mathbf{E}||\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}||_F^2 + 6{\eta_c}^2\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + 3{\eta_c}^2\sigma^2,$$

该不等式始终成立，因为右侧始终为正。因此，对于 $K = 1$，引理显然得证。接下来考虑 $K \geq 2$的情形，并且

$$\begin{aligned}
ne_{k,t} &:= \mathbf{E}||\mathbf{X}^{(t)+k} - \bar{\mathbf{X}}^{(t)}||_F^2 \\
&= \mathbf{E}||\mathbf{X}^{(t)+k-1} - \eta_c\Big(\nabla F(\mathbf{X}^{(t)+k-1}; \xi^{(t)+k}) + \mathbf{C}^{(t)}\Big) - \bar{\mathbf{X}}^{(t)}||_F^2 \\
&\leq (1 + \frac{1}{K-1})\mathbf{E}||\mathbf{X}^{(t)+k-1} - \bar{\mathbf{X}}^{(t)}||_F^2 + n{\eta_c}^2\sigma^2 \\
&\quad + K{\eta_c}^2\mathbf{E}||\nabla f(\mathbf{X}^{(t)+k-1}) - \nabla f(\bar{\mathbf{X}}^{(t)}) + \mathbf{C}^{(t)} + \nabla f(\bar{\mathbf{X}}^{(t)})(\mathbf{I} - \mathbf{J}) + \nabla f(\bar{\mathbf{X}}^{(t)})\mathbf{J}||_F^2 \\
&\leq \underbrace{(1 + \frac{1}{K-1} + 4K{\eta_c}^2L^2)}_{:=\mathcal{C}}\mathbf{E}||\mathbf{X}^{(t)+k-1} - \bar{\mathbf{X}}^{(t)}||_F^2 + 4K{\eta_c}^2L^2n\gamma_t + 2K{\eta_c}^2n\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + n{\eta_c}^2\sigma^2 \\
&\leq \mathcal{C}^k\mathbf{E}||\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}||_F^2 + \sum_{r=0}^{k-1}\mathcal{C}^r\Big(4K{\eta_c}^2L^2n\gamma_t + 2K{\eta_c}^2n\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + n{\eta_c}^2\sigma^2\Big)
\end{aligned}$$

若 $\eta_c \leq \frac{1}{8KL}$，则 $4K(\eta_cL)^2 \leq \frac{1}{16K} < \frac{1}{16(K-1)}$。由于 $\mathcal{C} > 1$，因此 $\mathcal{C}^k \leq \mathcal{C}^K \leq (1 + \frac{1}{K-1} + \frac{1}{16})^{(K-1)} - K \leq e^{1+\frac{1}{16}} \leq 3$，且 $\sum_r^{k-1}\mathcal{C}^r \leq K\mathcal{C}^K \leq 3K$. 我们可以在 k^{th} 本地步处重新写出客户端漂移的界，

$$ne_{k,t} \leq 3\Xi_t + 3K\Big(4K{\eta_c}^2L^2n\gamma_t + 2K{\eta_c}^2n\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + n{\eta_c}^2\sigma^2\Big) \tag{9}$$

显然，在不等式（9）中，右侧与时间步 $k \in [0,\ K)$ 无关。因此，本地步 $\mathcal{E}_t$ 内累积的进展可表示为

$$\mathcal{E}_t := \sum_{k=0}^{K-1} e_{k,t} \leq 3(K\Xi_t) + 12K^2{\eta_c}^2L^2(K\gamma_t) + 6K^2{\eta_c}^2(K\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2) + 3K^2{\eta_c}^2\sigma^2$$

□

共识距离

接下来，我们考察本地训练后，通信模型的共识距离如何在两次通信之间变化。

引理 C.7。 对于任意有效步长 $\eta = \eta_s\eta_c$，我们有关于 Ξ_t 的下降引理如下

$$\Xi_{t+1} \leq (1 - \frac{p}{2})\Xi_t + \frac{6K\eta^2L^2}{p}\mathcal{E}_t + \frac{6K^2\eta^2L^2}{p}\gamma_t + K\eta^2\sigma^2.$$

证明。 我们知道，两个通信轮次之间的更新如下，

$$\mathbf{X}^{(t+1)} = \Big(\mathbf{X}^{(t)} - K\eta\mathbf{Z}^{(t)}\Big)\mathbf{W},$$

其中 $\mathbf{Z}^{(t)} = \frac{1}{K}\sum_k \left(\nabla F(\mathbf{X}^{(t)+k};\xi^{(t)+k}) + \mathbf{C}^{(t)}\right)$ 那么，时刻（$t+1$）的共识距离可通过以下方式衡量：

$$
\begin{aligned}
n\Xi_{t+1} &= \mathbf{E}||\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t+1)}||_F^2 \\
&= \mathbf{E}||\Big(\mathbf{X}^{(t)} - \eta\sum_{k=0}^{K-1}(\nabla F(\mathbf{X}^{(t)+k};\xi^{(t)+k}) + \mathbf{C}^{(t)})\Big)(\mathbf{W}-\mathbf{J})||_F^2 \\
&\le (1-p)\mathbf{E}||\Big(\mathbf{X}^{(t)} - \eta\sum_{k=0}^{K-1}(\nabla f(\mathbf{X}^{(t)+k}) + \mathbf{C}^{(t)})\Big)(\mathbf{I}-\mathbf{J})||_F^2 + nK\eta^2\sigma^2 \\
&\le nK\eta^2\sigma^2 + (1+\alpha)(1-p)\mathbf{E}||\mathbf{X}^{(t)}(\mathbf{I}-\mathbf{J})||_F^2 \\
&\quad + (1+\frac{1}{\alpha})\eta^2\mathbf{E}||\sum_{k=0}^{K-1}\nabla f(\mathbf{X}^{(t)+k})(\mathbf{I}-\mathbf{J}) \pm K\nabla f(\bar{\mathbf{X}}^{(t)})(\mathbf{I}-\mathbf{J}) + K\mathbf{C}^{(t)}||_F^2 \\
&\underset{\alpha=\frac{p}{2},\frac{1}{p}\le 1}{\le} nK\eta^2\sigma^2 + (1-\frac{p}{2})\mathbf{E}||\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}||_F^2 + \frac{6}{p}\Big(K\eta^2L^2||\mathbf{I}-\mathbf{J}||^2\sum_{k=0}^{K-1}\mathbf{E}||\mathbf{X}^{(t)+k} - \bar{\mathbf{X}}^{(t)}||_F^2 \\
&\quad + K^2\eta^2\frac{L^2}{L^2}\mathbf{E}||f(\bar{\mathbf{X}}^{(t)})(\mathbf{I}-\mathbf{J}) + \mathbf{C}^{(t)}||_F^2\Big) \\
&\le (1-\frac{p}{2})n\Xi_t + \frac{6K\eta^2L^2}{p}n\mathcal{E}_t + \frac{6K^2\eta^2L^2}{p}n\gamma_t + nK\eta^2\sigma^2
\end{aligned}
$$

□

校正质量度量

现在我们来界定校正项的质量度量。校正项旨在描述通信时，本地梯度与全局理想平均模型$\mathbf{X}$ 的梯度之间的偏差。也就是说，校正项的质量度量定义为 $\gamma_t = \frac{1}{nL^2}\mathbf{E}||\mathbf{C}^{(t)} + \nabla f(\bar{\mathbf{X}}^{(t)}) - \nabla f(\bar{\mathbf{X}}^{(t)})\mathbf{J}||_F^2$，其中$\mathbf{J} = \frac{1}{n}\mathbf{1}\mathbf{1}^T$。

如何估计校正项 项，K-梯度跟踪与周期性梯度跟踪有不同的选项，具体讨论见第4.1节 .

引理 C.8。 对于任意有效步长 $\eta = \eta_s\eta_c \le \frac{\sqrt{p}}{\sqrt{6KL}}$ 周期性梯度跟踪中的 γ 下降引理如下：

$$
K\gamma_{t+1} \le (1-\frac{p}{2})K\gamma_t + \frac{24}{p}(Ke_{K-1,t}) + \frac{2}{p}\mathcal{E}_t + \frac{12K^2\eta}{p}K\eta\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + \frac{2K\sigma^2}{L^2},
$$

如果我们在 K-梯度跟踪的校正项中改为不使用本地步的平均值，则 γ 的下降引理如下：

$$
K\gamma_{t+1} \le (1-\frac{p}{2})K\gamma_t + \frac{30}{p}\mathcal{E}_t + \frac{12K^2\eta^2}{p}K\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + \frac{2\sigma^2}{L^2}.
$$

证明。 两个连续通信轮次之间的平均校正项满足

$$
\mathbf{C}^{(t+1)}\mathbf{J} = \mathbf{C}^{(t)}\mathbf{J} + \frac{1}{K\eta_c}(\mathbf{X}^{(t)} - \mathbf{X}^{(t)+K})(\mathbf{W}-\mathbf{I})\mathbf{J} = \mathbf{C}^{(t)}\mathbf{J}
$$

对于 K-梯度跟踪。我们假设校正项以任意值初始化，只要其全局平均值始终为零，即 $\mathbf{C}^{(t)}\mathbf{J} = \mathbf{C}^{(0)}\mathbf{J} = \mathbf{0}$。回顾 γ_t的定义，注意到

$$
(\mathbf{C}^{(t)} + \nabla f(\bar{\mathbf{X}}^{(t)}) - \nabla f(\bar{\mathbf{X}}^{(t)}\mathbf{J})\mathbf{J} = \mathbf{C}^{(t)}\mathbf{J} + \nabla f(\bar{\mathbf{X}}^{(t)})(\mathbf{J}-\mathbf{J}) = \mathbf{0}.
$$

容易验证，er o ca 具有相同的性质。 h

那么对于 K-梯度跟踪，质量度量的递推关系可表示如下：

$$
\begin{aligned}
nL^2\gamma_{t+1} &:= \mathbf{E}||\mathbf{C}^{(t+1)} + \nabla f(\bar{\mathbf{X}}^{(t+1)})(\mathbf{I}-\mathbf{J})||_F^2 \\
&= \mathbf{E}||\mathbf{C}^{(t)}\mathbf{W} + \frac{1}{K}\sum_{k=0}^{K-1}\nabla F(\mathbf{X}^{(t+k)};\xi^{(t+k)})(\mathbf{W}-\mathbf{I}) + \nabla f(\bar{\mathbf{X}}^{(t+1)})(\mathbf{I}-\mathbf{J})||_F^2 \\
&= \mathbf{E}||\Big(\mathbf{C}^{(t)} + \nabla f(\bar{\mathbf{X}}^{(t)})(\mathbf{I}-\mathbf{J})\Big)\mathbf{W} \\
&\quad + \Big(\frac{1}{K}\sum_{k=0}^{K-1}\nabla f(\mathbf{X}^{(t+k)}) - \nabla f(\bar{\mathbf{X}}^{(t)})\Big)(\mathbf{W}-\mathbf{I}) \\
&\quad + \Big(\nabla f(\bar{\mathbf{X}}^{(t+1)}) - \nabla f(\bar{\mathbf{X}}^{(t)})\Big)(\mathbf{I}-\mathbf{J})||_F^2 + \frac{n\sigma^2}{K} \\
&\le (1+\alpha)(1-p)nL^2\gamma_t + 2(1+\frac{1}{\alpha})\Big(||\mathbf{W}-\mathbf{I}||^2\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{E}||\nabla f(\mathbf{X}^{(t+k)}) - \nabla f(\bar{\mathbf{X}}^{(t)})||_F^2 \\
&\quad + ||\mathbf{I}-\mathbf{J}||^2\mathbf{E}||\nabla f(\bar{\mathbf{X}}^{(t+1)}) - \nabla f(\bar{\mathbf{X}}^{(t)})||_F^2\Big) + \frac{n\sigma^2}{K} \quad \text{(due to } ||\mathbf{W}-\mathbf{I}|| \le 2,\ ||\mathbf{I}-\mathbf{J}|| \le 1) \\
&\underset{\alpha=\frac{p}{2},\ \frac{1}{p}\le 1}{\le} (1-\frac{p}{2})nL^2\gamma_t + \frac{6}{p}\Big(\frac{4}{K}\sum_{k=0}^{K-1}L^2\mathbf{E}||\mathbf{X}^{(t+k)} - \bar{\mathbf{X}}^{(t)}||_F^2 + nL^2\mathbf{E}||\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}||^2\Big) + \frac{n\sigma^2}{K} \\
&\le (1-\frac{p}{2})n\gamma_t + \frac{6L^2}{pK}\Big(4n\mathcal{E}_t + 2K^2\eta^2L^2n\mathcal{E}_t + 2K^2\eta^2n(K\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2) + K^2\eta^2\sigma^2\Big) + \frac{n\sigma^2}{K}
\end{aligned}
$$

周期性梯度跟踪使用 $\nabla F(\mathbf{X}^{(t+K-1)};\ \xi^{(t+K-1)})$ 替代校正更新中的 $\frac{1}{K}\sum_{k=0}^{K-1}\nabla F(\mathbf{X}^{(t+k)};\ \xi^{(t+k)})$。通过几乎相同的分析，我们可以得到一个非常相似的等式。此外，如果在周期性梯度跟踪中用全梯度替代随机梯度，即 $\nabla f(\mathbf{X}^{(t+K-1)})$，则可以降低噪声项。

进一步地，如果 $\eta = \eta_s\eta_c \le \frac{\sqrt{p}}{\sqrt{6KL}}$，那么 $\frac{6K^2\eta^2L^2}{p} \le 1$ 从而完成证明。

□

注 3：由上述结果可知，周期性梯度跟踪的 γ_t 与 K-梯度跟踪的量化结果仅在随机噪声的系数上有所不同。使用全批量梯度可以将周期性梯度跟踪中的随机噪声改善到与 K-梯度跟踪相同的水平。

通信轮次之间的进展

我们研究如何界定通信轮次之间的进展。

引理 C.9. 我们可以将任意轮次 $t \ge 0$ 中，以及任意 $\eta = \eta_s\eta_c \ge 0$ 的通信之间的平均进展界定如下：

$$
\mathbf{E}||\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}||^2 \le 2K\eta^2L^2\mathcal{E}_t + 2K^2\eta^2\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + \frac{(K\eta)^2\sigma^2}{nK}\,.
$$

证明。 根据前面的分析，我们可以保证 $\frac{1}{n}\sum_i\mathbf{c}_i^{(t)} = \mathbf{0}$。那么，通信之间的平均进展可以改写为

$$
\begin{aligned}
\mathbf{E}||\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}||^2 &= \eta^2\mathbf{E}||\frac{1}{n}\sum_{i,k}\nabla F_i(\mathbf{x}_i^{(t)+k};\xi^{(t)+k}) + \frac{K}{n}\sum_i\mathbf{c}_i^{(t)}||^2 \\
&\le \frac{K\eta^2}{n}\sum_{i,k}2\mathbf{E}||\nabla f_i(\mathbf{x}_i^{(t)+k}) - \nabla f_i(\bar{\mathbf{x}}^{(t)})||^2 + 2K^2\eta^2\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + \frac{K\eta^2\sigma^2}{n} \\
&\le \frac{2K\eta^2L^2}{n}\sum_{i,k}\mathbf{E}||\mathbf{x}_i^{(t)+k} - \bar{\mathbf{x}}^{(t)}||^2 + 2K^2\eta^2\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + \frac{K\eta^2\sigma^2}{n}
\end{aligned}
$$

在第一个不等式中，注意到 K 随机变量 $\{\xi^{(t)+k}\}_{k=0}^{K-1}$ 在以通信为条件时(t) 彼此之间不一定独立，但根据假设 σ^2，其方差均小于2，因此我们可以应用引理C.4。接下来的不等式是通过反复应用三角不等式得到的。

□

非凸情形的下降引理

引理 C.10。当函数 f 是 L-光滑时，算法 $\bar{\mathbf{x}}^{(t)}$ 的迭代平均值*1* 和算法*2* 在常数步长$\eta_c < \frac{1}{4\eta_s KL}$下满足

$$\mathbf{E}f(\bar{\mathbf{x}}^{(t+1)}) - \mathbf{E}f(\bar{\mathbf{x}}^{(t)}) \leq -\frac{K\eta}{4}\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + \eta L^2\mathcal{E}_t + \frac{K\eta^2 L}{2n}\sigma^2$$

证明。 由于局部函数 $\{f_i(\mathbf{x})\}$ 根据假设 L是 -光滑的 1，显然可得全局函数 $f(\mathbf{x})$ 同样是 L-光滑的。

$$\begin{aligned}\mathbf{E}f(\bar{\mathbf{x}}^{(t+1)}) &= \mathbf{E}f\Big(\bar{\mathbf{x}}^{(t)} - \frac{\eta}{n}\sum_{i,k}(\nabla F_i(\mathbf{x}_i^{(t)+k};\xi_i^{(t)+k}) + \mathbf{c}_i^{(t)})\Big)\\ &\leq \mathbf{E}f(\bar{\mathbf{x}}^{(t)}) + \underbrace{\mathbf{E}\Big\langle\nabla f(\bar{\mathbf{x}}^{(t+1)}), -\frac{\eta}{n}\sum_{i,k}(\nabla F_i(\mathbf{x}_i^{(t)+k};\xi_i^{(t)+k}) + \mathbf{c}_i^{(t)})\Big\rangle}_{U} + \frac{L}{2}\mathbf{E}||\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}||^2\end{aligned}$$

根据我们之前的分析，我们知道 $\frac{1}{n}\sum_i \mathbf{c}_i^{(t)} = 0,\ \forall t \geq 0$对于$K$-梯度跟踪和周期性梯度跟踪（若采用所提出算法中指出的初始化）

$$\begin{aligned}U :&= \mathbf{E}\Big\langle\nabla f(\bar{\mathbf{x}}^{(t+1)}), -\frac{\eta}{n}\sum_{i,k}(\nabla F_i(\mathbf{x}_i^{(t)+k};\xi_i^{(t)+k}) + \mathbf{c}_i^{(t)})\Big\rangle\\ &= \mathbf{E}\Big\langle\nabla f(\bar{\mathbf{x}}^{(t)}),\ -\frac{\eta}{n}\sum_{i,k}\mathbf{E}_{\xi_i^{(t)+k}}\nabla F_i(\mathbf{x}_i^{(t)+k};\xi_i^{(t)+k})\Big\rangle\\ &= -K\eta\mathbf{E}\Big\langle\nabla f(\bar{\mathbf{x}}^{(t)}), \frac{1}{nK}\sum_{i,k}\nabla f_i(\mathbf{x}_i^{(t)+k}) - \nabla f(\bar{\mathbf{x}}^{(t)}) + f(\bar{\mathbf{x}}^{(t)})\Big\rangle\\ &= -K\eta\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + \frac{1}{nK}\sum_{i,k}K\eta\mathbf{E}\Big\langle\nabla f(\bar{\mathbf{x}}^{(t)}), \Big(\nabla f_i(\mathbf{x}_i^{(t)+k}) - \nabla f_i(\bar{\mathbf{x}}^{(t)})\Big)\Big\rangle\\ &\leq -\frac{K\eta}{2}\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + \frac{K\eta}{2nK}\sum_{i,k}\mathbf{E}||\nabla f_i(\mathbf{x}_i^{(t)+k}) - \nabla f_i(\bar{\mathbf{x}}^{(t)})||^2\\ &\leq -\frac{K\eta}{2}\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + \frac{KL^2\eta}{2nK}\sum_{i,k}\mathbf{E}||\mathbf{x}_i^{(t)+k} - \bar{\mathbf{x}}^{(t)}||^2\end{aligned}$$

然后再代入引理 C.9 关于$\mathbf{E}||\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}||^2$，可得

$$\mathbf{E}f(\bar{\mathbf{x}}^{(t+1)}) \leq \mathbf{E}f(\bar{\mathbf{x}}^{(t)}) + (-\frac{K\eta}{2} + K^2\eta^2 L)\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||_F^2 + (\frac{\eta L^2}{2} + K\eta^2 L^3)\mathcal{E}_t + \frac{K\eta^2 L}{2n}\sigma^2$$

那么，取值 $\eta \leq \frac{1}{4KL}$即可完成证明。 □

主递推式

我们首先构造一个势函数 $\mathcal{H}_t = \mathbf{E}f(\bar{\mathbf{x}}^{(t)}) - \mathbf{E}f(\mathbf{x}^{(\star)}) + A\frac{(K\eta_c)^3 L^4}{p\eta_s^2}\gamma_t + \frac{B}{6v^2}\frac{K\eta_c L^2}{p}\Xi_t$，其中常数 A、B 和 v 可通过以下引理得到。

引理 C.11 (关于 K-梯度跟踪的递推)。 对于满足 $\eta_s = O(\frac{p}{KL})$ 和 $\eta_c = O(p)$ 的算法 *1* 的任意有效步长，存在满足 A、 B、 v 的常数 $D > 0$ 和 $D_5 \geq 0$。于是我们得到以下递推关系

$$\mathcal{H}_{t+1} - \mathcal{H}_t \leq -DK\eta\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + \frac{D_5 L^2}{pK}(K\eta)^3\sigma^2 + \frac{L}{2nK}(K\eta)^2\sigma^2.$$

证明。 首先，根据先前对误差项 γ_t、 Ξ_t 和 $\mathcal{E}_t$的界，我们可以得到 K-梯度跟踪中 $\mathcal{H}_{t+1}$ 与$\mathcal{H}_t$ 之间差值的界，同时还代入了 $C > 0$ the

$$0 \leq -C\eta_c L^2\mathcal{E}_t + 3C(K\eta_c L^2\Xi_t) + 12C(K\eta_c)^3 L^4\gamma_t + C\frac{6(K\eta_c)^2 L^2}{\eta_s}K\eta\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 + 3C(K\eta_c)^3\frac{L^2\sigma^2}{K}.$$

于是，我们得到 K-梯度跟踪的不等式递推如下，

$$\begin{aligned}\mathcal{H}_{t+1}-\mathcal{H}_t \le & \underbrace{\Big(-\frac{A}{2}+B\frac{\eta_s^2}{v^2p^2}+12C\Big)}_{\le D_1}(K\eta_c)^3L^4\gamma_t \\ & +\underbrace{\Big(-\frac{B}{12v^2}+3C\Big)}_{\le D_2}K\eta_cL^2\Xi_t \\ & +\underbrace{\Big(A\frac{30(K\eta_cL)^2}{p^2K}+B\frac{K^2\eta^2L^2}{v^2p^2}-C+\eta_s\Big)}_{\le D_3}\eta_cL^2\mathcal{E}_t \\ & +\underbrace{\Big(-\frac{1}{4}+A\frac{12\eta_s(K\eta_c)^4L^4}{p}+C\frac{6K^2{\eta_c}^2L^2}{\eta_s}\Big)}_{\le D_4}K\eta\mathbf{E}||\nabla f(\bar{x}^{(t)})||^2 \\ & +\underbrace{\Big(A\frac{2L^2}{pK}+B\frac{\eta_s^2L^2}{6v^2pK}+C\frac{3L^2}{K}\Big)}_{\le \frac{D_5L^2}{pK}}(K\eta_c)^3\sigma^2+\frac{L}{2nK}(K\eta)^2\sigma^2\end{aligned}$$

只要 $\eta_c \le \frac{p}{96vKL}$，~~$\eta_s = v\cdot p \to \eta_c\eta_s \le \eta = \frac{p}{96KL}$~~ 且 $A = 72v^3p+48vp,\ B = 36v^3p,\ C = vp$ 就存在常数 $v>1$，使得 $D_1,\ D_2,\ D_3 \le 0, D_4 < 0$。并且 $D_4 \le -D < 0$，以及 $D_5 \ge 0$，从而完成证明。 □

~~**引理 C.12** (周期性梯度跟踪的递推)。 对于满足 $\eta_s = O(\frac{p}{KL})$ 和 $\eta_c = O(p)$的算法 *1* 的任意有效步长，存在常数 $A,\ B,\ v$，且满足 $D>0$ 和 $D_5 \ge 0$。于是我们得到递推关系~~

$$\mathcal{H}_{t+1}-\mathcal{H}_t \le -DK\eta\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2+\frac{D_5L^2}{p}(K\eta)^3\sigma^2+\frac{L}{2nK}(K\eta)^2\sigma^2.$$

证明。 与 K-梯度跟踪相比，周期性梯度跟踪的不等式集合仅在随机噪声方面有所不同。然后按照与 K-梯度跟踪相同的原理应用，可得其势函数的递推关系如下

$$\begin{aligned}\mathcal{H}_{t+1}-\mathcal{H}_t \le & D_1(K\eta_c)^3L^4\gamma_t+D_2K\eta_cL^2\Xi_t+D_3\eta_cL^2\mathcal{E}_t+D_4K\eta\mathbf{E}||\nabla f(\bar{x}^{(t)})||^2 \\ & +\underbrace{\Big(A\frac{2L^2}{p}+B\frac{\eta_s^2L^2}{6v^2pK}+C\frac{3L^2}{K}\Big)}_{\le \frac{D_5L^2}{p}}(K\eta_c)^3\sigma^2+\frac{L}{2nK}(K\eta)^2\sigma^2-\frac{C}{2}Ke_{K-1,t}\end{aligned}$$

分析的其余部分可以参考引理C.11。 □

注释 4。 使用全梯度改进周期性梯度跟踪，其递推式与 K-梯度跟踪相同。

求解主递推式

以 K-梯度跟踪为例。对势函数进行望远镜求和，可得

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\Big(\mathcal{H}_{t+1}-\mathcal{H}_t\Big) &= \frac{1}{T+1}\Big(\mathcal{H}_{T+1}-\mathcal{H}_0\Big) \\ &\underset{\eta=\eta_s\eta_c}{\le} -DK\eta\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2+\frac{D_5L^2}{pK\eta_s^3}(K\eta)^3\sigma^2+\frac{L}{2nK}(K\eta)\sigma^2 \\ \underset{\eta_s=v\cdot p}{\Rightarrow} \frac{1}{T+1}\sum_{t=0}^{T}\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 &\le \frac{\mathcal{H}_0-\mathcal{H}_{T+1}}{(T+1)D}\frac{1}{K\eta}+\frac{L\sigma^2}{2nKD}(K\eta)+\frac{D_5L^2\sigma^2}{v^3p^4KD}(K\eta)^2\end{aligned}$$

不失一般性，我们假设 $f(\mathbf{x})$ 非负。因此，我们可以忽略 $-\mathcal{H}_{T+1}$ 的影响。

引理 C.13。 存在一个常数步长，使得

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 = \mathcal{O}\Big(\sqrt{\frac{\sigma^2 L\mathcal{H}_0}{nKT}} + (\frac{\sigma L\mathcal{H}_0}{p^2 KT})^{\frac{2}{3}} + \frac{L\mathcal{H}_0}{p^2 T}\Big).$$

证明。 非负序列 $\{\mathcal{H}_t\}_{t=0}^{T+1}$和 $\{\mathbf{E}||\nabla f(\bar{\mathbf{x}}^t)||\}_{t=0}^{T}$在 $K\eta$和 $(K\eta)^2$前均具有正系数，并满足引理 C.5中的条件。然后我们利用引理 C.5调整步长。于是，梯度累积的平均值可以上界为

$$\begin{aligned}\Rightarrow \frac{1}{T+1}\sum_{t=0}^{T}\mathbf{E}||\nabla f(\bar{\mathbf{x}}^{(t)})||^2 &\leq \mathcal{O}\Big(2(\frac{\frac{L\sigma^2}{2nK}\mathcal{H}_0}{T+1})^{\frac{1}{2}} + 2(\frac{L^2\sigma^2}{p^4 K})^{\frac{1}{3}}(\frac{\mathcal{H}_0(\mathbf{x})}{T+1})^{\frac{2}{3}} + \frac{\mathcal{H}_0(\mathbf{x})}{K\eta_{max}(T+1)}\Big)\\ &= \mathcal{O}\Big(\sqrt{\frac{\sigma^2 L\mathcal{H}_0}{nKT}} + (\frac{\sigma L\mathcal{H}_0}{p^2 KT})^{\frac{2}{3}} + \frac{L\mathcal{H}_0}{p^2 T}\Big)\end{aligned}$$

□

那么，收敛速率取决于势函数 $\mathcal{H}_0$的初始值。根据引理 C.11中势函数的定义， $\mathcal{H}_0$是 $f(\mathbf{x}^0)$、$\mathbf{E}||\mathbf{X}^{(0)} - \bar{\mathbf{X}}^{(0)}||_F^2$和 $\mathbf{E}||\mathbf{C}^{(0)} + \nabla f(\bar{\mathbf{X}}^{(0)})(-\mathbf{J}+\mathbf{I})||_F^2$初始值的组合。

我们假设每个节点都保证以相同的模型 $\mathbf{x}^{(0)} = \mathbf{x}_i^{(0)}$, $\forall i \in [n]$ 初始化。这样我们很容易得到 $\mathbf{E}||\mathbf{X}^{(0)} - \bar{\mathbf{X}}^{(0)}||_F^2 = 0$。如果我们将校正项初始化为 $\mathbf{c}_i^{(0)} = -\nabla f_i(\mathbf{x}^{(0)}) + \frac{1}{n}\sum_i \nabla f_i(\mathbf{x}^{(0)})$，那么 $\mathbf{E}||\mathbf{C}^{(0)} + \nabla f(\mathbf{X}^{(0)})(-\mathbf{J} + \mathbf{I})||_F^2 = 0$。

$$\begin{aligned}\mathcal{H}_0 &= f(\mathbf{x}^0) - f(\mathbf{x}^\star) + A\frac{(K\eta_c)^3 L^4}{p\eta_s^2}\gamma_0 + \frac{B}{6v^2}\frac{K\eta_c L^2}{p}\Xi_0\\ &= f(\mathbf{x}^0) - f(\mathbf{x}^\star) := F_0\end{aligned} \tag{10}$$

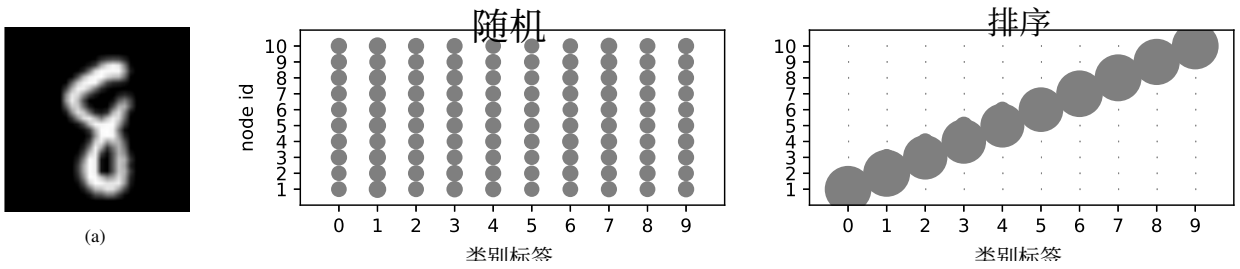
因此，对于任意准确率误差 $\epsilon > 0$，达到目标精度所需的通信轮次上界为

$$T \leq \mathcal{O}\Big(\frac{\sigma^2}{nK}\frac{1}{\epsilon^2} + \frac{\sigma}{p^2\sqrt{K}}\frac{1}{\epsilon^{\frac{3}{2}}} + \frac{1}{p^2}\frac{1}{\epsilon}\Big)\cdot LF_0,$$

由此完成了定理 3.2 关于 K-梯度跟踪的证明。定理 4.1 关于周期性梯度跟踪的证明也可以用相同原理轻松推导得出。

D 实验细节

D.1 基准数据集可视化



(a)

(b)

图4：数据可视化。(a) MNIST 数据集中的示例。(b) 有 $n = 10$ 个分布式节点时，随机和排序情形下每个节点的数据划分。点大小表示分配给每个节点的每类样本数量。

我们展示了一个来自 MNIST 数据集的图像示例，以及不同标签的数据在随机和排序情形下的划分方式。显然，在随机情形下，不同标签的数据被随机且均匀地分配到各个节点；而在排序情形下，每个节点只包含一个标签的图像，并且各节点获得的标签互不重叠。

D.2 模型结构

对于我们的非凸实　验，我们使用一个 4 层卷积神经网络（CNN），其详细信息列
　　于表3中。

表3：基准实验的模型架构。对于卷积层（二维卷积），我们按输入和输出维度、卷积核大小及步幅的顺序列出参数。对于最大池化层（二维最大池化），我们列出卷积核大小和步幅。对于全连接层（FC），我们列出输入和输出维度。对于随机失活（Dropout），我们列出概率参数。

层	详情
1	二维卷积(1, 10, 5, 1), 二维最大池化(2), ReLU
2	二维卷积(10, 10, 5, 1), 二维随机失活(0.5), 二维最大池化(2), ReLU
3	全连接层(320, 50), ReLU
4	全连接层(50, 10)