arXiv:2302.06103v1 [cs.LG] 13 Feb 2023

FedDA：通过重启式对偶平均实现更快速的局部自适应梯度方法框架

JunyiLi*、FeihuHuang†、HengHuang ‡

摘要

联邦学习（FL）是一种应对大规模分布式数据的新兴学习范式。在联邦学习中，一组客户端在服务器的协调下共同执行一项机器学习任务。FedAvg 算法是解决联邦学习问题最广泛使用的方法之一。在 FedAvg 中，学习率是固定的，而不是自适应变化的。自适应梯度方法相比固定学习率调度表现出更优的性能；然而，目前仍没有通用框架能够将自适应梯度方法纳入联邦学习场景。本文提出了**FedDA**，一种用于局部自适应梯度方法的新型框架。该框架采用重启式对偶平均技术，并且能够灵活结合各种梯度估计方法和自适应学习率形式。具体而言，我们分析了**FedDA-MVR**这一框架实例，并证明其在寻找驻点 ϵ 时可实现梯度复杂度 $\tilde{O}(\epsilon^{-1.5})$ 和通信复杂度 $\tilde{O}(\epsilon^{-1})$。这一结果达到了已知一阶联邦学习算法的最佳速率，而**FedDA-MVR** 是首个达到该速率的自适应联邦学习算法。我们还进行了大量数值实验，以验证所提方法的有效性。

1 引言

联邦学习是指分布式部署的一组客户端在中央服务器的协调下，利用各自持有的私有数据共同执行机器学习任务的过程。联邦学习中一种广泛使用的方法是 FedAvg（本地随机梯度下降） [31] 算法。顾名思义，FedAvg 在每个客户端上执行（随机）梯度下降步骤，并定期对局部状态求平均。当客户端的数据分布同质或异质性有界时，可以证明该方法能够收敛 [37, 14, 43] 。近年来，大量研究致力于加速 FedAvg。特别是，许多研究采用基于动量的方法来加速联邦学习，并在降低梯度复杂度和通信复杂度方面取得了显著进展 [8, 18, 22]。然而，另一类重要方法——自适应梯度方法——受到的关注要少得多，目前仍缺乏将自适应梯度方法纳入联邦学习场景的通用框架。

*美国匹兹堡匹兹堡大学电气与计算机工程系。电子邮箱：juny- ili.ai@gmail.com

†中国南京南京航空航天大学计算机科学与技术学院；同时任职于中国南京工业和信息化部模式分析与机器智能重点实验室。电子邮箱：huangfeihu2018@gmail.com

‡美国匹兹堡匹兹堡大学电气与计算机工程系。电子邮箱：henghuanghh@gmail.com

Adagrad [10], Adam [24] 和 AMSGrad [35] 等自适应梯度方法在非分布式环境中得到了广泛应用。梯度下降法使用固定学习率或固定的学习率调度。相比之下，自适应梯度方法将学习率设为与梯度幅值成反比，从而能够结合问题的局部曲率结构。自适应梯度方法在实践中表现出色；同时，它们还具有有益的理论性质，使其优于标准梯度下降法 [10, 13]。例如，近期一项研究 [36] 表明，自适应梯度有助于逃离鞍点。此外，一些研究 [27, 3] 表明，自适应梯度能够提升模型的泛化性能。

自适应梯度方法可视为一类广义镜像下降 [15] 方法，其关联的镜像映射根据自适应学习率定义。然而，镜像映射是动态的，并且在每个训练步骤都会发生变化。作为一个特例，梯度下降法可视为一种镜像下降法，其镜像映射为 L_2 距离函数。按照镜像下降文献中的惯例，我们将参数空间称为原始空间，将梯度空间称为对偶空间。在自适应梯度方法中，原始空间和对偶空间不同；而在梯度下降法中，二者重合。我们可以利用这种原始-对偶视角来理解现有的联邦学习算法并设计新算法。FedAvg 实际上利用了对偶状态平均值的作用。在 FedAvg 中，梯度平均值近似于在客户端状态平均点处计算的真实梯度，其近似误差以上界由客户端状态之间的差异给出；因此，客户端可以在不进行通信的情况下执行多个本地步。尽管在 FedAvg 中，由于原始空间和对偶空间相同，我们不对二者加以区分，但对于自适应梯度方法，对偶状态平均值与原始状态平均值并不等价。

现有文献中的联邦自适应梯度方法，要么只在服务器端执行自适应梯度步骤，要么在支持本地自适应梯度步骤时忽视原始-对偶这一细微差别。早期的一项工作是 [34], 作者提出在服务器平均步骤中应用自适应梯度，同时在本地执行常规梯度下降更新。该方法易于实现，且性能优于 FedAvg，但本地更新过程中并未利用自适应信息，削弱了自适应梯度的作用。近期的一些工作 [16, 4] 在本地更新步骤中利用了自适应信息；然而，这些方法有一个共同特点：在同步步骤中对问题的原始状态（参数）求平均。这会引发一些问题。首先，由于自适应学习率定义了镜像映射，在本地更新自适应学习率会导致对偶空间彼此不对齐，因此我们无法直接对原始状态求平均。其次，即使本地固定自适应学习率，原始空间相对于对偶空间也可能是非线性的关于 对偶空间，例如，在求解约束优化问题时。总之，我们提出了在联邦学习中应用自适应梯度的两项原则。第一，各本地对偶空间应彼此对齐；第二，我们应对对偶状态求平均。

更具体地说，我们提出了联邦学习自适应梯度框架 **FedDA**，它是 **Federated Dual-averaging Adaptive-gradient**（联邦对偶平均自适应梯度）的缩写。在 **FedDA** 的每个全局轮次中，客户端在本地聚合梯度（对偶状态），服务器则在同步步骤中对各客户端的对偶状态取平均。本地权重（原始状态）用作本地更新中的梯度查询点，并通过逆镜像映射（由自适应梯度定义）恢复。全局原始状态基于平均对偶状态和逆镜像映射进行更新。此外，我们采用重启技术，确保所有客户端在本地更新期间共享同一对偶空间；具体而言，我们在每个全局轮次重新计算自适应梯度，并在本地更新中使用固定的自适应梯度。我们的 **FedDA** 框架具有通用性，并且可以纳入

表1：用于求解目标函数（1）的 ϵ-驻点的代表性联邦学习算法比较即， $\|\nabla f(x)\|^2 \leq \epsilon$ 或其等价变体。 $Gc(f,\ \epsilon)$ 表示针对 $k \in [K]$ 的梯度查询次数（关于 $f^{(k)}(x)$）； $Cc(f,\ \epsilon)$ 表示通信轮数；**状态** 表示算法在本地维护的状态（原始/对偶）；**局部自适应** 表示算法是否在本地执行自适应梯度下降；**有约束** 表示算法能否同时求解有约束和无约束问题。前三种算法不是自适应梯度方法，后四种方法支持某种形式的自适应梯度。

算法	$Gc(f,\ \epsilon)$	$Cc(f,\ \epsilon)$	状态	局部自适应	有约束
FedAvg [31]	$O(\epsilon^{-2})$	$O(\epsilon^{-1.5})$	原始-对偶		
FedCM [44]	$\tilde{O}(\epsilon^{-1.5})$	$\tilde{O}(\epsilon^{-1})$	原始-对偶		
STEM [22]	$\tilde{O}(\epsilon^{-1.5})$	$\tilde{O}(\epsilon^{-1})$	原始-对偶		
FedAdam [34]	$O(\epsilon^{-2})$	$O(\epsilon^{-1})$	原始		
Local-AMSGrad [4]	$O(\epsilon^{-2})$	$O(\epsilon^{-1.5})$	原始	√	
MIME-MVR [16]	$\tilde{O}(\epsilon^{-1.5})$	$O(\epsilon^{-1.5})$	原始	√	
FedDA-MVR（我们的方法）	$\tilde{O}(\epsilon^{-1.5})$	$\tilde{O}(\epsilon^{-1})$	Dual	√	√

将一大类自适应梯度方法应用于联邦学习场景。具体而言，**FedDA-MVR**是我们框架的一种具体实现，在联邦学习场景中实现了已知最优的梯度复杂度和通信复杂度。最后，我们将**贡献** 概括如下：

(i) 我们提出了**FedDA**，一种用于联邦自适应梯度方法的框架。该框架采用重启式对偶平均技术，并将一大类自适应梯度方法应用于联邦学习场景；

(ii) **FedDA-MVR**作为我们框架的一种具体实现，实现了 $\tilde{O}(\epsilon^{-1.5})$ 的梯度复杂度和 $\tilde{O}(\epsilon^{-1})$ 的通信复杂度。这达到了非自适应联邦算法的最优速率，并优于现有的自适应联邦算法。**FedDA-MVR**采用基于动量的方差缩减梯度估计，并使用梯度平方的指数移动平均作为自适应学习率；

(iii) 我们通过在 CIFAR10 和 FEMNIST 数据集上执行结直肠癌预测任务和分类任务，实证验证了框架 **FedDA** 的有效性。

符号说明。 $\nabla f(x)$ ($\nabla f(k)(x)$) 表示函数 $f(x)$ ($f(k)(x)$) 关于变量 x 的一阶导数。 ξ表示一个随机样本， $\nabla f(x;\ \xi)$($\nabla f^{(k)}(x;\ \xi)$) 是 $\nabla f(x)$ ($\nabla f(k)(x)$) 的随机估计。 $O(\cdot)$ 表示大 O 表示法， $\tilde{O}(\cdot)$ 隐去了对数项。 I_d表示一个 d 维单位矩阵。 $Diag(x)$ 表示对角线为向量 x 的矩阵。 $\|\cdot\|$ 分别表示向量的 ℓ_2 范数和矩阵的谱范数。 $\langle\cdot,\cdot\rangle$ 表示欧几里得内积。 [K] 表示 $\{1, 2, ..., K\}$ 的集合。对于随机变量 X， $\mathbb{E}[X]$ 表示其期望。

2 相关工作

联邦学习中的优化算法。 联邦学习一词最早由 [31], 提出，其任务是在服务器的协调下，从一组分布式客户端中进行学习

。在论文 [31], 中，作者提出了 FedAvg 算法：每个客户端使用本地数据执行多步梯度下降，然后将更新后的模型发送至服务器进行平均。FedAvg 算法的思想与本地随机梯度下降算法相似，后者已在更一般的分布式环境中研究了更长时间 [30]。文献中对本地随机梯度下降方法的收敛性进行了大量分析 [37, 19, 9, 21, 46, 43, 42, 12]。最近， [12] 证明了本地随机梯度下降在凸环境下的收敛速率达到了下界。在基础的本地随机梯度下降方法之上，人们研究了各种加速方法；下面列举其中几个代表性方法。 [17] 将非分布式有限和问题中的方差缩减技术应用于此：利用包含历史完整梯度信息的“控制变量”来校正本地梯度的偏差。随后， [16],提出了一个通用框架（MIME），将集中式优化器（包括自适应梯度方法）迁移到联邦学习环境中。在 MIME 中，优化器的状态在本地更新步骤期间保持固定，仅在服务器平均步骤中更新。 [8, 23], 将基于动量的方差缩减应用于联邦学习，以控制随机梯度的噪声。 [8], 维护了服务器动量状态和客户端动量状态，而在 [23], 中，作者维护了一个动量状态，并像原始状态一样定期对动量进行平均。

自适应梯度方法在联邦学习（FL）设置下也得到了研究。“自适应联邦优化” [34] 方法提出在服务器端使用自适应梯度，同时使用本地梯度更新自适应梯度方法的状态。在 [4], 中，作者首先展示了朴素本地 AMSGrad 方法的发散性；该方法会定期直接对原始状态进行平均。随后，作者提出了 Local-AMSGrad，这是一种由客户端在本地更新自适应学习率，并在同步步骤中进行平均的方法。在服务器平均步骤中，原始状态和本地自适应学习率都会进行平均，以替换旧状态。最后，另一类研究 [39, 38, 28, 2] 通过压缩方法考察联邦自适应学习率：这些方法在每一步都通信本地梯度，但利用压缩技术降低通信成本。

非分布式学习中的自适应梯度。自适应梯度方法广泛应用于非分布式机器学习环境。首个自适应梯度方法即Adagrad 在 [10], 中提出，该方法被证明在稀疏梯度环境下优于随机梯度下降。由于 Adagrad 在稠密梯度环境和非凸环境下表现不佳，因此研究者提出了若干变体，例如 SC-Adagra [32] 和 SAdagrad [6]。此外，Adam [24] 和 YOGI [48] 提出使用指数移动平均来取代 Adagrad 中使用的算术平均。Adam/YOGI 广泛应用于深度学习应用并取得了显著成功；然而，Adam 在某些环境下会发散，且梯度信息会迅速消失。因此，研究者提出了 AMSGrad [35] ，并引入一个额外的“长期记忆”变量来保留历史梯度信息，以解决 Adam 的收敛性问题。文献 [5, 49, 26, 13, 15] 也研究了 Adam 类方法的收敛性。研究者还提出了泛化性能良好的自适应梯度方法，例如 AdamW [27],、Padam [3], 、Adabound [29],、Adabelief [50] 和 AaGrad-Norm [41]。

3 预备知识

本节将在介绍我们的框架之前，先介绍一些预备知识。首先，我们考虑如下联邦学习表述：

$$\min_{x\in\mathcal{X}\subset\mathbb{R}^d}\left\{f(x)\coloneqq\frac{1}{K}\sum_{k=1}^{K}\left\{f^{(k)}(x)\coloneqq\mathbb{E}_{\xi^{(k)}\sim\mathcal{D}^{(k)}}[f^{(k)}(x;\xi^{(k)})]\right\}\right\}. \tag{1}$$

其中包含 K 个客户端。对于第 k_{th} 个客户端，我们优化损失目标函数 $f^{(k)}(x)$: $\mathcal{X}\to\mathbb{R}$ 该函数光滑且可能是非凸的， x 表示所关注的变量。 $\mathcal{X}\subset\mathbb{R}^d$ 是一个紧凸集。 $\xi^{(k)}\sim\mathcal{D}^{(k)}$ 是遵循未知数据分布 $\mathcal{D}^{(k)}$的随机样本。公式 (1) 同时包含同质情形即 $f^{(k)}(x)=f^{(j)}(x)$ 对任意 $k,\ j\in[K]$，以及异质情形即 $f^{(k)}(x)\neq f^{(j)}(x)$ 对某些 $k,\ j\in[K]$。

接下来，我们从镜像下降视角介绍一些自适应梯度方法的基础知识。一般而言，镜像下降与镜像映射 $\Phi(x)$ 相关。给定目标函数 $f(x)$ 和在第 t_{th} 步时的原始状态 $x_t\in\mathcal{X}$，我们首先将原始状态映射到镜像空间中，即 $y_t=\nabla\Phi(x_t)$，然后在镜像空间中执行梯度下降步骤： $y_{t+1}=y_t-\eta\nabla f(x)$，其中 η是学习率；最后，我们将 y_{t+1}映射回原始空间，即 $x_{t+1}=\arg\min D_\Phi(x,\ y_{t+1})$，其中 $D_\Phi(x,\ y)_{x\in\mathcal{X}}$

表示与 Φ 相关的 Bregman散度，即 $D_\Phi(x,\ y)=f(x)-f(y)-\langle\nabla f(y)\ x-y\rangle$。总之，镜像下降步骤可写成如下 Bregman近端梯度步骤：

$$x_{t+1}=\arg\min_{x\in\mathcal{X}}\ \eta\langle\nabla f(x_t),x\rangle+D_\Phi(x,x_t)$$

对于自适应梯度方法，我们使用以下镜像映射： $\Phi(x)=\frac{1}{2}x^THx$，其中 H是自适应矩阵，且为正定。许多自适应梯度方法都可以写成以下近端梯度下降形式：

$$x_{t+1}=\arg\min_{x\in\mathcal{X}}\ \eta\langle\nu_t,x\rangle+\frac{1}{2}(x-x_t)^TH_t(x-x_t), \tag{2}$$

我们将梯度 $\nabla f(x)$ 替换为广义梯度估计 ν_t；此外，基于自适应矩阵在每一步都会更新这一事实，我们将 H替换为 H_t 。接下来，我们展示一些可用上述形式表述的自适应梯度方法示例。对于 Adagrad [10] 方法，我们设

$$\nu_t=\nabla f(x_t,\xi_t),\ H_t=Diag(\sqrt{\mu_t}),\ \mu_t=\frac{1}{t}\sum_{i=1}^{t}\nu_i^2 \tag{3}$$

对于 Adam [24], , 我们有：

$$\begin{aligned}&\hat{\nu}_t=(1-\beta_1)\nabla f(x_t,\xi_t)+\beta_1\hat{\nu}_{t-1},\\&\hat{\mu}_t=(1-\beta_2)\nabla f(x_t,\xi_t)^2+\beta_2\hat{\mu}_{t-1}\\&\nu_t=\hat{\nu}_t/(1-\gamma_1^t),\ \mu_t=\hat{\mu}_t/(1-\gamma_2^t),\\&H_t=Diag(\sqrt{\mu_t}+\epsilon)\end{aligned} \tag{4}$$

其中 $\beta_1,\ \beta_2,\ \gamma_1,\ \gamma_2$ 是一些常数。关于其他自适应梯度方法，请参阅 [15] .

算法1 FedDA-服务器

1：输入：全局轮次数量 E**，调优参数** $\{\beta_\tau\}_{i=1}^{E}$**；2：初始化：选择** $x_0 \in \mathcal{X}$**并计算** $\nu_0 = \frac{1}{K}\sum_{j=1}^{K} \nabla f^{(j)}(x_0, \mathcal{B}_0^{(k)})$**，其中** $\{\mathcal{B}_0^{(k)}\}_{k=1}^{K}$**是从每个** K **个客户端中选取的小批量随机点；3：对于** $\tau = 0$ **到** $E-1$**，执行 4：服务器从** r **个客户端中均匀随机地无放回选取一个集合** $\mathcal{S}_\tau$**；5：并行处理客户端** $k \in \mathcal{S}_\tau$**：6：(**$z_{\tau+1}^{(k)}{}_{,I}$**,** $\nu_{\tau+1}^{(k)}{}_{,I}$**) =FedDA-客户端(**$x_\tau$**,** ν_τ**,** H_τ**) 7：结束循环；8：计算** $z_{\tau+1} = \frac{1}{r}\sum_{k\in\mathcal{S}_\tau} z_{\tau+1}^{(k)}{}_{,I}$**；9：计算** $x_{\tau+1} = \operatorname{arg\,min}_{x\in\mathcal{X}} \{-\langle x, z_{\tau+1}\rangle + \frac{1}{2\lambda}(x - x_\tau)^T H_\tau (x - x_\tau)\}$**；10：计算** $\nu_{\tau+1} = \frac{1}{r}\sum_{k\in\mathcal{S}_\tau} \nu_{\tau+1}^{(k)}{}_{,I}$**；11：计算** $H_{\tau+1} = \mathcal{V}(H_\tau, z_{\tau+1})$**；12：结束循环**

算法 2 FedDA-客户端（x_τ， ν_τ， H_τ）

1：输入：本地步数 I、调优参数 $\{\eta_{\tau+1,i}\}_{i=0}^{I-1}$、 $\{\alpha_{\tau+1,i}\}_{i=1}^{I}$；2：初始化： $x_{\tau+1}^{(k)}{}_{,0} = x_\tau$； $\nu_{\tau+1}^{(k)}{}_{,0} = \nu_\tau$； $z_{\tau+1}^{(k)}{}_{,0} = 0$；3：对于 $i = 0$ 到 $I-1$，执行 4：计算 $z_{\tau+1}^{(k)}{}_{,i+1} = z_{\tau+1}^{(k)}{}_{,i} - \eta_{\tau+1,i}\nu_{\tau+1}^{(k)}{}_{,i}$；5：计算 $x_{\tau+1}^{(k)}{}_{,i+1} = \operatorname{arg\,min}_{x\in\mathcal{X}} \{-\langle x, z_{\tau+1,i+1}^{(k)}\rangle + \frac{1}{2\lambda}(x - x_{\tau+1}^{(k)}{}_{,0})^T H_\tau (x - x_{\tau+1}^{(k)}{}_{,0})\}$；6：计算 $\nu_{\tau+1}^{(k)}{}_{,i+1} = \mathcal{U}(\nu_{\tau+1}^{(k)}{}_{,i}, x_{\tau+1}^{(k)}{}_{,i+1}, x_{\tau+1}^{(k)}{}_{,i}; \alpha_{\tau+1,i+1}, \mathcal{B}_{\tau+1}^{(k)}{}_{,i+1})$，其中 $\mathcal{B}_{\tau+1}^{(k)}{}_{,i+1}$ 是来自客户端 k 的随机样本小批次；7：结束循环 8：输出：将 $z_{\tau+1}^{(k)}{}_{,I}$、 $\nu_{\tau+1}^{(k)}{}_{,I}$ 发送至服务器。

4 通过对偶平均实现的局部自适应梯度

本节将介绍**FedDA**，这是一个联邦自适应梯度方法框架。**FedDA** 的流程总结于算法1中。

在算法1中，我们执行 E 个全局步，并在每个全局步选择一部分客户端进行训练。每一步中，所有入选客户端都会运行算法 2。在算法 2 中，客户端接收当前模型权重 x_τ、梯度估计 ν_τ 和自适应梯度矩阵 H_τ。随后，客户端执行 I 个本地训练步：即算法 2 中第 3 行至第 7 行。每一步中，我们首先在变量 $z_{\tau\,i}^{(k)}{}_{,}$ 中累积对偶状态（第 4 行），然后计算局部原始状态 $x_{\tau\,i}^{(k)}{}_{,}$（第 5 行），这是一种类似于式（2）的近端梯度步。此步骤的作用是将聚合对偶状态 $z_{\tau\,i}^{(k)}{}_{,}$ 映射回原始空间，并使用原始状态查询梯度，以更新梯度估计 $\nu_{\tau\,i}^{(k)}{}_{,}$（第 6 行）。注意，在本地步期间，我们使用固定的自适应矩阵 H_τ，这使客户端共享同一个对偶空间。在算法 2 的第 6 行，我们更新梯度估计 $\nu_{\tau\,i}^{(k)}{}_{,}$。更新规则 $\mathcal{U}(\cdot)$ 具有通用性，例如，基于动量的方差

red更新（5）和动量更新（6）如下（$\alpha_{\tau,i}$ 为某个常数）：

$$\nu_{\tau+1,i+1}^{(k)} = \nabla f^{(k)}(x_{\tau+1,i+1}^{(k)}, \mathcal{B}_{\tau+1,i+1}^{(k)}) + (1-\alpha_{\tau+1,i+1})(\nu_{\tau+1,i}^{(k)} - \nabla f^{(k)}(x_{\tau+1,i}^{(k)}, \mathcal{B}_{\tau+1,i+1}^{(k)})) \tag{5}$$

and

$$\nu_{\tau+1,i+1}^{(k)} = \alpha_{\tau+1,i+1}\nabla f^{(k)}(x_{\tau+1,i+1}^{(k)}, \mathcal{B}_{\tau+1,i+1}^{(k)}) + (1-\alpha_{\tau+1,i+1})\nu_{\tau+1,i}^{(k)} \tag{6}$$

客户端运行算法 2 后，会将聚合的局部对偶状态 $z_{\tau+1,\ I}^{(k)}$ 以及局部梯度估计 $\nu_{\tau+1,\ I}^{(k)}$ 返回给服务器。服务器首先对局部对偶状态取平均（算法1的第 8 行），得到 $z_{\tau+1}$。由于所有客户端都处于公共对偶空间中，因此可以对局部对偶状态取平均。随后，服务器按照算法1的第 9 行计算新的原始状态 $x_{\tau+1}$。接下来，通过对局部状态取平均来更新梯度估计 ν_τ（算法1的第 10 行）。最后，我们更新自适应矩阵 H_τ（算法1的第 11 行）。更新规则 $\mathcal{V}$ 具有一般性，例如，

$$\begin{aligned} \mu_{\tau+1} &= \beta_{\tau+1} z_{\tau+1}^2 / \eta_{\tau+1,I-1}^2 + (1-\beta_{\tau+1})\mu_{\tau+1}, \\ H_{\tau+1} &= Diag(\sqrt{\mu_{\tau+1}} + \epsilon) \end{aligned} \tag{7}$$

and

$$\begin{aligned} \mu_{\tau+1} &= \beta_{\tau+1} ||z_{\tau+1}|| / \eta_{\tau+1,I-1} + (1-\beta_{\tau+1})\mu_{\tau+1}, \\ H_{\tau+1} &= (\mu_{\tau+1} + \epsilon) I_d \end{aligned} \tag{8}$$

其中我们设定 $\mu_0 = 0$，ϵ 为某个常数。总之，算法1在每个全局轮次都会聚合并平均对偶状态。自适应矩阵 H_τ 在本地更新期间保持固定，并在每个全局轮次于服务器端刷新。由于该算法在每个全局轮次都会使用新的镜像映射（自适应梯度矩阵），我们将这一框架称为重启式对偶平均。

注1。 不同于我们的对偶平均策略，一些现有的自适应联邦学习算法 *[33]*会对局部原始状态进行平均。在无约束情形下，原始空间与对偶空间彼此线性相关；但在有约束情形下，这种线性关系并不存在，因此在原始空间和对偶空间中进行平均并不等价。正如我们在后续理论分析中所展示的，对偶平均能够保证有约束情形下的收敛性。

注2。 请注意，在更新自适应矩阵时（算法*1*第*11*行），我们使用了平均对偶状态 $z_{\tau+1}$。另一种选择是使用最新的梯度 *[33]*。相比之下，对偶状态汇总了整个轮次的信息，能够更平滑地估计问题的局部曲率。另一种可能的选择是采用 *Local-AMSGrad* 方法 *[4],* 中的做法，在本地更新自适应矩阵的状态 μ_τ（见（7）），然后在服务器同步步骤中进行平均。这种方法的局限性在于， μ_τ 相对于 梯度不是线性的，因此，对 μ_τ 进行平均无法带来相对于 客户端数量的线性加速；相比之下，对偶状态满足线性关系。

注 3。 通过选择不同的更新规则 $\mathcal{U}$ 和 $\mathcal{V}$，我们可以构造出 *FedDA* 的许多变体。一个具有代表性的变体是 ***FedDA-MVR***，其中我们通过基于动量的方差缩减（(5)）来更新 $\nu_{\tau\ i}^{(k)}$，，并通过对梯度平方进行指数平均（(7)）来更新自适应矩阵 H_τ。在接下来的讨论中，我们将重点研究这一变体，并进行理论分析和实证分析。

5 理论分析

本节将对我们的**FedDA** 框架进行理论分析；更具体地说，我们重点分析 **FedDA-MVR**。FedDA-MVR 使用 (7) 更新自适应矩阵H_τ，并使用 (5) 更新梯度估计 $\nu_{\tau\,i}^{(k)}$，。首先，我们陈述分析所需的假设：

5.1 一些温和假设

假设 1（有界客户端异质性）。 不同工作节点之间的梯度差异有界：

$$\|\nabla f^{(k)}(x) - \nabla f^{(\ell)}(x)\|^2 \le \zeta^2,\ \forall k, \ell \in [K].$$

我们根据梯度差异度来衡量客户端的异质性。上述假设或其类似形式也被用于其他联邦学习算法的分析中，例如 [22, 8]。

A假设 2。 函数 f（x） 在 $\mathcal{X}$ 中有下界，即 $f^* =$下确界$_{x\in\mathcal{X}}\, f$（x）。

假设 3（无偏且方差有界的随机梯度）。 随机梯度是无偏的，且方差有界，即

$$\mathbb{E}[\nabla f^{(k)}(x;\xi^{(k)})] = \nabla f^{(k)}(x)$$

并且存在一个常数 σ，使得

$$\mathbb{E}\|\nabla f^{(k)}(x;\xi^{(k)}) - \nabla f^{(k)}(x)\|^2 \le \sigma^2,$$
$$\forall\ \xi^{(k)} \sim \mathcal{D}^{(k)},\ \forall\ k \in [K]$$

假设 2 保证了联邦学习问题 (1) 的可行性，而假设 3 在随机优化分析中得到了广泛应用。

假设 4. 自适应矩阵 H_τ 是对称正定的，即 存在一个常数$\rho > 0$，使得

$$H_\tau \succeq \rho I_d \succ 0,\ \forall t \ge 1,$$

在我们的分析中，我们假设自适应矩阵是正定的，而许多自适应梯度方法都可以轻松满足这一要求。首先，大多数自适应梯度方法的自适应学习率始终非负，例如式 (3) 和式 (4)。为了使其为正，我们可以添加偏置项 ϵ，如 Adam 更新规则 (4) 中所示。

假设 5 (样本梯度利普希茨光滑性**).** 随机函数 $f^{(k)}(x,\xi^{(k)})$满足 $\xi^{(k)} \sim \mathcal{D}^{(k)}$ （对所有 $k \in [K]$），并具有均方光滑性，即我们有

$$\mathbb{E}\|\nabla f^{(k)}(x;\xi^{(k)}) - \nabla f^{(k)}(y;\xi^{(k)})\|^2 \le L^2\|x-y\|^2$$
$$for\ all\ x, y \in \mathbb{R}^d$$

上述光滑性假设比标准光滑条件略强，但这一假设广泛应用于方差缩减方法的分析中，例如 SPIDER [11] 和 STORM [7]。

假设 **6.** 每一步中所有客户端都参与训练，即 在算法*1*中选择 $r = K$ 。

为简化理论结果的阐述，我们采用全参与假设。所呈现的所有结果都可以轻松推广到部分参与情形。

5.2 Fed-MVR收敛性质

在本小节中，我们给出**FedDA-MVR** 变体的收敛性质。为便于讨论，我们重新定义下标 $t = \tau I + i$，即 将 t 步称为全局轮次中的 i 本地步。类似地，我们将运行总步数记为 $T = EI$。我们通过以下测度分析算法:

$$\mathcal{G}_t = \frac{\rho^2}{\lambda^2 \eta_t^2} ||\tilde{x}_t - \tilde{x}_{t+1}||^2 + ||\bar{\nu}_t - \nabla f(\tilde{x}_t)||^2 \tag{9}$$

其中， $\bar{\nu}_t$ 表示 t 步的平均梯度估计， $\tilde{x}_t$ 表示 t 步的虚拟全局原始状态（形式化定义见附录B.2节）。在附录的注7中，我们讨论测度 $\mathcal{G}_t$ 的直观含义。具体而言，在无约束情形即 当 $\mathcal{X} = R^d$ 时，该测度是梯度平方范数的上界。因此，测度 $\mathcal{G}_t$ 收敛意味着收敛到一阶驻点。现在，我们给出收敛定理的主要结果。

定理 1。 在算法*1*中，我们选择参数为 $\kappa = \frac{\rho K^{2/3}}{\lambda L}$, $c = \frac{96\lambda^2 L^2}{K\rho^2} + \frac{\rho}{72\kappa^3 \lambda L I^2}$, $w_t = \max \left\{ 48^3 I^6 K^2 - t - I, 14^3 K^{0.5} \right\}$, $\lambda > 0$, 并选择 $\eta_t = \frac{\kappa}{(\omega_t + t + I)^{1/3}}$, 则有:

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\mathcal{G}_t] \leq & \left[\frac{96LI^2}{T} + \frac{2L}{K^{2/3} T^{2/3}} \right] (f(x_0) - f^*) + \left[\frac{72I^4}{bT} + \frac{3I^2}{2bK^{2/3} T^{2/3}} \right] \sigma^2 \\ & + 192^2 \times \left(\frac{48I^2}{T} + \frac{1}{K^{2/3} T^{2/3}} \right) \left(\frac{\sigma^2}{4b_1} + \frac{2\zeta^2}{21} \right) \log(T+1). \end{aligned}$$

注意，通过选择适当的本地更新次数 I ，并在第一次迭代中使用一批样本来降低噪声，我们的结果达到了随机联邦梯度方法已知的最佳收敛速率 [22], 即 我们的算法具有 $\tilde{O}(\epsilon^{-1.5})$ 的梯度复杂度和 $\tilde{O}(\epsilon^{-1})$ 的通信复杂度；此外，相对于客户端数量 K，我们实现了线性加速 。更正式地说，我们有以下推论:

推论1. 假设在算法*1*中，我们设置 $I = O((T/K^2)^{1/6})$，并在初始化时使用大小为 $O(I^2)$ 的样本小批量，则有:

$$\frac{1}{T} \sum_{t=1}^{T} \left(\mathbb{E}[\mathcal{G}_t] \right) = \tilde{O}(\frac{1}{K^{2/3} T^{2/3}})$$

并且要达到一个 ϵ驻点，我们需要进行 $\tilde{O}$（$\epsilon^{-1.5}/K$）步，并需要 $\tilde{O}$（ϵ^{-1}）轮通信。

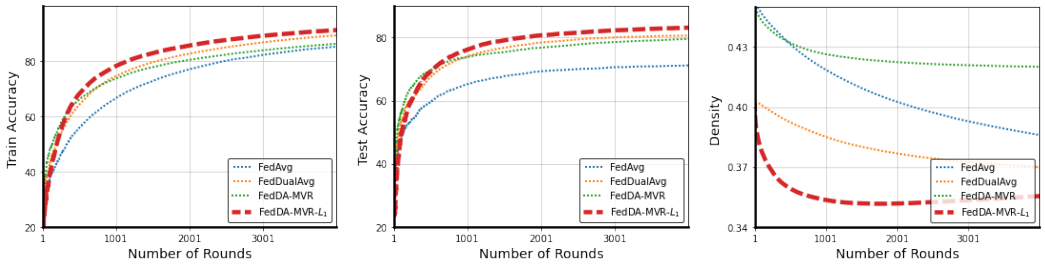

图1：PATHMNIST [45] 数据集的结果。图中分别展示了训练准确率、测试准确率、密度与轮数（E 算法1）。 L_1 的后缀表示我们考虑了 L_1 约束。 I 取值为5。

6 数值实验

在本节中，我们进行数值实验，以验证所提出的自适应联邦学习框架，即 **FedDA**的有效性。具体而言，我们在此考察**FedDA-MVR** 变体，并将其他变体的实验留待附录的 A节介绍。我们进行了两组实验。第一组实验考察一项生物医学预测任务：预测结直肠癌患者的生存情况。在该任务中，我们施加了 L_1 稀疏性约束。 L_1 约束提升了模型的可解释性，这对生物医学应用至关重要。第二组实验考察一项联邦多类别图像分类任务。具体而言，我们使用两个数据集：CIFAR10 [25] 和 FEMNIST [1]。所有实验均在一台配备英特尔至强金牌6248处理器和 4 块英伟达特斯拉V100 GPU 的机器上运行。代码使用 PyTorch 编写。我们通过 Pytorch.distributed 软件包模拟联邦学习环境。

6.1 带稀疏约束的结直肠癌生存预测

在本小节中，我们在 PATHMNST 数据集上考虑一个结直肠癌预测任务， [45, 20],该数据集包含9个不同类别。该数据集有89996张训练图像，我们将训练集随机均分给10个客户端。我们使用原始测试集计算指标。在该任务中，我们施加L_1 稀疏性约束，以提高模型的可解释性。

在此任务中，我们将以下基线方法进行比较：FedAvg [31] 和 FedDualAvg [47]。FedDualAvg 是一种近期提出的、用于处理复合优化问题的联邦算法。在 FedDualAvg 中，客户端在本地维护对偶状态，但不使用自适应梯度。对于我们的 FedDA-MVR，我们分别在有和没有 L_1 约束的情况下进行训练。在此任务中，我们使用一个 4 层卷积神经网络，每层包含 32 个滤波器。我们设置 10 个客户端，运行 20000 步（T），每隔 5 步对状态求平均（I），并将小批量大小设为 16。此外，我们使用 0.01 的阈值计算密度。对于其他超参数，我们通过网格搜索为每种方法选择最佳设置。具体而言，对于随机梯度下降方法，我们使用学习率 0.01；对于 FedDualAvg 算法，我们使用本地学习率 0.1、全局学习率 0.1、 L_1 约束 0.01；对于我们的 FedDA-MVR，我们使用学习率 0.01、 w 为 100000、 c 为 5000000、 β 为 0.999、 τ为 0.01；对于 L_1 正则化版本 FedDA-MVR-L_1，我们还添加了 L_1 约束 0.01。对于 FedDA 的其他变体：对于 FedDA-2-1，我们使用学习率 0.001、 α 为 0.9、 β 为 0.999、 τ 为 0.01；对于 FedDA-1-2，我们使用学习率 1、 w 为 10000、 c 为 200、 β 为 0.999、 τ 为 0.001、 L_1 约束 0.01；

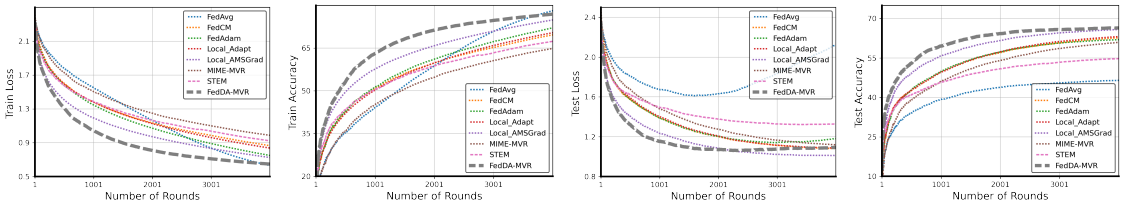

图2：CIFAR10数据集的结果。从左到右，我们分别展示了训练损失、训练准确率、测试损失和测试准确率与轮数（算法1中的E）的关系。 I 取值为5。

对于 FedDA-2-2，我们使用学习率 0.01、 α 0.9、 β 为 0.999、 τ 为 0.01、 L_1 约束 0.01。

结果汇总于图1中，图中的曲线取5次独立运行的平均值后进行了平滑处理。在图1中，FedDualAvg和FedDA-MVR-L_1 考虑了 L_1 约束，而FedAvg和FedDA-MVR未考虑约束。我们展示了训练/测试准确率，以及参数中非零（低于某一阈值）元素的数量（即图1最右侧的图）。如图所示，FedDA-MVR-L_1 在所有指标上均优于无约束的FedDA-MVR。这表明，在联邦学习中考虑有约束问题十分重要。此外，FedDA-MVR-L_1 在所有指标上也优于FedAvg和FedDualAvg。这表明，我们的算法能够在约束情形下有效利用自适应梯度信息。有关该实验的更多细节（例如超参数选择），请参阅附录的A节。

6.2 CIFAR10 和 FEMNIST 图像分类任务

在本小节中，我们考虑同质和异质情形下的无约束图像分类任务。具体而言，我们考虑两个数据集：CIFAR10 [25] 和 FEMNIST [1]。CIFAR10 是一种广泛使用的图像分类基准数据集，包含 50000 张训练图像，我们基于该数据集构造同质和异质情形。对于同质情形，我们将这些图像均匀随机地分配给 10 个客户端。对于异质情形，我们按如下方式在训练集内引入异质性：假设有 10 个客户端，对于 i_{th} 客户端，我们分配 ρ% 的 i_{th} 类别样本，以及其他类别的 $(1-\rho)/9$% 样本，其中 $0 < \rho \leq 1$。注意，当 ρ 接近 1 时， i_{th} 客户端将主要由 i_{th} 类别的图像构成，因此客户端之间的数据分布将大不相同。在实验中，我们选择 $\rho = 0.8$。这意味着 i_{th} 客户端拥有 4000 张 i_{th} 类别的图像，以及 111 张其他类别的图像。这造成了高度异质性。注意，我们使用 CIFAR10 的原始测试集。FEMNIST 是一个联邦手写数字数据集，包含 3550 位用户的手写数字（在实验中，我们随机抽取 500 位用户）。由于人们的书写风格各异，FEMNIST 的数据分布具有异质性。注意，我们在每个全局轮次随机抽取 50 位用户。我们运行 20000 步（T），每隔 5 步对状态求平均（I），并使用大小为 16 的小批量。我们使用一个 4 层卷积神经网络，每层有 64 个滤波器。对于其他超参数，我们进行网格搜索，并为每种方法选择最佳设置。

在本任务中，我们将方法与以下基线方法进行比较：非自适应方法：FedAvg [31], FedCM [44], STEM [22] 以及自适应方法：FedAdam [34], Local-Adapt [40], Local-

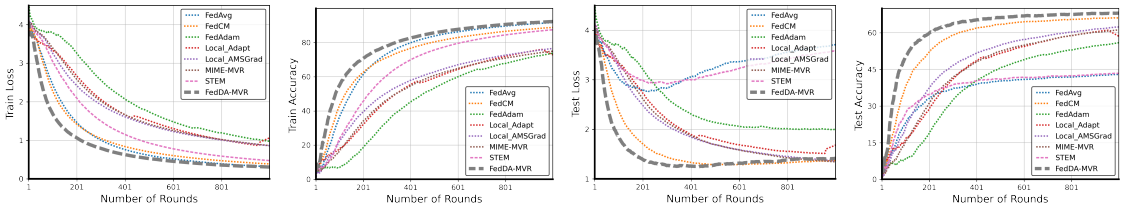

图3：FEMNIST数据集的结果。从左到右，依次展示训练损失、训练准确率、测试损失和测试准确率与轮数（算法1中的E）的关系。 I 取5。

AMSGrad [4], MIME-MVR [33]。对于所有方法，我们都会调节其超参数，以找到最佳设置。结果汇总于图2和图11（CIFAR10）以及图3（FEMNIST）中，图表数据取5次运行的平均值后进行了平滑处理。请注意，在图11中，**FedDA-i-j** 表示我们框架的不同变体，其定义请参阅附录A。特别地，**FedDA-1-1**表示**FedDA-MVR**。如图所示，我们的FedDA-MVR优于所有基线方法。此外，FedAvg 算法具有较强的训练性能；然而，它往往会严重过拟合训练数据，并且受异质性的影响最为严重。随后我们观察到，自适应方法通常能够获得更好的训练和测试性能。最后，我们的方法相较于三个自适应基线方法所展现出的优越性能表明，我们的方法能够更好地利用自适应信息；例如，MIME-MVR同样采用了基于动量的方差缩减技术，但它在本地更新期间固定所有优化器状态；相比之下，我们只固定自适应矩阵，同时在每一步更新动量 $\nu_t^{(k)}, k \in [K]$ 。关于更多细节，包括超参数的选择，请参阅附录的A节。

7 结论

本文提出了FedDA框架，将自适应梯度引入联邦学习环境。具体而言，我们采用自适应梯度的镜像下降视角，并提出在训练过程中维护和平均对偶状态；同时，我们在本地训练期间固定自适应矩阵，使所有客户端共享同一个对偶空间。我们还分析了该框架的收敛性质：对于FedDA-MVR变体，我们证明它以 $\tilde{O}(\epsilon^{-1.5})$次梯度查询和 $\tilde{O}(\epsilon^{-1})$轮通信达到一个 ϵ-最优驻点。这些结果达到了非凸情形下随机联邦算法已知的最佳梯度复杂度和通信复杂度。最后，我们在有约束和无约束任务上验证了该算法。数值结果表明，与多种基线方法相比，我们的算法性能更优。

参考文献

[1] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith 和 A. Talwalkar。LEAF：联邦设置的基准。*arXiv* 预印本 *arXiv:1812.01097*，2018年。

[2] C. Chen, L. Shen, H. Huang, W. Liu 和 Z.-Q. Luo。Efficient-Adam：具有复杂度分析的高通信效率分布式 Adam。2020年。[3] J. Chen, D. Zhou, Y. Tang, Z. Yang 和 Q. Gu。缩小深度神经网络训练中自适应梯度方法的泛化差距。*arXiv* 预印本 *arXiv:1806.06763*，2018年。[4] X. Chen, X. Li 和 P. Li。迈向通信高效的自适应梯度方法。载于*2020*年 *ACM-IMS* 数据科学基础会议论文集，第 119–128 页，2020年。[5] X. Chen, S. Liu, R. Sun 和 M. Hong。关于一类用于非凸优化的 Adam 型算法的收敛性。载于第七届国际学习表征会议（*ICLR 2019*），2019年。[6] Z. Chen, Y. Xu, E. Chen 和 T. Yang。SAdagrad：强自适应随机梯度方法。载于国际机器学习会议，第 913–921 页。PMLR，2018年。[7] A. Cutkosky 和 F. Orabona。非凸随机梯度下降中的基于动量的方差缩减。载于神经信息处理系统进展，第 15236–15245 页，2019年。[8] R. Das, A. Acharya, A. Hashemi, S. Sanghavi, I. S. Dhillon 和 U. Topcu。通过全局与局部动量加速非凸联邦学习。*arXiv* 预印本 *arXiv:2012.04061*，2020年。[9] A. Dieuleveut 和 K. K. Patel。大步长本地随机梯度下降的通信权衡。神经信息处理系统进展，第 32 卷，2019年。[10] J. Duchi, E. Hazan 和 Y. Singer。用于在线学习和随机优化的自适应次梯度方法。机器学习研究杂志，12(7)，2011年。[11] C. Fang, C. J. Li, Z. Lin 和 T. Zhang。SPIDER：通过随机路径积分微分估计器实现近乎最优的非凸优化。载于神经信息处理系统进展，第 689–699 页，2018年。[12] M. R. Glasgow, H. Yuan 和 T. Ma。联邦平均（本地随机梯度下降）的紧确界与连续视角。载于国际人工智能与统计会议，第 9050–9090 页。PMLR，2022年。[13] Z. Guo, Y. Xu, W. Yin, R. Jin 和 T. Yang。关于用于非凸优化的随机移动平均估计器。*arXiv* 预印本 *arXiv:2104.14840*，2021年。[14] F. Haddadpour 和 M. Mahdavi。关于联邦学习中局部下降方法的收敛性。*arXiv* 预印本 *arXiv:1910.14425*，2019年。[15] F. Huang, J. Li 和 H. Huang。Super-Adam：更快且通用的自适应梯度框架。*arXiv* 预印本 *arXiv:2106.08208*，2021年。[16] S. P. Karimireddy, M. Jaggi, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich 和 A. T. Suresh。Mime：在联邦学习中模仿集中式随机算法。*arXiv* 预印本 *arXiv:2008.03606*，2020年。

[17] S. P. Karimireddy、S. Kale、M. Mohri、S. Reddi、S. Stich 和 A. T. Suresh。Scaffold：用于联邦学习的随机控制平均。在国际机器学习会议，第 5132–5143 页。PMLR，2020年。[18] S. P. Karimireddy、S. Kale、M. Mohri、S. J. Reddi、S. U. Stich 和 A. T. Suresh。Scaf- fold：用于设备端联邦学习的随机控制平均。*arXiv* 预印本 *arXiv:1910.06378*，2019年。[19] S. P. Karimireddy、Q. Rebjock、S. Stich 和 M. Jaggi。误差反馈修正 signsgd 及其他梯度压缩方案。在 国际机器学习会议，第 3252–3261 页。PMLR，2019年。[20] J. N. Kather、J. Krisam、P. Charoentong、T. Luedde、E. Herpel、C.-A. Weis、T. Gaiser、A. Marx、N. A. Valous、D. Ferber 等。使用深度学习从结直肠癌组织学切片预测生存：一项回顾性多中心研究。*PLoS*医学，16(1):e1002730，2019年。[21] A. Khaled、K. Mishchenko 和 P. Richt´arik。关于相同数据和异质数据上的本地 SGD 的更紧理论。在 国际人工智能与统计会议，第 4519–4529 页。PMLR，2020年。[22] P. Khanduri、P. Sharma、H. Yang、M. Hong、J. Liu、K. Rajawat 和 P. Varshney。STEM：一种随机双侧动量算法，实现联邦学习中近乎最优的样本和通信复杂度。神经信息处理系统进展，34，2021年。[23] P. Khanduri、S. Zeng、M. Hong、H.-T. Wai、Z. Wang 和 Z. Yang。通过双重动量实现随机双层优化的近乎最优算法。*arXiv* 预印本 *arXiv:2102.07367*，2021年。[24] D. P. Kingma 和 J. Ba。Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014年。[25] A. Krizhevsky、G. Hinton 等。学习来自微型图像的多层特征。2009年。[26] L. Liu、H. Jiang、P. He、W. Chen、X. Liu、J. Gao 和 J. Han。关于自适应学习率的方差及其扩展。*arXiv* 预印本 *arXiv:1908.03265*，2019年。[27] I. Loshchilov 和 F. Hutter。解耦权重衰减正则化。在 国际学习表征会议，2018年。[28] Y. Lu、C. Li、M. Zhang、C. De Sa 和 Y. He。通过 0/1 Adam 最大化大规模训练的通信效率。*arXiv* 预印本 *arXiv:2202.06009*，2022年。[29] L. Luo、Y. Xiong、Y. Liu 和 X. Sun。具有学习率动态界限的自适应梯度方法。*arXiv* 预印本 *arXiv:1902.09843*，2019年。[30] O. L. Mangasarian 和 M. V. Solodov。通过确定性非单调扰动最小化实现反向传播收敛。神经信息处理系统进展，6，1993年。

[31] 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。在 人工智能与统计学，第 1273–1282 页。PMLR，2017年。[32] M. C. Mukkamala 和 M. Hein。具有对数遗憾界的 RMSProp 和 AdaGrad 变体。在 国际机器学习会议，第 2545–2553 页。PMLR，2017年。[33]S. Praneeth Karimireddy、M. Jaggi、S. Kale、M. Mohri、S. J. Reddi、S. U. Stich 和 A. Theertha Suresh。Mime：在联邦学习中模仿集中式随机算法。*arXiv* 电子论文，第 arXiv–2008 页，2020年。[34] S. Reddi、Z. Charles、M. Zaheer、Z. Garrett、K. Rush、J. Koneˇcn`y、S. Kumar 和 H. B. McMahan。自适应联邦优化。*arXiv* 预印本 *arXiv:2003.00295*，2020年。[35] S. J. Reddi、S. Kale 和 S. Kumar。Adam及其扩展的收敛性。在国际学习表征会议，2018年。[36] M. Staib、S. Reddi、S. Kale、S. Kumar 和 S. Sra。使用自适应梯度方法逃离鞍点。在 国际机器学习会议，第 5956–5965 页。PMLR，2019年。[37] S. U. Stich。本地 SGD 快速收敛且通信量少。*arXiv* 预印本 *arXiv:1805.09767*，2018年。[38] H. Tang、S. Gan、A. A. Awan、S. Rajbhandari、C. Li、X. Lian、J. Liu、C. Zhang 和 Y. He。1-bit Adam：以 Adam 的收敛速度实现高效通信的大规模训练。在国际机器学习会议，第 10118–10129 页。PMLR，2021年。[39] H. Tang、S. Gan、S. Rajbhandari、X. Lian、J. Liu、Y. He 和 C. Zhang。Apmsqueeze：一种通信高效的 Adam 预条件动量 SGD 算法。*arXiv* 预印本 *arXiv:2008.11343*，2020年。[40] J. Wang、Z. Xu、Z. Garrett、Z. Charles、L. Liu 和 G. Joshi。联邦学习中的局部自适应性：收敛性与一致性。*arXiv* 预印本 *arXiv:2106.02305*，2021年。[41] R. Ward、X. Wu 和 L. Bottou。AdaGrad 步长：非凸景观上的尖锐收敛。在 国际机器学习会议，第 6677–6686 页。PMLR，2019年。[42]B. Woodworth。分布式优化的极小极大复杂度。*arXiv* 预印本 *arXiv:2109.00534*，2021年。[43] B. Woodworth、K. K. Patel、S. Stich、Z. Dai、B. Bullins、B. Mcmahan、O. Shamir 和 N. Srebro。本地 SGD 是否优于小批量 SGD？在国际机器学习会议，第 10334–10343 页。PMLR，2020年。[44] J. Xu、S. Wang、L. Wang 和 A. C.-C. Yao。FedCM：具有客户端级动量的联邦学习。*arXiv* 预印本 *arXiv:2106.10874*，2021年。[45] J. Yang、R. Shi、D. Wei、Z. Liu、L. Zhao、B. Ke、H. Pfister 和 B. Ni。MedMNIST v2：用于二维和三维生物医学图像分类的大规模轻量级基准。*arXiv* 预印本 *arXiv:2110.14795*，2021年。

[46] H. Yu、S. Yang 和 S. Zhu。具有更快收敛速度和更少通信量的并行重启随机梯度下降：揭示模型平均为何适用于深度学习。载于*AAAI*人工智能大会论文集，第33卷，第5693–5700页，2019年。[47] H. Yuan、M. Zaheer 和 S. Reddi。联邦复合优化。载于国际机器学习会议，第12253–12266页。PMLR，2021年。[48] M. Zaheer、S. Reddi、D. Sachan、S. Kale 和 S. Kumar。用于非凸优化的自适应方法。载于神经信息处理系统进展，第9793–9803页，2018年。[49] D. Zhou、J. Chen、Y. Cao、Y. Tang、Z. Yang 和 Q. Gu。关于非凸优化中自适应梯度方法的收敛性。*arXiv*预印本 *arXiv:1808.05671*，2018年。[50] J. Zhuang、T. Tang、Y. Ding、S. C. Tatikonda、N. Dvornek、X. Papademetris 和 J. Duncan。Adabelief优化器：根据对观测梯度的置信度调整步长。神经信息处理系统进展，第33卷，2020年。

更多实验细节与结果

本节中，我们补充了更多实验。在A.1节中，我们考察**FedDA**的更多变体，除了**FedDA-MVR**之外。具体而言，我们考察**FedDA**的四种变体。我们针对式（7）和式（8）中的自适应矩阵更新引入两种情况，分别称为情况1和情况2；类似地，我们将式（5）和式（6）分别称为梯度估计的情况1和情况2。因此，我们有四种不同的变体，记为**FedDA-**i**-**j，其中 i，$j \in \{1,2\}$，i 表示梯度估计的选择，j 表示自适应矩阵更新规则的选择。请注意，**FedDA-MVR**对应于**FedDA-1-1**，因为在算法1中，我们选择了梯度估计的情况1和自适应矩阵更新的情况1。我们还介绍了更多细节，例如超参数选择。随后，在A.2节中，我们进行了若干消融研究，并更详细地将FedDA与其他基线方法进行比较；在A.3节中，我们纳入了基于CIFAR10构建异质数据集的实验；最后，在A.4节中，我们展示了当 $I = 1$时FedDA的形式，即没有本地步数。

A.1 第6节任务中的 FedDA 其他变体

A.1.1 具有稀疏约束的结直肠癌生存预测

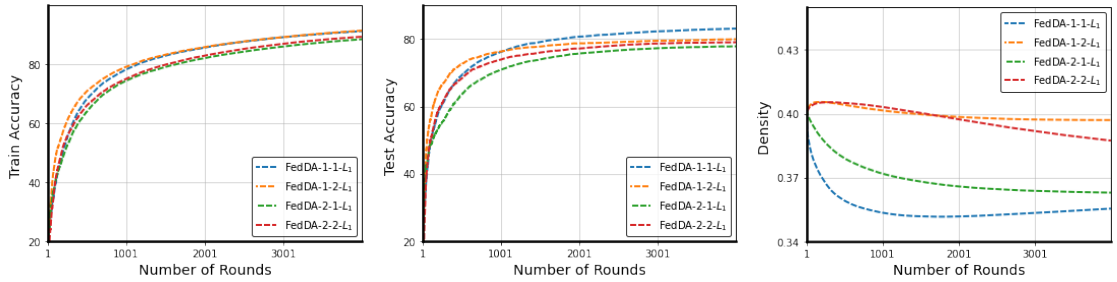

图4：PATHMNIST 数据集的结果。图中分别展示了训练准确率、测试准确率、密度与轮数（E 算法1中的）的关系。L_1 的后缀表示我们考虑了 L_1 约束。

不同 FedDA 变体的实验结果总结于图4中。如图所示，FedDA 的所有变体都取得了良好性能，但我们发现，按照密度指标测量，FedDA-MVR（FedDA-1-1）得到的模型最为稀疏。

A.1.2 使用 CIFAR10 和 FEMNIST 的图像分类任务

在这个无约束的联邦图像分类任务中，我们使用一个每层包含 64 个滤波器的 4 层卷积神经网络。对于 FEMNIST 数据集，每个全局轮次随机抽取 50 个用户。我们运行 20000 步（T），每隔 5 步（I）对状态取平均，并将小批量大小设为 16。对于其他超参数，我们采用网格搜索，并为每种方法选择最佳设置。在 CIFAR10 相关实验中，对于随机梯度下降方法，我们使用学习率 0.005；对于 FedCM算法，我们使用学习率 0.01，动量系数 α 设为 0.9；对于 FedAdam算法，我们使用本地学习率 0.001、全局学习率 0.002、动量系数 0.9，

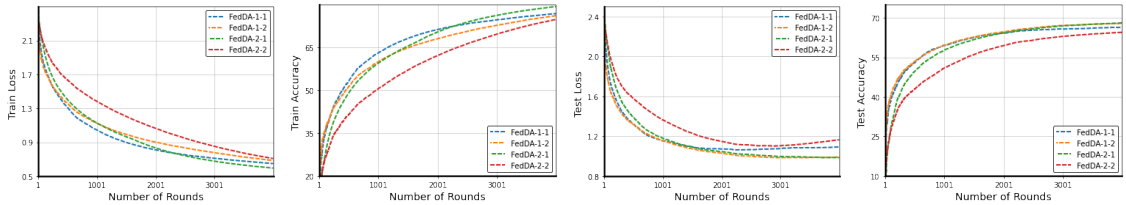

图5：CIFAR10数据集的结果。从左到右，分别展示训练损失、训练准确率、测试损失和测试准确率相对于全局轮次数量（算法1中的 E）的变化。

自适应矩阵的系数 β 设为 0.999；对于 Local-Adapt算法，我们使用本地学习率 0.001、全局学习率 0.002、动量系数 0.9，自适应矩阵的系数 β 设为 0.999；对于 Local-AMSGrad算法，我们使用学习率 0.001、动量系数 0.9、自适应矩阵系数 0.999；对于 MIME-MVR算法，我们使用学习率 0.1， w 100， c设为 2000；对于 STEM 算法，我们使用学习率 0.1， w 100 和 c 2000；对于我们的 FedDA-MVR，我们使用学习率 0.02， w 设为 10000， c设为 1000000， β 设为 0.999， τ 设为 0.01。对于 FedDA 的其他变体：对于 FedDA-2-1，我们使用学习率 0.001， α 设为 0.9， β 设为 0.999， τ 设为 0.01；对于 FedDA-1-2，我们使用学习率 1， w 设为 5000， c 设为 100， β 设为 0.999， τ 设为 0.01；对于 FedDA-2-2，我们使用学习率 0.01， α 0.9， β 设为 0.999， τ 设为 0.01。

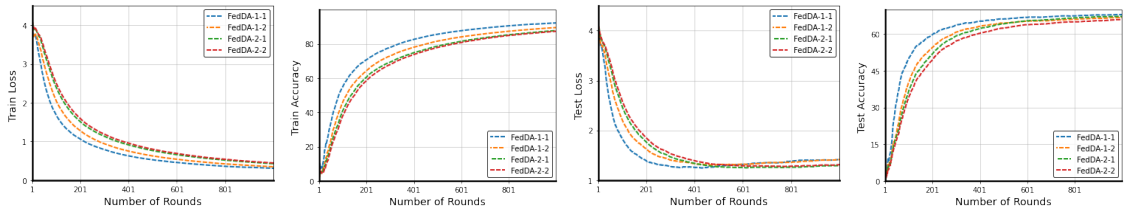

图6：FEMNIST数据集的结果。从左到右，分别展示训练损失、训练准确率、测试损失和测试准确率相对于全局轮次数量（算法1中的 E）的变化。

然后，在 FEMNIST 实验中，对于随机梯度下降方法，我们使用学习率 0.1；对于 FedCM算法，我们使用学习率 0.1，动量系数 α 为 0.9；对于 FedAdam算法，我们使用本地学习率 0.02、全局学习率 0.04、动量系数 0.9、自适应矩阵系数 β 为 0.999；对于 Local-Adapt算法，我们使用本地学习率 0.02、全局学习率 0.02、动量系数 0.9、自适应矩阵系数 β 为 0.999；对于 Local-AMSGrad算法，我们使用学习率 0.0005、动量系数 0.9、自适应矩阵系数 0.999；对于 MIME-MVR算法，我们使用学习率 1， w 10000， c 为 400；对于 STEM算法，我们使用学习率 1， w 10000 和 c 400；对于我们的 FedDA-MVR，我们使用学习率 0.02， w 为 10000， c 为 1000000， β 为 0.999， τ 为 0.01。对于 FedDA 的其他变体：对于 FedDA-2-1，我们使用学习率 0.001， α 为 0.9， β 为 0.999， τ 为 0.01；对于 FedDA-1-2，我们使用学习率 1， w 为 5000， c 为 100， β 为 0.999， τ 为 0.01；对于 FedDA-2-2，我们使用学习率 0.01， α 0.9， β 为 0.999， τ 为 0.01。

该 e FedDA 不同变体的实验结果总结于图5和图6中。A s

从图中可以看出，FedDA 的所有变体都表现良好。在大多数指标上，FedDA-MVR（FedDA-1-1）的表现最佳；我们观察到，其测试损失在训练后期出现了一定程度的过拟合。

A.2 实验结果的进一步讨论

在本小节中，我们将 FedDA 与其他基线方法进行了更详细的比较（实验基于同质 CIFAR10 数据集）。在图7中，我们比较了不同本地步数 I 下 FedCM、FedDA-2-1 和 FedDA-2-2 的表现。由于 FedDA-2-1 和 FedDA-2-2 未采用方差缩减加速，其优异表现表明，在我们的框架中使用自适应梯度是有效的。接下来，在图8中，我们比较了不同 I 下 Local-AMSGrad 与 FedDA-2-1 的表现。对于所有 I，FedDA-2-1 的表现都优于 Local-AMSGrad，且 I 越大，优势越明显。请注意，Local-AMSGrad 和 FedDA-2-1 都使用 Adam式自适应梯度（(6) 和 (7)），并且每个训练轮次的通信成本相同。在图9中，我们将 FedAdam 和 Local-Adapt 与 FedDA-2-1 进行了比较。所有方法都使用 Adam式自适应梯度。FedAdam 仅在服务器上执行自适应梯度计算；Local-Adapt 则同时执行本地和全局自适应梯度计算，但每个训练轮次都会刷新本地自适应梯度的状态。我们有两点观察：第一，Local-Adapt 相较于 FedAdam 的改进非常有限，这表明 Local-Adapt 使用的重启策略不如我们的方法有效；第二，FedAdam 和 Local-Adapt 都很少从增大 I 值中获益（与我们的 FedDA-2-1 相比）。对于 FedAdam，这表明仅在服务器层级应用自适应梯度存在局限。最后，在图10中，我们为 FedDA 的四种变体调整了 I。如图所示，我们的框架能够受益于增加本地步数。

A.3 异构 CIFAR10 上的图像分类任务

对于超参数，我们通过网格搜索为每种方法选择最佳设置。对于随机梯度下降方法，我们使用学习率 0.01；对于 FedCM算法，我们使用学习率 0.01，动量系数 α 为 0.9；对于 FedAdam算法，我们使用本地学习率 0.001、全局学习率 0.002、动量系数 0.9，自适应矩阵系数 β 为 0.999；对于 Local-Adapt算法，我们使用本地学习率 0.001、全局学习率 0.002、动量系数 0.9，自适应矩阵系数 β 为 0.999；对于 Local-AMSGrad算法，我们使用学习率 0.001、动量系数 0.9、自适应矩阵系数 0.999；对于 MIME-MVR算法，我们使用学习率 0.1，w 100，c 为 2000；对于 STEM算法，我们使用学习率 0.1，w 100 和 c 2000；对于我们的 FedDA-MVR/FedDA-1-1，我们使用学习率 0.02，w 为 10000，c 为 1000000，β 为 0.999，τ 为 0.01。对于 FedDA 的其他变体：对于 FedDA-2-1，我们使用学习率 0.001，α 为 0.9，β 为 0.999，τ 为 0.01；对于 FedDA-1-2，我们使用学习率 1，w 为 5000，c 为 100，β 为 0.999，τ 为 0.01；对于 FedDA-2-2，我们使用学习率 0.01，α 0.9，β 为 0.999，τ 为 0.01。

A.4 FedDA 的一个特殊情形：$I = 1$

为更好地说明 FedDA 的结构，本小节给出一个特殊情况的形式，即 $I = 1$。伪代码汇总于算法 3：在每个训练轮次，服务器首先通过式 (2) 获取新的原始状态（第 4 行）；随后，每个客户端（为简洁起见，我们假设所有客户端均参与）在本地更新梯度估计 ν_τ（第 6 行），服务器对这些状态取平均（第 8 行），并由服务器更新自适应矩阵（第 8 行）。

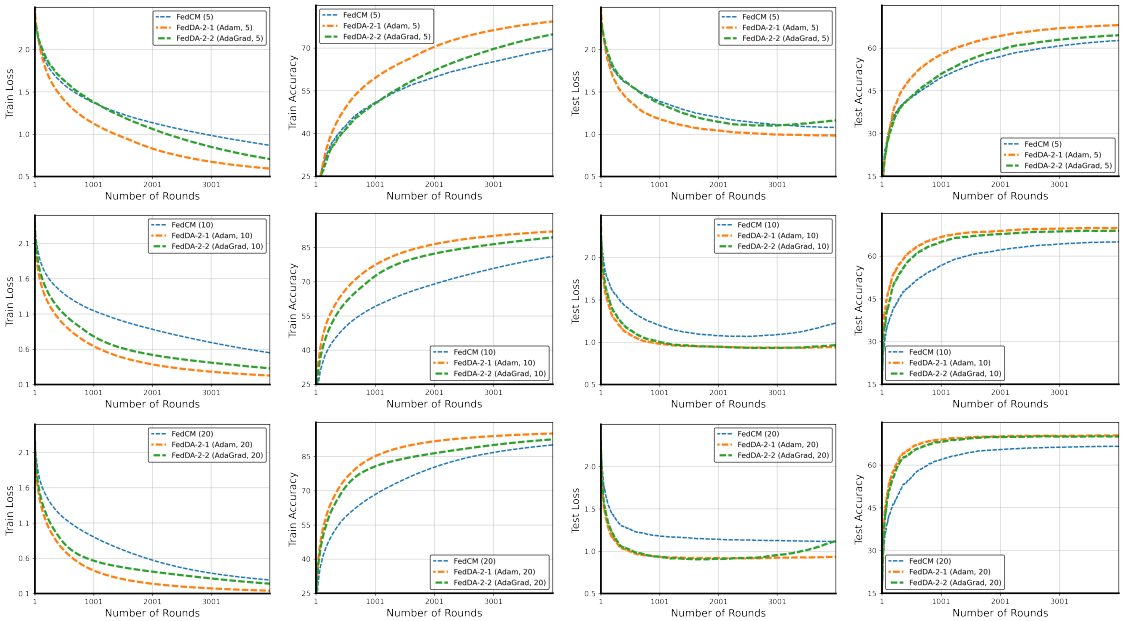

图 7：FedCM 与 FedDA-2-1 和 FedDA-2-2 的比较。从上到下分别展示$I = 5$、10、20。括号内的数字是 I 的值。

B 定理的证明

本节将对我们的算法进行收敛性分析。

B.1 预备命题

命题 1. 设 $\{\theta_k\}$、$k \in K$ 为 K 向量。则以下结论成立：$||\theta_i + \theta_j||^2 \leq (1+\lambda)||\theta_i||^2 + (1+\frac{1}{\lambda})||\theta_j||^2$ 对于任意 $a > 0$ 和 $||\sum_{k=1}^K \theta_k||^2 \leq K\sum_{k=1}^K ||\theta_k||^2$

命题 2. 对于有限序列 $z^{(k)} \in \mathbb{R}^d$，对 $k \in [K]$ 定义 $\bar{z} := \frac{1}{K}\sum_{k=1}^K z^{(k)}$，则有 $\sum_{k=1}^K \|z^{(k)} - \bar{z}\|^2 \leq \sum_{k=1}^K \|z^{(k)}\|^2$.

命题3. 设 $z_0 > 0$ 和 $z_1, z_2, \ldots, z_T \geq 0$。我们有 $\sum_{t=1}^T \frac{z_t}{z_0 + \sum_{i=1}^t z_i} \leq \log(1 + \frac{\sum_{i=1}^t z_i}{z_0})$。

这些命题都是标准结果。证明可参阅 [18] 中的引理 3（对应命题 1），以及 [22] 中的引理C.1和引理C.2（对应命题 2 和命题 3）。

B.2 局部更新中的预备引理

我们首先介绍一些记号。对于 $0 \leq i \leq I$，记作：

$$\psi_{\tau,i}^{(k)}(x) = -\langle x, z_{\tau,i}^{(k)} \rangle + \frac{1}{2\lambda}(x - x_{\tau,0}^{(k)})^T H_{\tau-1}(x - x_{\tau,0}^{(k)}), \tag{10}$$

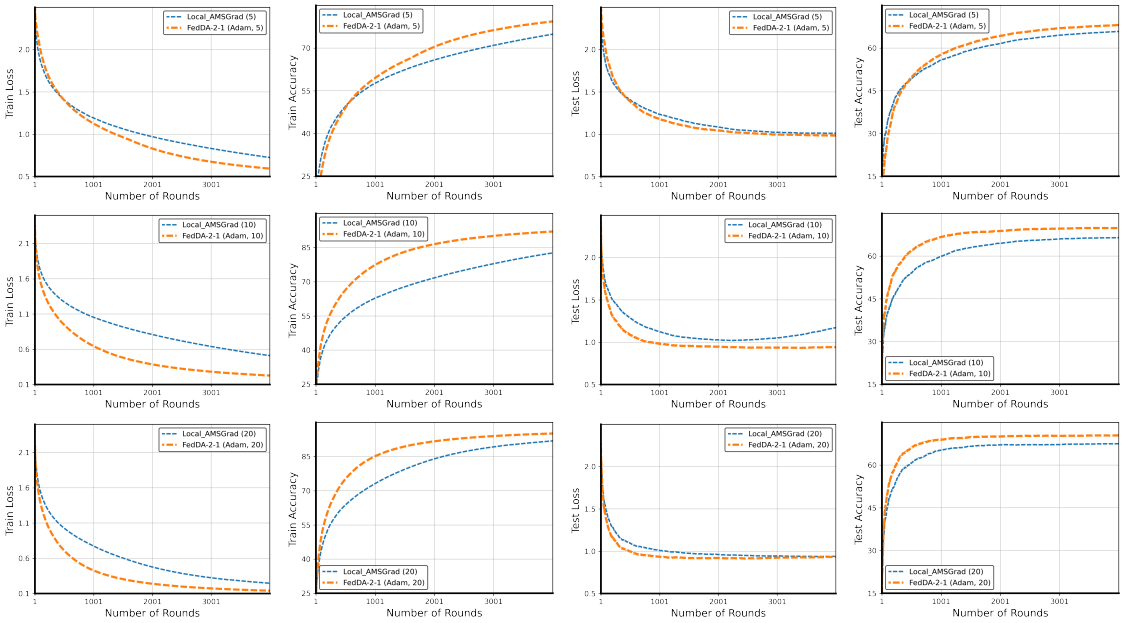

图8：Local-AMSGrad 与 FedDA-2-1 的比较。从上到下分别展示$I = 5$、10、20。括号内的数字是 I 的值。

然后，根据定义（算法 2 的第 4 行），我们有：

$$x_{\tau,i}^{(k)} = \arg\min_{x \in \mathcal{X}} \psi_{\tau,i}^{(k)}(x), \tag{11}$$

我们还定义

$$\tilde{\psi}_{\tau,i}(x) = -\langle x, \bar{z}_{\tau,i} \rangle + \frac{1}{2\lambda}(x - x_{\tau,0})^T H_{\tau-1}(x - x_{\tau,0}), \tag{12}$$

其中，$\bar{z}_{\tau,i} = \frac{1}{K}\sum_{k=1}^{K} z_{\tau i}^{(k)}$，是 $z_{\tau i}^{(k)}$，和 $x_{\tau,0} = x_\tau$ 的虚拟平均值。然后我们定义 e

$$\tilde{x}_{\tau,i} = \arg\min_{x \in \mathcal{X}} \tilde{\psi}_{\tau,i}(x), \tag{13}$$

注4。 在算法*1*中，在每个训练轮次 τ，我们仅从 K 个客户端中抽取 r 个客户端执行更新。对于 $k \notin \mathcal{S}_\tau$，我们为便于分析而定义相关变量，但实际上并不计算这些变量。

注5。 请注意，全局原始状态 $\tilde{x}_i$ 通常并非局部状态 $x_i^{(k)}$ 的算术平均值。

最后，我们还定义

$$\tilde{d}_{\tau,i} = \frac{1}{\eta_{\tau,i}}(\tilde{x}_{\tau,i} - \tilde{x}_{\tau,i+1}),\ d_{\tau,i}^{(k)} = \frac{1}{\eta_{\tau,i}}(x_{\tau,i}^{(k)} - x_{\tau,i+1}^{(k)}), k \in [K], i \in [I], \tag{14}$$

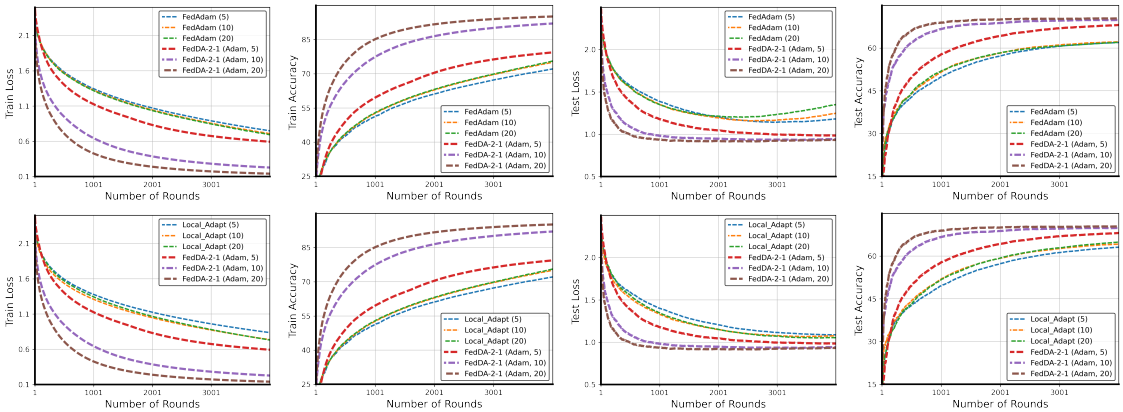

图 9：FedAdam 和 Local-Adapt 与 FedDA-2-1 的比较。括号内的数字是 I 的值。

算法 3 FedDA-分布式

1：输入：全局轮次数量 E**，调优参数** $\{\alpha_\tau, \beta_\tau, \eta_\tau\}_{i=1}^E$**；2：初始化：选择** $x_0 \in \mathcal{X}$**并计算** $\nu_0 = \frac{1}{K}\sum_{j=1}^K \nabla f^{(j)}(x_0, \mathcal{B}_0^{(k)})$**，其中** $\{\mathcal{B}_0^{(k)}\}_{k=1}^K$**是从每个** K **客户端中选取的小批量随机点；3：对于** $\tau = 0$ **到** $E-1$**：4：计算** $x_{\tau+1} = \arg\min_{x\in\mathcal{X}} \{\eta_{\tau+1}\langle x, \nu_\tau\rangle + \frac{1}{2\lambda}(x - x_\tau)^T H_\tau (x - x_\tau)\}$**；5：并行处理客户端** $k \in [K]$**：6：计算** $\nu_{\tau+1}^{(k)} = \mathcal{U}(\nu_\tau, x_{\tau+1}, x_\tau; \alpha_{\tau+1}, \mathcal{B}_{\tau+1}^{(k)})$**，其中** $\mathcal{B}_{\tau+1}^{(k)}$**是从客户端** k**中抽取的小批量随机样本；7：结束循环；8：计算** $\nu_{\tau+1} = \frac{1}{K}\sum_{k\in[K]} \nu_{\tau+1}^{(k)}$**和** $H_{\tau+1} = \mathcal{V}(H_\tau, \nu_\tau)$**；9：结束循环**

此外，回顾算法 2（第 6 行）的过程，我们有

$$\bar{\nu}_{\tau,i} = \frac{1}{\eta_{\tau,i}}(\bar{z}_{\tau,i} - \bar{z}_{\tau,i+1}),\ \nu_{\tau,i}^{(k)} = \frac{1}{\eta_{\tau,i}}(z_{\tau,i}^{(k)} - z_{\tau,i+1}^{(k)}),\ k \in [K],\ i \in [I], \tag{15}$$

注6。当上下文明确时，我们省略定义下标中的全局轮次 τ，即我们使用 $\psi_i^{(k)}(x)$、$\tilde{\psi}_i(x)$、$x_i^{(k)}$、$\tilde{x}_i$、$\tilde{d}_i$、$d_i^{(k)}$、$\bar{\nu}_i$、$\nu_i^{(k)}$ 和 H。

接下来，我们介绍一个与本地更新相关的引理。我们省略下标中的全局轮次编号 τ。

引理 1. 对于任意 $i \in [I]$ 和 $k \in [K]$，以下不等式成立：

1. $\lambda\langle \nu_i^{(k)}, d_i^{(k)}\rangle \geq \rho||d_i^{(k)}||^2, \lambda||\nu_i^{(k)}|| \geq \rho||d_i^{(k)}||$
2. $\lambda\langle \bar{\nu}_i, \tilde{d}_i\rangle \geq \rho||\tilde{d}_i||^2, \lambda||\bar{\nu}_i|| \geq \rho||\tilde{d}_i||;$

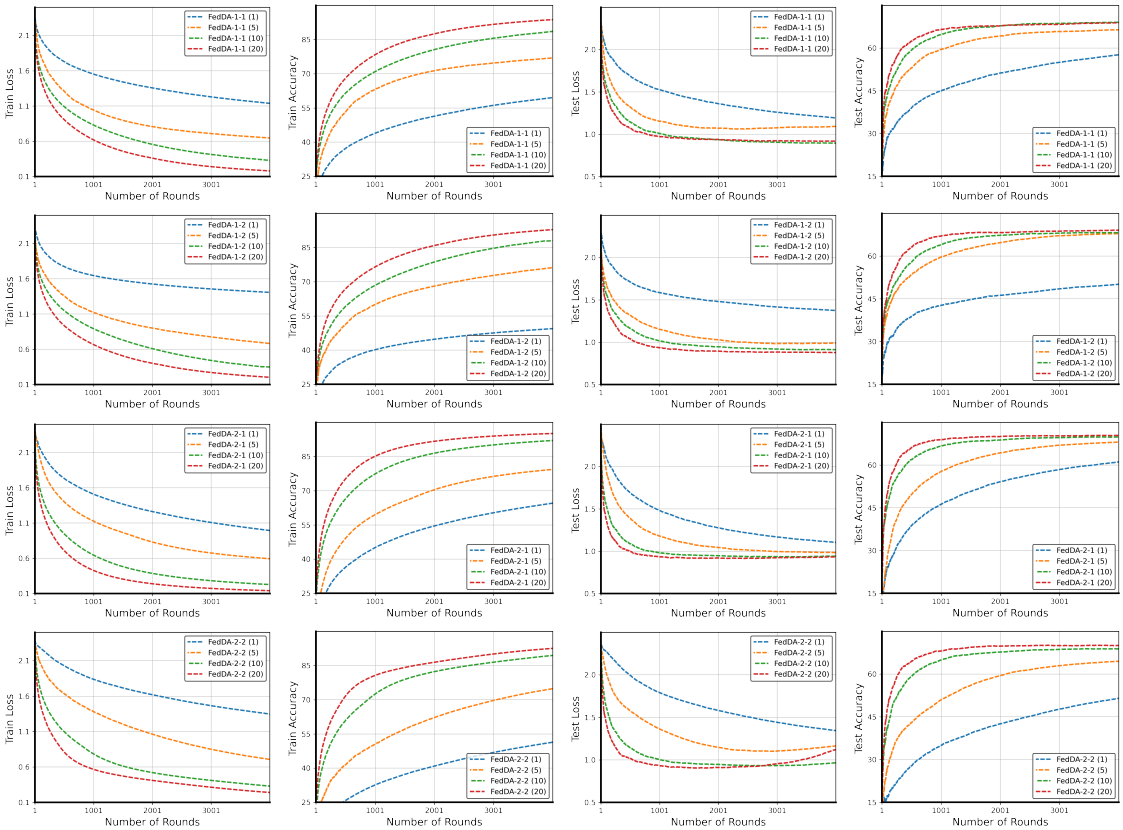

图10：本地步数 I 的消融实验。从上到下各行依次展示 FedDA-1-1、FedDA-1-2、FedDA-2-1 和 FedDA-2-2 的结果。括号内的数字是 I 的值。

3. $\lambda||z_i^{(k)} - \bar{z}_i|| \geq \rho||x_i^{(k)} - \tilde{x}_i||$;

证明. 第一项和第二项结论可通过类似的推导得到，这里仅给出第一项结论的推导。首先，若 $i = 1$, 则有

$$x_1^{(k)} = \arg\min_{x \in \mathcal{X}} \ -\langle x, z_1^{(k)} \rangle + \frac{1}{2\lambda}(x - x_0^{(k)})^T H(x - x_0^{(k)}),$$

根据一阶最优性条件，我们有:

$$\langle -z_1^{(k)} + \frac{1}{\lambda} H(x_1^{(k)} - x_0^{(k)}), u - x_1^{(k)} \rangle \geq 0, \ \forall \ u \in \mathcal{X},$$

取 $u = x_0^{(k)}$ 并利用事实 $z_1^{(k)} = -\eta_0 \nu_0$，我们得到:

$$\eta_0||\nu_0^{(k)}|| \times ||x_0^{(k)} - x_1^{(k)}|| \geq \eta_0 \langle \nu_0^{(k)}, x_0^{(k)} - x_1^{(k)} \rangle \geq \frac{1}{\lambda}(x_1^{(k)} - x_0^{(k)})^T H(x_1^{(k)} - x_0^{(k)}) \geq \frac{\rho}{\lambda}||x_0^{(k)} - x_1^{(k)}||^2$$

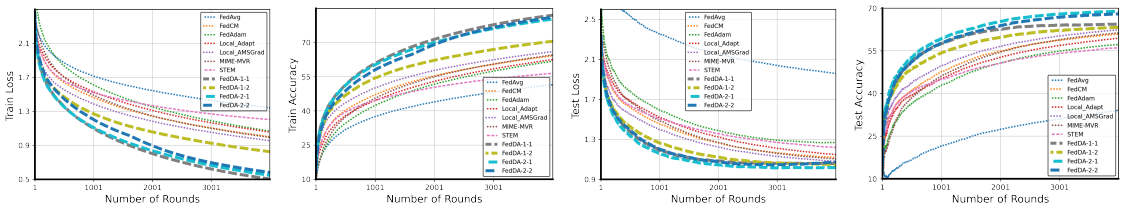

图11：异构 CIFAR10 数据集的结果。从左到右，分别展示了训练损失、训练准确率、测试损失和测试准确率关于 轮数（算法1中的 E）的变化。 I取值为 5。

在最左侧的不等式中，我们使用柯西-施瓦茨不等式；在最右侧的不等式中，我们使用自适应矩阵的强凸性假设，由此得到引理中的结论。

接下来，若 $i>0$，根据 $\psi_i^{(k)}(x)$ 的定义，我们有：

$$\psi_i^{(k)}(x_{i+1}^{(k)})-\psi_i^{(k)}(x_i^{(k)})=-\langle z_i^{(k)},x_{i+1}^{(k)}-x_i^{(k)}\rangle+\frac{1}{2\lambda}(x_{i+1}^{(k)}-x_i^{(k)})^TH(x_{i+1}^{(k)}+x_i^{(k)}-2x_0^{(k)}) \quad (16)$$

然后，根据 $x_i^{(k)}$ 的定义和一阶最优性条件，我们有

$$\langle -z_i^{(k)}+\frac{1}{\lambda}H(x_i^{(k)}-x_0^{(k)}),u-x_i^{(k)}\rangle\geq 0,\ \forall\ u\in\mathcal{X},$$

若取 $u=x_{i+1}^{(k)}$，则有 $-\langle z_i^{(k)},x_{i+1}^{(k)}-x_i^{(k)}\rangle\geq-\frac{1}{\lambda}(x_{i+1}^{(k)}-x_i^{(k)})^TH(x_i^{(k)}-x_0^{(k)})$，将该不等式代入式（16），我们有：

$$\begin{aligned}&\psi_i^{(k)}(x_{i+1}^{(k)})-\psi_i^{(k)}(x_i^{(k)})\\&\geq-\frac{1}{\lambda}(x_{i+1}^{(k)}-x_i^{(k)})^TH(x_i^{(k)}-x_0^{(k)})+\frac{1}{2\lambda}(x_{i+1}^{(k)}-x_i^{(k)})^TH(x_{i+1}^{(k)}+x_i^{(k)}-2x_0^{(k)})\\&\geq\frac{1}{2\lambda}(x_{i+1}^{(k)}-x_i^{(k)})^TH(x_{i+1}^{(k)}-x_i^{(k)})\end{aligned}$$

同理，对于 $\psi_{i+1}^{(k)}$，我们有：

$$\psi_{i+1}^{(k)}(x_{i+1}^{(k)})-\psi_{i+1}^{(k)}(x_i^{(k)})=-\langle z_{i+1}^{(k)},x_{i+1}^{(k)}-x_i^{(k)}\rangle+\frac{1}{2\lambda}(x_{i+1}^{(k)}-x_i^{(k)})^TH(x_{i+1}^{(k)}+x_i^{(k)}-2x_0^{(k)})$$

并且根据 $x_{i+1}^{(k)}$ 的定义和一阶最优性条件，我们可以得到

$$\langle -z_{i+1}^{(k)}+\frac{1}{\lambda}H(x_{i+1}^{(k)}-x_0^{(k)}),u-x_{i+1}^{(k)}\rangle\geq 0,\ \forall\ u\in\mathcal{X},$$

取 $u=x_i^{(k)}$，可得 $-\langle z_{i+1}^{(k)},x_{i+1}^{(k)}-x_i^{(k)}\rangle\leq-\frac{1}{\lambda}(x_{i+1}^{(k)}-x_i^{(k)})^TH(x_{i+1}^{(k)}-x_0^{(k)})$，将此不等式代入上述等式，可得：

$$\begin{aligned}&\psi_{i+1}^{(k)}(x_{i+1}^{(k)})-\psi_{i+1}^{(k)}(x_i^{(k)})\\&\leq-\frac{1}{\lambda}(x_{i+1}^{(k)}-x_i^{(k)})^TH(x_{i+1}^{(k)}-x_0^{(k)})+\frac{1}{2\lambda}(x_{i+1}^{(k)}-x_i^{(k)})^TH(x_{i+1}^{(k)}+x_i^{(k)}-2x_0^{(k)})\\&\leq-\frac{1}{2\lambda}(x_{i+1}^{(k)}-x_i^{(k)})^TH(x_{i+1}^{(k)}-x_i^{(k)})\end{aligned}$$

接下来，根据 $\psi_i^{(k)}(x)$ 和 $\psi_{i+1}^{(k)}(x)$ 的定义，可得：

$$\psi_{i+1}^{(k)}(x_{i+1}^{(k)}) - \psi_{i+1}^{(k)}(x_i^{(k)}) = \psi_i^{(k)}(x_{i+1}^{(k)}) - \psi_i^{(k)}(x_i^{(k)}) + \eta_i \langle \nu_i^{(k)}, x_{i+1}^{(k)} - x_i^{(k)} \rangle$$

最后，我们结合上述关系，得到：

$$\eta_i ||\nu_i^{(k)}|| \times ||x_i^{(k)} - x_{i+1}^{(k)}|| \geq \eta_i \langle \nu_i^{(k)}, x_i^{(k)} - x_{i+1}^{(k)} \rangle \geq \frac{1}{\lambda} (x_{i+1}^{(k)} - x_i^{(k)})^T H (x_{i+1}^{(k)} - x_i^{(k)}) \geq \rho ||x_i^{(k)} - x_{i+1}^{(k)}||^2$$

在最左侧的不等式中使用柯西-施瓦茨不等式，并在最右侧的不等式中使用自适应矩阵的强凸性假设，即可得到引理中所述的结论。

接下来，我们证明第三个结论。根据 $\psi_{i+1}^{(k)}$ 的定义，我们有：

$$\psi_{i+1}^{(k)}(x_{i+1}^{(k)}) - \psi_{i+1}^{(k)}(\tilde{x}_{i+1}) = -\langle z_{i+1}^{(k)}, x_{i+1}^{(k)} - \tilde{x}_{i+1} \rangle + \frac{1}{2\lambda} (x_{i+1}^{(k)} - \tilde{x}_{i+1})^T H (x_{i+1}^{(k)} + \tilde{x}_{i+1} - 2x_0^{(k)})$$

根据 $x_{i+1}^{(k)}$ 的定义和一阶最优性条件，我们有

$$\langle -z_{i+1}^{(k)} + \frac{1}{\lambda} H (x_{i+1}^{(k)} - x_0^{(k)}), u - x_{i+1}^{(k)} \rangle \geq 0, \ \forall \ u \in \mathcal{X},$$

取 $u = \tilde{x}_{i+1}$，我们有 $-\langle z_{i+1}^{(k)}, x_{i+1}^{(k)} - \tilde{x}_{i+1} \rangle \leq -\frac{1}{\lambda} (x_{i+1}^{(k)} - \tilde{x}_{i+1})^T H (x_{i+1}^{(k)} - x_0^{(k)})$。将此不等式代回上述不等式，可得：

$$\begin{aligned}
&\psi_{i+1}^{(k)}(x_{i+1}^{(k)}) - \psi_{i+1}^{(k)}(\tilde{x}_{i+1}) \\
&\leq -\frac{1}{\lambda} (x_{i+1}^{(k)} - \tilde{x}_{i+1})^T H (x_{i+1}^{(k)} - x_0^{(k)}) + \frac{1}{2\lambda} (x_{i+1}^{(k)} - \tilde{x}_{i+1})^T H (x_{i+1}^{(k)} + \tilde{x}_{i+1} - 2x_0^{(k)}) \\
&\leq -\frac{1}{2\lambda} (x_{i+1}^{(k)} - \tilde{x}_{i+1})^T H (x_{i+1}^{(k)} - \tilde{x}_{i+1})
\end{aligned}$$

˜那么，对于 $\psi_{i+1}(x)$，我们有：

$$\tilde{\psi}_{i+1}^{(k)}(x_{i+1}^{(k)}) - \tilde{\psi}_{i+1}(\tilde{x}_{i+1}) = -\langle \bar{z}_{i+1}, x_{i+1}^{(k)} - \tilde{x}_{i+1} \rangle + \frac{1}{2\lambda} (x_{i+1}^{(k)} - \tilde{x}_{i+1})^T H (x_{i+1}^{(k)} + \tilde{x}_{i+1} - 2\tilde{x}_0)$$

根据 $\tilde{x}_{i+1}$ 的定义和一阶最优性条件，我们有：

$$\langle -\bar{z}_{i+1} + \frac{1}{\lambda} H (\tilde{x}_{i+1} - \tilde{x}_0), u - \tilde{x}_{i+1} \rangle \geq 0, \ \forall \ u \in \mathcal{X},$$

取 $u = x_{i+1}^{(k)}$，我们有 $-\langle \bar{z}_{i+1}, x_{i+1}^{(k)} - \tilde{x}_{i+1} \rangle \geq -\frac{1}{\lambda} (x_{i+1}^{(k)} - \tilde{x}_{i+1})^T H (\tilde{x}_{i+1} - \tilde{x}_0)$。将此不等式代回上面的不等式，可得：

$$\begin{aligned}
&\tilde{\psi}_{i+1}(x_{i+1}^{(k)}) - \tilde{\psi}_{i+1}(\tilde{x}_{i+1}) \\
&\geq -\frac{1}{\lambda} (x_{i+1}^{(k)} - \tilde{x}_{i+1})^T H (\tilde{x}_{i+1} - \tilde{x}_0) + \frac{1}{2\lambda} (x_{i+1}^{(k)} - \tilde{x}_{i+1})^T H (x_{i+1}^{(k)} + \tilde{x}_{i+1} - 2\tilde{x}_0) \\
&\geq \frac{1}{2\lambda} (x_{i+1}^{(k)} - \tilde{x}_{i+1})^T H (x_{i+1}^{(k)} - \tilde{x}_{i+1})
\end{aligned}$$

接下来，由于我们有 $x_0^{(k)} = \tilde{x}_0$，根据 $\psi_{i+1}^{(k)}(x)$ 和 $\tilde{\psi}_{i+1}(x)$ 的定义，我们有 ve:

$$\psi_{i+1}^{(k)}(x_{i+1}^{(k)}) - \psi_{i+1}^{(k)}(\tilde{x}_{i+1}) = \tilde{\psi}_{i+1}(x_{i+1}^{(k)}) - \tilde{\psi}_{i+1}(\tilde{x}_{i+1}) - \langle z_{i+1}^{(k)} - \bar{z}_{i+1}, x_{i+1}^{(k)} - \tilde{x}_{i+1} \rangle$$

接下来，我们将上述关系结合起来，得到：

$$\begin{aligned} ||z_{i+1}^{(k)} - \bar{z}_{i+1}|| \times ||x_{i+1}^{(k)} - \tilde{x}_{i+1}|| &\geq \langle z_{i+!}^{(k)} - \bar{z}_{i+1}, x_{i+1}^{(k)} - \tilde{x}_{i+1} \rangle \\ &\geq \frac{1}{\lambda}(x_{i+1}^{(k)} - \tilde{x}_{i+1})^T H (x_{i+1}^{(k)} - \tilde{x}_{i+1}) \geq \rho ||x_{i+1}^{(k)} - \tilde{x}_{i+1}||^2 \end{aligned}$$

其中，第一个不等式由柯西-施瓦茨不等式得出，最后一个不等式则由 H的正定性得出。至此，引理中第一个不等式的证明完成。 □

B.3 状态共识误差

随着每个客户端执行本地更新，状态即 $z_{\tau\, i}^{(k)}$，和 $\nu_{\tau\, i}^{(k)}$，会逐渐偏离；以下引理将对这一差异进行界定。我们在下标中省略全局轮次编号 τ。

引理 2. 对于每个 $0 \le i \le I$，并假设迭代序列 $z_i^{(k)}$, $k \in [K]$ 由算法 *2* 生成，则有：

$$\sum_{k=1}^{K} \mathbb{E}\|z_i^{(k)} - \bar{z}_i\|^2 \le (I-1) \sum_{\ell=1}^{i-1} \eta_\ell^2 \sum_{k=1}^{K} \mathbb{E}\|\nu_\ell^{(k)} - \bar{\nu}_\ell\|^2,$$

其中，期望是关于算法随机性的期望。

证明。 根据算法 2，我们有 $z_0^{(k)} = \bar{z}_0 = 0$，此时引理中的不等式显然成立。否则，我们有

$$z_i^{(k)} = -\sum_{\ell=0}^{i-1} \eta_\ell \nu_\ell^{(k)} \quad \text{and} \quad \bar{z}_i = -\sum_{\ell=0}^{i-1} \eta_\ell \bar{\nu}_\ell.$$

因此，我们有：

$$\sum_{k=1}^{K} \|z_i^{(k)} - \bar{z}_i\|^2 = \sum_{k=1}^{K} \Big\| \sum_{\ell=1}^{i-1} \big(\eta_\ell \nu_\ell^{(k)} - \eta_\ell \bar{\nu}_\ell\big) \Big\|^2 \le (I-1) \sum_{\ell=1}^{i-1} \eta_\ell^2 \sum_{k=1}^{K} \|\nu_\ell^{(k)} - \bar{\nu}_\ell\|^2$$

其中，等号利用了以下事实：对于 $k \in [K]$, $\nu_0^{(k)} = \nu_0$；不等式利用了命题 1 以及我们有 $i \le I$ 这一事实。对上述不等式两边取期望，即可得到引理中的结论。证明完毕。

□

引理 3. 对于 $i \in [I]$，我们有：

$$\sum_{k=1}^{K} ||d_i^{(k)} - \tilde{d}_i||^2 \le \frac{4\lambda^2 (I-1)}{\rho^2 \eta_i^2} \sum_{\ell=1}^{i} \eta_\ell^2 \sum_{k=1}^{K} \mathbb{E}\|\nu_\ell^{(k)} - \bar{\nu}_\ell\|^2$$

其中，期望是关于算法随机性的。

证明。 首先，当 $i = 0$, $x_0^{(k)} = \tilde{x}_0$, $z_1^{(k)} = \bar{z}_1$ 时，根据算法 2 的第 5 行，我们有 $x_1^{(k)} = \tilde{x}_1$，进而得到 $\eta_0 d_0^{(k)} = x_0^{(k)} - x_1^{(k)} = \tilde{x}_0 - \tilde{x}_1 = \eta_t \tilde{d}_0$，引理中的不等式显然成立。

接下来，当 $i > 0$ 时，我们有：

$$\begin{aligned} \eta_i^2 ||d_i^{(k)} - \tilde{d}_i||^2 &= ||x_i^{(k)} - x_{i+1}^{(k)} - (\tilde{x}_i - \tilde{x}_{i+1})||^2 \le 2||x_i^{(k)} - \tilde{x}_i||^2 + 2||x_{i+1}^{(k)} - \tilde{x}_{i+1}||^2 \\ &\le \frac{2\lambda^2}{\rho^2} \big(||z_i^{(k)} - \bar{z}_i||^2 + ||z_{i+1}^{(k)} - \bar{z}_{i+1}||^2 \big) \end{aligned}$$

The 最后一个不等式用到了引理 1 的断言 3。对 $k \in [K]$ 求和并应用引理 2，我们有 ：

$$\begin{aligned} \rho^2 \eta_i^2 \sum_{k=1}^{K} ||d_i^{(k)} - \tilde{d}_i||^2 &\le 2\lambda^2 (I-1) \sum_{\ell=1}^{i-1} \eta_\ell^2 \sum_{k=1}^{K} \mathbb{E}\|\nu_\ell^{(k)} - \bar{\nu}_\ell\|^2 + 2\lambda^2 (I-1) \sum_{\ell=1}^{i} \eta_\ell^2 \sum_{k=1}^{K} \mathbb{E}\|\nu_\ell^{(k)} - \bar{\nu}_\ell\|^2 \\ &\le 4\lambda^2 (I-1) \sum_{\ell=1}^{i} \eta_\ell^2 \sum_{k=1}^{K} \mathbb{E}\|\nu_\ell^{(k)} - \bar{\nu}_\ell\|^2 \end{aligned}$$

证明至此完成。 □

B.4 下降引理

在本　　小节中，我们对虚拟序列 $\tilde{x}_{\tau,i}$ 上的函数值 $f(\tilde{x}_{\tau,i})$ 的下降幅度进行界定。

引理 4. 假设序列 $\{x_{\tau,i}^{(k)}\}_{i=0}^{I-1}$ 由算法 *2* 生成，则有

$$f(x_{\tau+1}) \leq f(x_\tau) - \sum_{i=0}^{I-1}\left(\frac{3\rho\eta_{\tau+1,i}}{4\lambda} - \frac{\eta_{\tau+1,i}^2 L}{2}\right)\|\tilde{d}_{\tau+1,i}\|^2 + \sum_{i=0}^{I-1}\frac{\lambda\eta_{\tau+1,i}}{\rho}\left\|\bar{e}_{\tau+1,i}\right\|^2,$$

其中 $\bar{e}_{\tau,i} = \bar{\nu}_{\tau,i} - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(\tilde{x}_{\tau,i})$。

证明。 由于函数 $f(x)$ 是 L-光滑的，我们有（为简化记号，省略全局轮次编号 τ）：

$$\begin{aligned}
f(\tilde{x}_{i+1}) &\leq f(\tilde{x}_i) + \langle\nabla f(\tilde{x}_i), \tilde{x}_{i+1} - \tilde{x}_i\rangle + \frac{L}{2}\|\tilde{x}_{i+1} - \tilde{x}_i\|^2 = f(\tilde{x}_i) - \eta_i\langle\nabla f(\tilde{x}_i), \tilde{d}_i\rangle + \frac{L\eta_i^2}{2}\|\tilde{d}_i\|^2 \\
&= f(\tilde{x}_i) - \eta_i\langle\bar{\nu}_i, \tilde{d}_i\rangle - \eta_i\langle\nabla f(\tilde{x}_i) - \bar{\nu}_i, \tilde{d}_i\rangle + \frac{L\eta_i^2}{2}\|\tilde{d}_i\|^2 \\
&\overset{(a)}{\leq} f(\tilde{x}_i) - (\frac{\rho\eta_i}{\lambda} - \frac{L\eta_i^2}{2})\|\tilde{d}_i\|^2 - \eta_i\langle\nabla f(\tilde{x}_i) - \bar{\nu}_i, \tilde{d}_i\rangle \\
&\overset{(b)}{\leq} f(\tilde{x}_i) - \left(\frac{\rho\eta_i}{\lambda} - \frac{\eta_i^2 L}{2}\right)\|\tilde{d}_i\|^2 + \frac{\rho\eta_i}{4\lambda}\|\tilde{d}_i\|^2 + \frac{\lambda\eta_i}{\rho}\|\bar{\nu}_i - \nabla f(\tilde{x}_i)\|^2 \\
&\overset{(c)}{\leq} f(\tilde{x}_i) - \left(\frac{3\rho\eta_i}{4\lambda} - \frac{\eta_i^2 L}{2}\right)\|\tilde{d}_i\|^2 + \frac{\lambda\eta_i}{\rho}\|\bar{e}_i\|^2
\end{aligned}$$

在不等式 (a) 中，我们使用引理 1 的结论 1；不等式 (b) 使用杨氏不等式；不等式 (c) 表示 $\bar{e}_i = \bar{\nu}_i - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(\tilde{x}_i)$。

对于 τ 全局轮次，我们对 $i = 0$ 到 $I-1$求和，可得：

$$f(\tilde{x}_{\tau+1,I}) \leq f(\tilde{x}_{\tau+1,0}) - \sum_{i=0}^{I-1}\left(\frac{3\rho\eta_{\tau+1,i}}{4\lambda} - \frac{\eta_{\tau+1,i}^2 L}{2}\right)\|\tilde{d}_{\tau+1,i}\|^2 + \sum_{i=0}^{I-1}\frac{\lambda\eta_{\tau+1,i}}{\rho}\left\|\bar{e}_{\tau+1,i}\right\|^2,$$

遵循算法1和算法 2 中的更新规则，我们有 $\tilde{x}_{\tau+1,\,0} = x_\tau$ 以及 $\tilde{x}_{\tau+1,\,I} = x_{\tau+1}$。证明完毕。

□

B.5 梯度误差收缩

在本小节中，我们对梯度估计误差 $\bar{e}_{\tau,\,i}$进行界定，其中 $\bar{e}_{\tau,\,i} = \bar{\nu}_{\tau,\,i} - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(\tilde{x}_{\tau,\,i})$ 的定义见引理 4；此外，我们还将全局梯度估计误差 e_τ定义为 $e_\tau = \nu_\tau - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(x_\tau) = \nu_\tau - \nabla f(x_\tau)$。注意，我们有 $e_\tau = \bar{e}_{\tau,\,I} = \bar{e}_{\tau+1,\,0}$。我们首先给出关于 $\bar{e}_0$（即初始梯度估计误差）的一个事实。

引理 5. 对于 $e_0 := \nu_0 - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(x_0)$，假设我们选择的小批量大小为 $|\mathcal{B}_0^{(k)}| = b, k \in [K]$, 则有： $\mathbb{E}\|e_0\|^2 \leq \frac{\sigma^2}{bK}$. —

证明。 根据算法1的第1行，我们有：

$$\begin{aligned}\mathbb{E}\|e_0\|^2 &= \mathbb{E}\left\|\nu_0 - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(x_0)\right\|^2 \\ &= \mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(x_0;\mathcal{B}_0^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(x_0)\right\|^2 \\ &\overset{(a)}{\leq} \frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla f^{(k)}(x_0;\mathcal{B}_0^{(k)}) - \nabla f^{(k)}(x_0)\right\|^2 \overset{(b)}{\leq} \frac{\sigma^2}{bK}.\end{aligned}$$

其中，(a)由以下内容得出：根据无偏梯度假设，我们有： $\mathbb{E}\big[\nabla f^{(k)}(x_0^{(k)};\mathcal{B}_0^{(k)})\big] = \nabla f^{(k)}(x_0^{(k)})$，对所有 $k \in [K]$均成立。此外，对于所有 k， $\ell \in [K]$ 和 $k \neq \ell$， k^{th} 和 ℓ^{th} 客户端的样本 $\mathcal{B}_0^{(k)}$ 和 $\mathcal{B}_0^{(\ell)}$ 均匀随机抽取，且彼此独立。

$$\begin{aligned}&\mathbb{E}\left[\left\langle (x_0^{(k)};\mathcal{B}_0^{(k)}) - \nabla f^{(k)}(x_0)\big), \big(\nabla f^{(\ell)}(x_0^{(\ell)};\mathcal{B}_0^{(\ell)}) - \nabla f^{(\ell)}(\bar{x}_0)\big)\right\rangle\right] \\ &= \mathbb{E}\left[\left\langle \underbrace{\mathbb{E}\big[\nabla f^{(k)}(x_0^{(k)};\mathcal{B}_0^{(k)}) - \nabla f^{(k)}(x_0^{(k)})\big]}_{=0}, \underbrace{\mathbb{E}\big[\nabla f^{(\ell)}(x_0^{(\ell)};\mathcal{B}_0^{(\ell)}) - \nabla f^{(\ell)}(x_0^{(\ell)})\big]}_{=0}\right\rangle\right] = 0.\end{aligned}$$

不等式 (c) 由有界方差假设得出。证明至此完成。 □

引理 6. 定义 $\bar{e}_{\tau,\,i} := \bar{\nu}_{\tau,\,i} - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(\tilde{x}_{\tau,\,i})$，则对于每个 $\tau \geq 1$ 和 $i \geq 0$，假设 $\alpha_i < 1$且客户端在训练中使用批量大小 b_1 ，那么我们有：

$$\begin{aligned}\mathbb{E}\|\bar{e}_{\tau,i}\|^2 \leq (1-\alpha_{\tau,i})^2\mathbb{E}\|\bar{e}_{\tau,i-1}\|^2 + \frac{40\lambda^2(I-1)L^2}{\rho^2K^2}\sum_{\ell=1}^{i-1}\eta_{\tau,\ell}^2\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau,\ell}^{(k)} - \bar{\nu}_{\tau,\ell}\|^2 \\ + \frac{8\eta_{\tau,i-1}^2L^2}{K}\mathbb{E}\|\tilde{d}_{\tau,i-1}\|^2 + \frac{4\alpha_{\tau,i}^2\sigma^2}{b_1K}\end{aligned}$$

其中期望是相对于算法的随机性而言的。

证明。考虑 考虑误差项 $\|\bar{e}_i\|^2, i \geq 1$ （为简化记法，我们省略全局轮次编号 τ），

我们有:

$$
\begin{aligned}
\mathbb{E}\|\bar{e}_i\|^2 &= \mathbb{E}\left\|\bar{\nu}_i - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(\tilde{x}_i)\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)}) + (1-\alpha_i)\left(\bar{\nu}_{i-1} - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})\right) - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(\tilde{x}_i)\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\Bigg(\Big(\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(\tilde{x}_i)\Big)\right. \\
&\qquad\qquad \left. - (1-\alpha_i)\Big(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(\tilde{x}_{i-1})\Big)\Bigg) + (1-\alpha_i)\bar{e}_{i-1}\right\|^2 \\
&= (1-\alpha_i)^2\mathbb{E}\|\bar{e}_{i-1}\|^2 + \frac{1}{K^2}\mathbb{E}\left\|\sum_{k=1}^{K}\Big[\Big(\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(\tilde{x}_i)\Big)\right. \\
&\qquad\qquad \left. - (1-\alpha_i)\left(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(\tilde{x}_{i-1})\right)\Big]\right\|^2 \\
&= (1-\alpha_i)^2\mathbb{E}\|\bar{e}_{i-1}\|^2 + \frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}\left\|\Big(\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(\tilde{x}_i)\Big)\right. \\
&\qquad\qquad \left. - (1-\alpha_i)\Big(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(\tilde{x}_{i-1})\Big)\right\|^2,
\end{aligned}
$$

其中，第一个等号使用了 $\bar{\nu}_i$ 的定义；最后一个等号是通过展开范数并计算 $k\in[K]$ 之间的内积得出的。由于不同工作节点上的样本是独立抽取的，交叉项的期望为零。现在我们来考虑上式中的第二项:

$$
\begin{aligned}
&\mathbb{E}\left\|\big(\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(\tilde{x}_i)\big) - (1-\alpha_i)\big(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(\tilde{x}_{i-1})\big)\right\|^2 \\
&= \mathbb{E}\left\|\big(\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(x_i^{(k)})\big) - (1-\alpha_i)\big(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(x_{i-1}^{(k)})\big)\right. \\
&\qquad \left. + \nabla f^{(k)}(x_i^{(k)}) - \nabla f^{(k)}(\tilde{x}_i) - (1-\alpha_i)\big(\nabla f^{(k)}(x_{i-1}^{(k)}) - \nabla f^{(k)}(\tilde{x}_{i-1})\big)\right\|^2 \\
&\le 2\mathbb{E}\left\|\big(\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(x_i^{(k)})\big) - (1-\alpha_i)\big(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(x_{i-1}^{(k)})\big)\right\|^2 \\
&\qquad + 2\mathbb{E}\left\|\nabla f^{(k)}(x_i^{(k)}) - \nabla f^{(k)}(\tilde{x}_i) - (1-\alpha_i)\big(\nabla f^{(k)}(x_{i-1}^{(k)}) - \nabla f^{(k)}(\tilde{x}_{i-1})\big)\right\|^2
\end{aligned}
$$

对于上述不等式的第一项，我们有：

$$
\begin{aligned}
&\mathbb{E}\left\|\left(\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_i^{(k)})\right)-(1-\alpha_i)\left(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_{i-1}^{(k)})\right)\right\|^2\\
&=\mathbb{E}\left\|(1-a_i)\left[\left(\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_i^{(k)})\right)-\left(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_{i-1}^{(k)})\right)\right]\right.\\
&\qquad\left.+\alpha_i\left(\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_i^{(k)})\right)\right\|^2\\
&\le 2(1-\alpha_i)^2\mathbb{E}\left\|\left(\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})\right)-\left(\nabla f^{(k)}(x_i^{(k)})-\nabla f^{(k)}(x_{i-1}^{(k)})\right)\right\|^2\\
&\qquad+2\alpha_i^2\mathbb{E}\left\|\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_i^{(k)})\right\|^2\\
&\le 2(1-\alpha_i)^2\mathbb{E}\left\|\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})\right\|^2+2\alpha_i^2\sigma^2/b_1\\
&\le 2(1-\alpha_i)^2L^2\mathbb{E}\|x_i^{(k)}-x_{i-1}^{(k)}\|^2+2a_i^2\sigma^2/b_1\le 2(1-\alpha_i)^2L^2\eta_{i-1}^2\mathbb{E}\|d_{i-1}^{(k)}\|^2+2\alpha_i^2\sigma^2/b_1\\
&\le 4(1-\alpha_i)^2L^2\eta_{i-1}^2\mathbb{E}\|d_{i-1}^{(k)}-\tilde{d}_{i-1}\|^2+4(1-\alpha_i)^2L^2\eta_{i-1}^2\mathbb{E}\|\tilde{d}_{i-1}\|^2+2\alpha_i^2\sigma^2/b_1
\end{aligned}
$$

其中，第一个不等式使用了命题 1，第二个不等式使用了有界方差假设。对于第二个不等式，我们有：

$$
\begin{aligned}
&\mathbb{E}\left\|\nabla f^{(k)}(x_i^{(k)})-\nabla f^{(k)}(\tilde{x}_i)-(1-\alpha_i)\left(\nabla f^{(k)}(x_{i-1}^{(k)})-\nabla f^{(k)}(\tilde{x}_{i-1})\right)\right\|^2\\
&\overset{(a)}{\le} 2\mathbb{E}\left\|\nabla f^{(k)}(x_i^{(k)})-\nabla f^{(k)}(\tilde{x}_i)\right\|^2+2\mathbb{E}\left\|(1-\alpha_i)\left(\nabla f^{(k)}(x_{i-1}^{(k)})-\nabla f^{(k)}(\tilde{x}_{i-1})\right)\right\|^2\\
&\le 2L^2\mathbb{E}\left\|x_i^{(k)}-\tilde{x}_i\right\|^2+2L^2(1-\alpha_i)^2\mathbb{E}\left\|x_{i-1}^{(k)}-\tilde{x}_{i-1}\right\|^2\\
&\overset{(b)}{\le}\frac{2\lambda^2L^2}{\rho^2}\mathbb{E}\left\|z_i^{(k)}-\bar{z}_i\right\|^2+\frac{2\lambda^2L^2(1-\alpha_i)^2}{\rho^2}\mathbb{E}\left\|z_{i-1}^{(k)}-\bar{z}_{i-1}\right\|^2
\end{aligned}
$$

其中 (a)　使用命题 1；(b) 使用引理 1 的断言 3；接下来，我们将上述不等式结合起来　　es

合并后得到:

$$\begin{aligned}
\mathbb{E}\|\bar{e}_i\|^2 &\le (1-\alpha_i)^2\mathbb{E}\|\bar{e}_{i-1}\|^2 + \frac{4\alpha_i^2\sigma^2}{b_1K} + \frac{8(1-\alpha_i)^2\eta_{i-1}^2L^2}{K^2}\sum_{k=1}^{K}\mathbb{E}\|d_{i-1}^{(k)} - \tilde{d}_{i-1}\|^2 \\
&\quad + \frac{8(1-\alpha_i)^2\eta_{i-1}^2L^2}{K}\mathbb{E}\|\tilde{d}_{i-1}\|^2 + \frac{4\lambda^2L^2}{K^2\rho^2}\sum_{k=1}^{K}\mathbb{E}\left\|z_i^{(k)} - \bar{z}_i\right\|^2 \\
&\quad + \frac{4\lambda^2L^2(1-\alpha_i)^2}{K^2\rho^2}\sum_{k=1}^{K}\mathbb{E}\left\|z_{i-1}^{(k)} - \bar{z}_{i-1}\right\|^2 \\
&\le (1-\alpha_i)^2\mathbb{E}\|\bar{e}_{i-1}\|^2 + \frac{4\alpha_i^2\sigma^2}{b_1K} + \frac{32\lambda^2(I-1)(1-\alpha_i)^2L^2}{K^2\rho^2}\sum_{\ell=1}^{i-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)} - \bar{\nu}_\ell\|^2 \\
&\quad + \frac{8(1-\alpha_i)^2\eta_{i-1}^2L^2}{K}\mathbb{E}\|\tilde{d}_{i-1}\|^2 + \frac{4\lambda^2(I-1)L^2}{K^2\rho^2}\sum_{\ell=1}^{i-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)} - \bar{\nu}_\ell\|^2 \\
&\quad + \frac{4\lambda^2(I-1)L^2(1-\alpha_{i-1})^2}{K^2\rho^2}\sum_{\ell=1}^{i-2}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)} - \bar{\nu}_\ell\|^2 \\
&\le (1-\alpha_i)^2\mathbb{E}\|\bar{e}_{i-1}\|^2 + \frac{4\alpha_i^2\sigma^2}{b_1K} + \frac{40\lambda^2(I-1)L^2}{K^2\rho^2}\sum_{\ell=1}^{i-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)} - \bar{\nu}_\ell\|^2 + \frac{8\eta_{i-1}^2L^2}{K}\mathbb{E}\|\tilde{d}_{i-1}\|^2,
\end{aligned}$$

第二个不等式使用引理 2 和引理 3，最后一个不等式使用假设$\alpha_i < 1$。证明至此完成。

□

引理 7. 对于 $\tau \ge 0$。假设我们选择 $\eta_{\tau,\,i} = \kappa/(\omega_i + i + \tau I)^{1/3}$，此外，假设满足 $\alpha_i < 1$,$w_i \le w_{i-1}$, $w_i \ge 2$, $\eta_{\tau,\,i} \le \frac{\rho}{48\lambda LI^2}$，则有:

$$\begin{aligned}
\frac{\rho K}{64L^2}\left(\frac{\mathbb{E}\|\bar{e}_{\tau+1}\|^2}{\eta_{\tau+1,I-1}} - \frac{\mathbb{E}\|\bar{e}_\tau\|^2}{\eta_{\tau,I-1}}\right) &\le -\sum_{i=0}^{I-1}\frac{3\eta_{\tau+1,i}}{2\rho}\mathbb{E}\|\bar{e}_{\tau+1,i}\|^2 + \sum_{i=0}^{I-1}\frac{\eta_{\tau+1,i}\rho}{8}\mathbb{E}\|\tilde{d}_{\tau+1,i}\|^2 + \sum_{i=0}^{I-1}\frac{\sigma^2c^2\eta_{\tau+1,i}^3\rho}{16L^2} \\
&\quad + \frac{5I(I-1)}{4K\rho}\sum_{\ell=1}^{I}\eta_{\tau+1,\ell}\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau+1,\ell}^{(k)} - \bar{\nu}_{\tau+1,\ell}\|^2
\end{aligned}$$

证明。 在全局轮次 $\tau - 1$ 应用引理 6，则对于 $i \ge 0$ （我们记 $\eta_{\tau,\,-1} = \eta_{\tau-1,\,I-1}$ 对所有

$\tau \geq 1$），我们有：

$$\begin{aligned}
&\frac{\mathbb{E}\|\bar{e}_{\tau,i+1}\|^2}{\eta_{\tau,i}} - \frac{\mathbb{E}\|\bar{e}_{\tau,i}\|^2}{\eta_{\tau,i-1}} \\
&\leq \left[\frac{(1-a_{\tau,i+1})^2}{\eta_{\tau,i}} - \frac{1}{\eta_{\tau,i-1}}\right]\mathbb{E}\|\bar{e}_{\tau,i}\|^2 + \frac{40\lambda^2(I-1)L^2}{\rho^2K^2\eta_{\tau,i}}\sum_{\ell=1}^{i}\eta_{\tau,\ell}^2\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau,\ell}^{(k)} - \bar{\nu}_{\tau,\ell}\|^2 \\
&\qquad + \frac{8L^2\eta_{\tau,i}}{K}\mathbb{E}\|\tilde{d}_{\tau,i}\|^2 + \frac{4a_{\tau,i+1}^2\sigma^2}{\eta_{\tau,i}b_1K} \\
&\overset{(a)}{\leq} \left(\eta_{\tau,i}^{-1} - \eta_{\tau,i-1}^{-1} - c\eta_{\tau,i}\right)\mathbb{E}\|\bar{e}_{\tau,i}\|^2 + \frac{80\lambda^2(I-1)L^2}{\rho^2K^2}\sum_{\ell=1}^{i}\eta_{\tau,\ell}\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau,\ell}^{(k)} - \bar{\nu}_{\tau,\ell}\|^2 \\
&\qquad + \frac{8L^2\eta_{\tau,i}}{K}\mathbb{E}\|\tilde{d}_{\tau,i}\|^2 + \frac{4\sigma^2c^2\eta_{\tau,i}^3}{b_1K},
\end{aligned}$$

其中，不等式（a）利用了以下事实：对于所有 $i \in [I]$，均有 $(1-\alpha_{\tau,i})^2 \leq 1-\alpha_{\tau,i} \leq 1$ 以及 $a_{\tau,i+1} = c\eta_{\tau,i}^2$；此外还利用了以下事实：假设我们选择 $\eta_{\tau,i} = \kappa/(\omega_i + i + \tau I)^{1/3}$，那么对于 $0 \leq l \leq i < I$，有：

$$\begin{aligned}
\frac{\eta_{\tau,l}}{\eta_{\tau,i}} &= \frac{(w_i+i+\tau I)^{1/3}}{(w_l+l+\tau I)^{1/3}} = \left(1 + \frac{w_i+i-w_l-l}{w_l+l+\tau I}\right)^{1/3} \\
&\leq \left(1 + \frac{(I-1)}{w_l+l+\tau I}\right)^{1/3} \leq 1 + \frac{(I-1)}{3(w_l+l+\tau I)} \leq 2
\end{aligned} \tag{17}$$

第一个不等式依据事实 $0 < i-l < I-1$ 得出；倒数第二个不等式利用了 $x^{1/3}$ 的凹性，即：$(x+y)^{1/3} - x^{1/3} \leq y/3x^{2/3}$；最后一个不等式则利用了以下事实：$w_l \geq 0$、$I \geq 1$、$l \geq 0$、$\tau \geq 1$。

对于差值 $\eta_i^{-1} - \eta_{i-1}^{-1}$，有：

$$\begin{aligned}
\frac{1}{\eta_{\tau,i}} - \frac{1}{\eta_{\tau,i-1}} &= \frac{(w_i+i+\tau I)^{1/3}}{\kappa} - \frac{(w_{i-1}+i-1+\tau I)^{1/3}}{\kappa} \\
&\overset{(a)}{\leq} \frac{(w_i+i+\tau I)^{1/3}}{\kappa} - \frac{(w_i+i-1+\tau I)^{1/3}}{\kappa} \\
&\overset{(b)}{\leq} \frac{1}{3\kappa(w_i+i-1+\tau I)^{2/3}} \overset{(c)}{\leq} \frac{2^{2/3}\kappa^2}{3\kappa^3(w_i+i+\tau I)^{2/3}} \overset{(d)}{=} \frac{2^{2/3}}{3\kappa^3}\eta_i^2 \overset{(e)}{\leq} \frac{\rho}{72\kappa^3\lambda LI^2}\eta_i,
\end{aligned} \tag{18}$$

其中，不等式（a）是因为我们选择了 $w_i \leq w_{i-1}$；（b）由 $x^{1/3}$ 的凹性得出，即：$(x+y)^{1/3} - x^{1/3} \leq y/(3x^{2/3})$；（$c$）利用了事实 $w_i \geq 2$；最后，（d）和（e）分别利用了 $\eta_{\tau,i}$ 的定义以及条件 $\eta_{\tau,i} \leq \frac{\rho}{48\lambda LI^2}$。因此，若我们选择 $c = \frac{96\lambda^2L^2}{K\rho^2} + \frac{\rho}{72\kappa^3\lambda LI^2}$，则有：$\eta_{\tau,i}^{-1} - \eta_{\tau,i-1}^{-1} - c\eta_{\tau,i} \leq -\frac{96\lambda^2L^2}{K\rho^2}\eta_{\tau,i}$，

因此，我们有：

$$\begin{aligned}
\frac{\mathbb{E}\|\bar{e}_{\tau,i+1}\|^2}{\eta_{\tau,i}} - \frac{\mathbb{E}\|\bar{e}_{\tau,i}\|^2}{\eta_{\tau,i-1}} &\leq -\frac{96\lambda^2L^2\eta_{\tau,i}}{K\rho^2}\mathbb{E}\|\bar{e}_{\tau,i}\|^2 + \frac{80\lambda^2(I-1)L^2}{\rho^2K^2}\sum_{\ell=1}^{i}\eta_{\tau,\ell}\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau,\ell}^{(k)} - \bar{\nu}_{\tau,\ell}\|^2 \\
&\qquad + \frac{8L^2\eta_{\tau,i}}{K}\mathbb{E}\|\tilde{d}_{\tau,i}\|^2 + \frac{4\sigma^2c^2\eta_{\tau,i}^3}{b_1K},
\end{aligned}$$

两边同时乘以 $\rho K/64\lambda L^2$，得到:

$$\frac{\rho K}{64\lambda L^2}\left(\frac{\mathbb{E}\|\bar{e}_{\tau,i+1}\|^2}{\eta_{\tau,i}}-\frac{\mathbb{E}\|\bar{e}_{\tau,i}\|^2}{\eta_{\tau,i-1}}\right)\leq-\frac{3\lambda\eta_{\tau,i}}{2\rho}\mathbb{E}\|\bar{e}_{\tau,i}\|^2+\frac{5\lambda(I-1)}{4K\rho}\sum_{\ell=1}^{i}\eta_{\tau,\ell}\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau,\ell}^{(k)}-\bar{\nu}_{\tau,\ell}\|^2 + \frac{\eta_{\tau,i}\rho}{8\lambda}\mathbb{E}\|\tilde{d}_{\tau,i}\|^2+\frac{\sigma^2c^2\eta_{\tau,i}^3\rho}{16\lambda L^2b_1}.$$

然后将上述不等式从 0 累加到 $I-1$，得到:

$$\begin{aligned}\frac{\rho K}{64\lambda L^2}\left(\frac{\mathbb{E}\|\bar{e}_{\tau,I}\|^2}{\eta_{\tau,I-1}}-\frac{\mathbb{E}\|\bar{e}_{\tau,0}\|^2}{\eta_{\tau-1,I-1}}\right)&\leq-\sum_{i=0}^{I-1}\frac{3\lambda\eta_i}{2\rho}\mathbb{E}\|\bar{e}_{\tau,i}\|^2+\sum_{i=0}^{I-1}\frac{5\lambda(I-1)}{4K\rho}\sum_{\ell=1}^{i}\eta_\ell\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau,\ell}^{(k)}-\bar{\nu}_{\tau,\ell}\|^2\\&\quad+\sum_{i=0}^{I-1}\frac{\eta_{\tau,i}\rho}{8\lambda}\mathbb{E}\|\tilde{d}_{\tau,i}\|^2+\sum_{i=0}^{I-1}\frac{\sigma^2c^2\eta_{\tau,i}^3\rho}{16\lambda L^2b_1}\\&\leq-\sum_{i=0}^{I-1}\frac{3\lambda\eta_i}{2\rho}\mathbb{E}\|\bar{e}_{\tau,i}\|^2+\frac{5\lambda I(I-1)}{4K\rho}\sum_{\ell=1}^{I}\eta_\ell\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau,\ell}^{(k)}-\bar{\nu}_{\tau,\ell}\|^2\\&\quad+\sum_{i=0}^{I-1}\frac{\eta_{\tau,i}\rho}{8\lambda}\mathbb{E}\|\tilde{d}_{\tau,i}\|^2+\sum_{i=0}^{I-1}\frac{\sigma^2c^2\eta_{\tau,i}^3\rho}{16\lambda L^2b_1}\end{aligned}$$

根据定义，我们有 $\bar{e}_{\tau,\ 0}=e_{\tau-1}$ 和 $\bar{e}_{\tau,\ I}=e_\tau$，然后将 τ 替换为 $\tau+1$，即可得到引理中的结果。

□

B.6 势函数下降

我们将势函数定义如下:

$$\Phi_\tau:=f(\tilde{x}_\tau)+\frac{\rho K}{64\lambda L^2}\frac{\|e_\tau\|^2}{\eta_{\tau-1,I-1}}. \tag{19}$$

接下来，我们刻画势函数的下降情况。

引理 8。 对于任意 $\tau\geq 0$，我们有:

$$\begin{aligned}\mathbb{E}[\Phi_{\tau+1}-\Phi_\tau]\leq-\sum_{i=0}^{I-1}\left(\frac{5\rho\eta_{\tau+1,i}}{8\lambda}-\frac{\eta_{\tau+1,i}^2L}{2}\right)\mathbb{E}\|\tilde{d}_i\|^2-\frac{\lambda}{2\rho}\sum_{i=0}^{I-1}\eta_{\tau+1,i}\mathbb{E}\|\bar{e}_{\tau+1,i}\|^2\\+\frac{\sigma^2c^2\rho}{16\lambda L^2b_1}\sum_{i=0}^{I-1}\eta_{\tau+1,i}^3+\frac{5\lambda I(I-1)}{4K\rho}\sum_{i=1}^{I}\eta_{\tau+1,i}\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau+1,i}^{(k)}-\bar{\nu}_{\tau+1,i}\|^2,\end{aligned}$$

其中，期望是相对于算法的随机性而言的。

证明。 我们可以通过结合引理 4 和引理 7 得到引理中的不等式

□

B.7 累积梯度误差

在这个 小节中，我们对由以下项给出的梯度共识误差进行界定： $\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau}^{(k)},_{i}-\bar{\nu}_{\tau,i}\|\quad{}^{2}$.

引理 9. 对于 $i\geq 1$ 和 $\alpha_i<1$，有：

$$\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau,i}^{(k)}-\bar{\nu}_{\tau,i}\|^2\leq(1+\frac{1}{I})\sum_{k=1}^{K}\mathbb{E}\big\|\nu_{\tau,i-1}^{(k)}-\bar{\nu}_{\tau,i-1}\big\|^2+8KIL^2\eta_{\tau,i-1}^2\mathbb{E}\|\tilde{d}_{\tau,i-1}\|^2+\frac{8KI\sigma^2c^2\eta_{\tau,i-1}^4}{b_1}$$
$$+16KI\zeta^2c^2\eta_{\tau,i-1}^4+\frac{96\lambda^2I^2L^2}{\rho^2}\sum_{\ell=1}^{i-1}\eta_{\tau,\ell^2}\sum_{k=1}^{K}\mathbb{E}\|\nu_{\tau,\ell}^{(k)}-\bar{\nu}_{\tau,\ell}\|^2$$

其中，期望是关于算法随机性的。

证明。 根据 $\nu_i^{(k)}$ 的更新规则（为方便起见，省略全局轮次步骤），有：

$$\begin{aligned}
&\mathbb{E}\|\nu_i^{(k)}-\bar{\nu}_i\|^2\\
&=\mathbb{E}\Bigg\|\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)})+(1-\alpha_i)\big(\nu_{i-1}^{(k)}-\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})\big)\\
&\qquad-\bigg(\frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_i^{(j)};\mathcal{B}_i^{(j)})+(1-\alpha_i)\big(\bar{\nu}_{i-1}-\frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)};\mathcal{B}_i^{(j)})\big)\bigg)\Bigg\|^2\\
&=\mathbb{E}\Bigg\|(1-\alpha_i)\big(\nu_{i-1}^{(k)}-\bar{\nu}_{i-1}\big)+\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_i^{(j)};\mathcal{B}_i^{(j)})\\
&\qquad-(1-\alpha_i)\bigg(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)};\mathcal{B}_i^{(j)})\bigg)\Bigg\|^2\\
&\leq(1+\beta)(1-\alpha_i)^2\mathbb{E}\Big\|\nu_{i-1}^{(k)}-\bar{\nu}_{i-1}\Big\|^2+\Big(1+\frac{1}{\beta}\Big)\mathbb{E}\Bigg\|\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_i^{(j)};\mathcal{B}_i^{(j)})\\
&\qquad-(1-\alpha_i)\bigg(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)};\mathcal{B}_i^{(j)})\bigg)\Bigg\|^2,
\end{aligned}\tag{20}$$

其中，最后一个不等式使用了命题 1。

接下来，我们考虑第二项：

$$
\begin{aligned}
&\mathbb{E}\Bigg\|\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)}) - \frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_i^{(j)};\mathcal{B}_i^{(j)}) \\
&\qquad\qquad - (1-\alpha_i)\left(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)}) - \frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)};\mathcal{B}_i^{(j)})\right)\Bigg\|^2 \\
&\overset{(a)}{\leq} 2\mathbb{E}\Bigg\|\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)}) - \frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_i^{(j)};\mathcal{B}_i^{(j)}) \\
&\qquad\qquad - \left(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)}) - \frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)};\mathcal{B}_i^{(j)})\right)\Bigg\|^2 \\
&\qquad\qquad + 2\alpha_i^2\mathbb{E}\Bigg\|\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)}) - \frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)};\mathcal{B}_i^{(j)})\Bigg\|^2 \\
&\overset{(b)}{\leq} 2\mathbb{E}\Bigg\|\left(\nabla f^{(k)}(x_i^{(k)};\mathcal{B}_i^{(k)}) - \nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})\right)\Bigg\|^2 \\
&\qquad\qquad + 2\alpha_i^2\mathbb{E}\Bigg\|\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)}) - \frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)};\mathcal{B}_i^{(j)})\Bigg\|^2 \\
&\overset{(c)}{\leq} 2L^2\mathbb{E}\left\|x_i^{(k)} - x_{i-1}^{(k)}\right\|^2 + 2\alpha_i^2\mathbb{E}\Bigg\|\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)}) - \frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)};\mathcal{B}_i^{(j)})\Bigg\|^2, \qquad (21)
\end{aligned}
$$

其中，不等式 (a) 使用了命题 1；不等式 (b) 使用了命题 2；不等式 (c) 使用了光滑性假设。

接下来，我们考虑上面式（21）中的第二项，有

$$
\begin{aligned}
&\mathbb{E}\left\|\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)};\mathcal{B}_i^{(j)})\right\|^2\\
&=\mathbb{E}\Bigg\|\left(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_{i-1}^{(k)})\right)\\
&\qquad-\frac{1}{K}\sum_{j=1}^{K}\left(\nabla f^{(j)}(x_{i-1}^{(j)};\mathcal{B}_i^{(j)})-\nabla f^{(j)}(x_{i-1}^{(j)})\right)+\nabla f^{(k)}(x_{i-1}^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)})\Bigg\|^2\\
&\le 2\mathbb{E}\Bigg\|\left(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_{i-1}^{(k)})\right)\\
&\qquad\qquad-\frac{1}{K}\sum_{j=1}^{K}\left(\nabla f^{(j)}(x_{i-1}^{(j)};\mathcal{B}_i^{(j)})-\nabla f^{(j)}(x_{i-1}^{(j)})\right)\Bigg\|^2\\
&\qquad\qquad\qquad+2\mathbb{E}\left\|\nabla f^{(k)}(x_{i-1}^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)})\right\|^2\\
&\overset{(a)}{\le} 2\mathbb{E}\left\|\left(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_{i-1}^{(k)})\right)\right\|^2+2\mathbb{E}\left\|\nabla f^{(k)}(x_{i-1}^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)})\right\|^2\\
&\le 2\mathbb{E}\left\|\left(\nabla f^{(k)}(x_{i-1}^{(k)};\mathcal{B}_i^{(k)})-\nabla f^{(k)}(x_{i-1}^{(k)})\right)\right\|^2+4\mathbb{E}\left\|\nabla f^{(k)}(\tilde{x}_{i-1})-\nabla f(\tilde{x}_{i-1})\right\|^2\\
&\qquad+8\mathbb{E}\left\|\nabla f^{(k)}(x_{i-1}^{(k)})-\nabla f^{(k)}(\tilde{x}_{i-1})\right\|^2+8\mathbb{E}\left\|\nabla f(\tilde{x}_{i-1})-\frac{1}{K}\sum_{j=1}^{K}\nabla f^{(j)}(x_{i-1}^{(j)})\right\|^2\\
&\overset{(b)}{\le}\frac{2\sigma^2}{b_1}+\frac{4}{K}\sum_{j=1}^{K}\mathbb{E}\|\nabla f^{(k)}(\tilde{x}_{i-1})-\nabla f^{(j)}(\bar{x}_{i-1})\|^2\\
&\qquad+8L^2\mathbb{E}\|x_{i-1}^{(k)}-\tilde{x}_{i-1}\|^2+\frac{8L^2}{K}\sum_{j=1}^{K}\mathbb{E}\|x_{i-1}^{(j)}-\tilde{x}_{i-1}\|^2\\
&\overset{(c)}{\le}\frac{2\sigma^2}{b_1}+4\zeta^2+8L^2\mathbb{E}\|x_{i-1}^{(k)}-\tilde{x}_{i-1}\|^2+\frac{8L^2}{K}\sum_{j=1}^{K}\mathbb{E}\|x_{i-1}^{(j)}-\tilde{x}_{i-1}\|^2,
\end{aligned}
\tag{22}
$$

其中，不等式（a）使用了命题 2；不等式（b）利用了有界方差假设；（c）使用了有界异质性假设。最后，将（22）和（21）代入（20），并求和

所有 K 个工作节点后，可得

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\|\nu_i^{(k)}-\bar{\nu}_i\|^2\\
&\leq(1-\alpha_i)^2(1+\beta)\sum_{k=1}^{K}\mathbb{E}\left\|\nu_{i-1}^{(k)}-\bar{\nu}_{i-1}\right\|^2+2L^2\left(1+\frac{1}{\beta}\right)\sum_{k=1}^{K}\mathbb{E}\|x_i^{(k)}-x_{i-1}^{(k)}\|^2\\
&\qquad+\frac{4K\sigma^2}{b_1}\left(1+\frac{1}{\beta}\right)\alpha_i^2+8K\zeta^2\left(1+\frac{1}{\beta}\right)\alpha_i^2+32L^2\left(1+\frac{1}{\beta}\right)\alpha_i^2\sum_{k=1}^{K}\mathbb{E}\|x_{i-1}^{(k)}-\tilde{x}_{i-1}\|^2\\
&\leq(1-\alpha_i)^2(1+\beta)\sum_{k=1}^{K}\mathbb{E}\left\|\nu_{i-1}^{(k)}-\bar{\nu}_{i-1}\right\|^2+2L^2\eta_{i-1}^2\left(1+\frac{1}{\beta}\right)\sum_{k=1}^{K}\mathbb{E}\|d_{i-1}^{(k)}\|^2\\
&\qquad+\frac{4K\sigma^2}{b_1}\left(1+\frac{1}{\beta}\right)\alpha_i^2+8K\zeta^2\left(1+\frac{1}{\beta}\right)\alpha_i^2+\frac{32\lambda^2L^2a_i^2}{\rho^2}\left(1+\frac{1}{\beta}\right)\sum_{k=1}^{K}\mathbb{E}\|z_{i-1}^{(k)}-\bar{z}_{i-1}\|^2
\end{aligned}
$$

其中，第二个不等式使用了引理 1 的断言 3。

接下来利用引理 2，可得:

$$
\begin{aligned}
\sum_{k=1}^{K}\mathbb{E}\|\nu_i^{(k)}-\bar{\nu}_i\|^2\leq(1-\alpha_i)^2(1+\beta)\sum_{k=1}^{K}\mathbb{E}\left\|\nu_{i-1}^{(k)}-\bar{\nu}_{i-1}\right\|^2+2L^2\eta_{i-1}^2\left(1+\frac{1}{\beta}\right)\sum_{k=1}^{K}\mathbb{E}\|d_{i-1}^{(k)}\|^2\\
+\frac{4K\sigma^2}{b_1}\left(1+\frac{1}{\beta}\right)\alpha_i^2+8K\zeta^2\left(1+\frac{1}{\beta}\right)\alpha_i^2\\
+\frac{32\lambda^2L^2a_i^2}{\rho^2}\left(1+\frac{1}{\beta}\right)(I-1)\sum_{\ell=1}^{i-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)}-\bar{\nu}_\ell\|^2 \qquad (23)
\end{aligned}
$$

对于上述不等式中的第二项，我们有:

$$
\begin{aligned}
&2L^2\eta_{i-1}^2\left(1+\frac{1}{\beta}\right)\sum_{k=1}^{K}\mathbb{E}\|d_{i-1}^{(k)}\|^2\\
&\leq4L^2\eta_{i-1}^2\left(1+\frac{1}{\beta}\right)\sum_{k=1}^{K}\mathbb{E}\|d_{i-1}^{(k)}-\tilde{d}_{i-1}\|^2+4KL^2\eta_{i-1}^2\left(1+\frac{1}{\beta}\right)\mathbb{E}\|\tilde{d}_{i-1}\|^2\\
&\leq\frac{16\lambda^2L^2(I-1)}{\rho^2}\left(1+\frac{1}{\beta}\right)\sum_{\ell=1}^{i-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)}-\bar{\nu}_\ell\|^2+4KL^2\eta_{i-1}^2\left(1+\frac{1}{\beta}\right)\mathbb{E}\|\tilde{d}_{i-1}\|^2
\end{aligned}
$$

其中，　第一个不等式使用命题 1，第二个不等式使用引理 3。接下来，将　ug

上述不等式代入式 (23)，可得：

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}\|\nu_i^{(k)}-\bar{\nu}_i\|^2 &\le (1-\alpha_i)^2(1+\beta)\sum_{k=1}^{K}\mathbb{E}\left\|\nu_{i-1}^{(k)}-\bar{\nu}_{i-1}\right\|^2+4KL^2\eta_{i-1}^2\left(1+\frac{1}{\beta}\right)\mathbb{E}\|\tilde{d}_{i-1}\|^2\\ &\quad+\frac{4K\sigma^2}{b_1}\left(1+\frac{1}{\beta}\right)\alpha_i^2+8K\zeta^2\left(1+\frac{1}{\beta}\right)\alpha_i^2\\ &\quad+\frac{16\lambda^2L^2(1+2a_i^2)(I-1)}{\rho^2}\left(1+\frac{1}{\beta}\right)\sum_{\ell=1}^{i-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)}-\bar{\nu}_\ell\|^2\\ &\le (1+\frac{1}{I})\sum_{k=1}^{K}\mathbb{E}\left\|\nu_{i-1}^{(k)}-\bar{\nu}_{i-1}\right\|^2+8KIL^2\eta_{i-1}^2\mathbb{E}\|\tilde{d}_{i-1}\|^2+\frac{8KI\sigma^2c^2\eta_{i-1}^4}{b_1}\\ &\quad+16KI\zeta^2c^2\eta_{i-1}^4+\frac{96\lambda^2I^2L^2}{\rho^2}\sum_{\ell=1}^{i-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)}-\bar{\nu}_\ell\|^2,\end{aligned}$$

在最后一个不等式中，我们取 $\beta=1/I$，于是有 $(1+1/\beta)\le(1+I)\le 2I$。我们还用到了以下事实：$(1-\alpha_i)^2<1$ 和 $a_i=c\eta_{i-1}^2<1$。证明完毕。 □

引理 10. 对于 $\eta_i \le \frac{\rho}{48LI^2}$，于是有

$$\frac{I^2}{\rho K}\sum_{i=1}^{I}\eta_i\sum_{k=1}^{K}\mathbb{E}\|\nu_i^{(k)}-\bar{\nu}_i\|^2 \le \frac{\rho}{84}\sum_{i=0}^{I-1}\eta_i\mathbb{E}\|\tilde{d}_i\|^2+\left(\frac{\rho\sigma^2c^2}{84b_1L^2}+\frac{\rho\zeta^2c^2}{42L^2}\right)\sum_{i=0}^{I-1}\eta_i^3$$

证明。 由引理 9（为方便起见，我们省略全局轮次编号）可得：

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}\|\nu_i^{(k)}-\bar{\nu}_i\|^2 &\le (1+\frac{1}{I})\sum_{k=1}^{K}\mathbb{E}\left\|\nu_{i-1}^{(k)}-\bar{\nu}_{i-1}\right\|^2+8KIL^2\eta_{i-1}^2\mathbb{E}\|\tilde{d}_{i-1}\|^2+\frac{8KI\sigma^2c^2\eta_{i-1}^4}{b_1}\\ &\quad+16KI\zeta^2c^2\eta_{i-1}^4+\frac{96\lambda^2I^2L^2}{\rho^2}\sum_{\ell=1}^{i-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)}-\bar{\nu}_\ell\|^2\\ &\le (1+\frac{1}{I})\sum_{k=1}^{K}\mathbb{E}\left\|\nu_{i-1}^{(k)}-\bar{\nu}_{i-1}\right\|^2+\frac{KL\rho\eta_{i-1}}{6\lambda I}\mathbb{E}\|\tilde{d}_{i-1}\|^2+\frac{K\rho\sigma^2c^2\eta_{i-1}^3}{6\lambda ILb_1}\\ &\quad+\frac{K\rho\zeta^2c^2\eta_{i-1}^3}{3\lambda IL}+\frac{96\lambda^2I^2L^2}{\rho^2}\sum_{\ell=1}^{i-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)}-\bar{\nu}_\ell\|^2,\end{aligned}\tag{24}$$

其中在第二个不等式中，我们使用了条件 $\eta_i \le \frac{\rho}{48\lambda LI^2}$。将 (24) 从 1 递归应用至 i，可得：

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}\|\nu_i^{(k)}-\bar{\nu}_i\|^2 &\le \frac{KL\rho}{6\lambda I}\sum_{\ell=0}^{i-1}\left(1+\frac{1}{I}\right)^{i-1-\ell}\eta_\ell\mathbb{E}\|\tilde{d}_\ell\|^2+\frac{K\rho\sigma^2c^2}{6\lambda ILb_1}\sum_{\ell=0}^{i-1}\left(1+\frac{1}{I}\right)^{i-1-\ell}\eta_\ell^3\\ &\quad+\frac{K\rho\zeta^2c^2}{3\lambda IL}\sum_{\ell=0}^{i-1}\left(1+\frac{1}{I}\right)^{i-1-\ell}\eta_\ell^3\\ &\quad+\frac{96\lambda^2L^2I^2}{\rho^2}\sum_{\ell=0}^{i-1}\left(1+\frac{1}{I}\right)^{i-1-\ell}\sum_{\bar{\ell}=0}^{\ell}\eta_{\bar{\ell}}^2\sum_{k=1}^{K}\mathbb{E}\|\nu_{\bar{\ell}}^{(k)}-\bar{\nu}_{\bar{\ell}}\|^2\\ &\overset{(a)}{\le} \frac{KL\rho}{6\lambda I}\left(1+\frac{1}{I}\right)^{I}\sum_{\ell=0}^{i-1}\eta_\ell\mathbb{E}\|\tilde{d}_\ell\|^2+\frac{K\rho\sigma^2c^2}{6\lambda ILb_1}\left(1+\frac{1}{I}\right)^{I}\sum_{\ell=0}^{i-1}\eta_\ell^3\\ &\quad+\frac{K\rho\zeta^2c^2}{3\lambda IL}\left(1+\frac{1}{I}\right)^{I}\sum_{\ell=0}^{i-1}\eta_\ell^3+\frac{96\lambda^2L^2I^3}{\rho^2}\left(1+\frac{1}{I}\right)^{I}\sum_{\bar{\ell}=0}^{i-1}\eta_{\bar{\ell}}^2\sum_{k=1}^{K}\mathbb{E}\|\nu_{\bar{\ell}}^{(k)}-\bar{\nu}_{\bar{\ell}}\|^2\\ &\overset{(b)}{\le} \frac{KL\rho}{2\lambda I}\sum_{\ell=0}^{i-1}\eta_\ell\mathbb{E}\|\tilde{d}_\ell\|^2+\frac{K\rho\sigma^2c^2}{2\lambda ILb_1}\sum_{\ell=0}^{i-1}\eta_\ell^3+\frac{K\rho\zeta^2c^2}{\lambda IL}\sum_{\ell=0}^{i-1}\eta_\ell^3\\ &\quad+\frac{288\lambda^2L^2I^3}{\rho^2}\sum_{\ell=0}^{i-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)}-\bar{\nu}_\ell\|^2,\end{aligned}\tag{25}$$

其中，不等式（a）由以下事实得出：$1+1/I>1$，且对于 $i\in[I]$，$i-1-\ell\le I$；同时，不等式（b）是因为 $(1+1/I)^I \le \mathrm{e} < 3$。

接下来，将（25）两边乘以 η_i，并对 $i=1$ 到 I 求和：

$$\begin{aligned}
\sum_{i=1}^{I}\eta_i\sum_{k=1}^{K}\mathbb{E}\|\nu_i^{(k)}-\bar{\nu}_i\|^2 &\le \frac{KL\rho}{2\lambda I}\sum_{i=1}^{I}\eta_i\sum_{\ell=0}^{i-1}\eta_\ell\mathbb{E}\|\tilde{d}_\ell\|^2+\frac{K\rho\sigma^2c^2}{2\lambda ILb_1}\sum_{i=1}^{I}\eta_i\sum_{\ell=0}^{i-1}\eta_\ell^3\\
&\quad+\frac{K\rho\zeta^2c^2}{\lambda IL}\sum_{i=1}^{I}\eta_i\sum_{\ell=0}^{i-1}\eta_\ell^3+\frac{288\lambda^2L^2I^3}{\rho^2}\sum_{i=1}^{I}\eta_i\sum_{\ell=0}^{i-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)}-\bar{\nu}_\ell\|^2\\
&\overset{(a)}{\le}\frac{KL\rho}{2\lambda I}\Big(\sum_{i=1}^{I}\eta_i\Big)\sum_{\ell=0}^{I-1}\eta_\ell\mathbb{E}\|\tilde{d}_\ell\|^2+\Big(\frac{K\rho\sigma^2c^2}{2\lambda ILb_1}+\frac{K\rho\zeta^2c^2}{\lambda IL}\Big)\Big(\sum_{i=1}^{I}\eta_i\Big)\sum_{\ell=0}^{I-1}\eta_\ell^3\\
&\quad+\frac{288\lambda^2L^2I^3}{\rho^2}\Big(\sum_{i=1}^{I}\eta_i\Big)\sum_{\ell=0}^{I-1}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)}-\bar{\nu}_\ell\|^2\\
&\overset{(b)}{\le}\frac{K\rho^2}{96\lambda^2I^2}\sum_{i=0}^{I-1}\eta_i\mathbb{E}\|\tilde{d}_i\|^2+\Big(\frac{K\rho^2\sigma^2c^2}{96\lambda^2I^2L^2b_1}+\frac{K\rho^2\zeta^2c^2}{48\lambda^2I^2L^2}\Big)\sum_{i=0}^{I-1}\eta_i^3+\frac{1}{8}\sum_{\ell=1}^{I-1}\eta_\ell\sum_{k=1}^{K}\mathbb{E}\|\nu_\ell^{(k)}-\bar{\nu}_\ell\|^2
\end{aligned}$$

其中，不等式 (a) 利用了 $i\le I$ 这一事实，而 (b) 则利用了我们选择 $\eta_i\le\rho/(48\lambda LI^2)$ 这一点。整理各项后可得：

$$\frac{7}{8}\sum_{i=1}^{I}\eta_i\sum_{k=1}^{K}\mathbb{E}\|\nu_i^{(k)}-\bar{\nu}_i\|^2\le\frac{K\rho^2}{96\lambda^2I^2}\sum_{i=0}^{I-1}\eta_i\mathbb{E}\|\tilde{d}_i\|^2+\Big(\frac{K\rho^2\sigma^2c^2}{96\lambda^2I^2L^2b_1}+\frac{K\rho^2\zeta^2c^2}{48\lambda^2I^2L^2}\Big)\sum_{i=0}^{I-1}\eta_i^3$$

两边乘以 $8\lambda I^2/(7K\rho)$，可得：

$$\frac{\lambda I^2}{K\rho}\sum_{i=1}^{I}\eta_i\sum_{k=1}^{K}\mathbb{E}\|\nu_i^{(k)}-\bar{\nu}_i\|^2\le\frac{\rho}{84\lambda}\sum_{i=0}^{I-1}\eta_i\mathbb{E}\|\tilde{d}_i\|^2+\Big(\frac{\rho\sigma^2c^2}{84\lambda L^2b_1}+\frac{\rho\zeta^2c^2}{42\lambda L^2}\Big)\sum_{i=0}^{I-1}\eta_i^3$$

证明至此完成。 □

B.8 主要收敛定理的证明

本小节中，我们证明定理 1 和推论 5.7。为证明定理 1，我们首先证明以下定理成立：

定理 2. 选择参数为 $\kappa=\dfrac{\rho K^{2/3}}{\lambda L}$，$c=\dfrac{96\lambda^2L^2}{K\rho^2}+\dfrac{\rho}{72\kappa^3\lambda LI^2}$，$w_t=\max\Big\{48^3I^6K^2-t-I,14^3K^{0.5}\Big\}$，$\lambda>0$，并选择 $\eta_t=\dfrac{\kappa}{(\omega_t+t+I)^{1/3}}$，则有：

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}\|\tilde{d}_t\|^2+\frac{\lambda^2}{\rho^2}\mathbb{E}\|\bar{e}_t\|^2\Big)\\
&\le\Big[\frac{96\lambda^2LI^2}{\rho^2T}+\frac{2\lambda^2L}{\rho^2K^{2/3}T^{2/3}}\Big](f(x_0)-f^*)+\Big[\frac{72\lambda^2I^4}{b\rho^2T}+\frac{3\lambda^2I^2}{2b\rho^2K^{2/3}T^{2/3}}\Big]\sigma^2\\
&\quad+\frac{192^2\lambda^2}{\rho^2}\times\Big(\frac{48I^2}{T}+\frac{1}{K^{2/3}T^{2/3}}\Big)\times\Big(\frac{\sigma^2}{4b_1}+\frac{2\zeta^2}{21}\Big)\log(T+1).
\end{aligned}$$

证明. 根据定义，有 $\eta_t \le \eta_0 < \kappa/w_0^{1/3} = \rho K^{2/3}/48\lambda L I^2 K^{2/3} = \rho/48\lambda L I^2$，于是 $c = \lambda^2 L^2\left(\frac{96}{K\rho^2} + \frac{1}{72K^2\rho^2 I^2}\right) \le \frac{192\lambda^2 L^2}{K\rho^2}$ 且：

$$c\eta_t^2 \le c\eta_0^2 < \frac{192\lambda^2 L^2}{K\rho^2} * \frac{\kappa^2}{w_0^{2/3}} = \frac{192L^2}{K\rho^2} * \frac{\rho^2 K^{4/3}}{L^2 w_0^{2/3}} = \frac{192K^{1/3}}{w_0^{2/3}} \le \frac{192K^{1/3}}{196K^{1/3}} < 1,$$

因此我们有 $\alpha_t < 1$，于是引理 8-引理 10 的条件均得到满足。

首先，将引理 10 中的梯度共识误差代入引理 8，可得势函数的下降量为：

$$\begin{aligned}
\mathbb{E}[\Phi_{\tau+1} - \Phi_\tau] \le & -\sum_{i=0}^{I-1}\left(\frac{5\rho\eta_{\tau+1,i}}{8\lambda} - \frac{\eta_{\tau+1,i}^2 L}{2}\right)\mathbb{E}\|\tilde{d}_{\tau+1,i}\|^2 - \frac{\lambda}{2\rho}\sum_{i=0}^{I-1}\eta_{\tau+1,i}\mathbb{E}\|\bar{e}_{\tau+1,i}\|^2 \\
& + \frac{\sigma^2 c^2 \rho}{16\lambda L^2 b_1}\sum_{i=0}^{I-1}\eta_{\tau+1,i}^3 + \frac{\rho}{42\lambda}\sum_{i=0}^{I-1}\eta_{\tau+1,i}\mathbb{E}\|\tilde{d}_{\tau+1,i}\|^2 + \left(\frac{\rho\sigma^2 c^2}{42\lambda L^2 b_1} + \frac{\rho\zeta^2 c^2}{21\lambda L^2}\right)\sum_{i=0}^{I-1}\eta_{\tau+1,i}^3 \\
\le & -\sum_{i=0}^{I-1}\left(\frac{3\rho\eta_{\tau+1,i}}{5\lambda} - \frac{\eta_{\tau+1,i}^2 L}{2}\right)\mathbb{E}\|\tilde{d}_{\tau+1,i}\|^2 - \frac{\lambda}{2\rho}\sum_{i=0}^{I-1}\eta_{\tau+1,i}\mathbb{E}\|\bar{e}_{\tau+1,i}\|^2 \\
& + \left(\frac{\rho\sigma^2 c^2}{8\lambda L^2 b_1} + \frac{\rho\zeta^2 c^2}{21\lambda L^2}\right)\sum_{i=0}^{I-1}\eta_{\tau+1,i}^3 \\
\overset{(a)}{\le} & -\sum_{i=0}^{I-1}\frac{\rho\eta_i}{2\lambda}\mathbb{E}\|\tilde{d}_{\tau+1,i}\|^2 - \frac{\lambda}{2\rho}\sum_{i=0}^{I-1}\eta_{\tau+1,i}\mathbb{E}\|\bar{e}_{\tau+1,i}\|^2 + \left(\frac{\rho\sigma^2 c^2}{8\lambda L^2 b_1} + \frac{\rho\zeta^2 c^2}{21\lambda L^2}\right)\sum_{i=0}^{I-1}\eta_{\tau+1,i}^3,
\end{aligned}$$

其中，(a) 根据 $\eta_i \le \frac{\rho}{48\lambda L I^2} \le \frac{\rho}{48\lambda L}$ 这一事实得出。 ——

假设我们将 $T = EI$ 记为 $t \ge 0$ 的 $t = \tau I + i$，将 $\tau \ge 0$ 记为 $\eta_t = \eta_{\tau+1,\,i}$，$\tilde{d}_t = d_{\tau+1,\,i}$，$\bar{e}_t = \bar{e}_{\tau+1,\,i}$。特别地，为方便起见，我们将 $\eta_{-1} = \eta_0$ 记为。

然后，我们将上述不等式对 τ 从 0 到 $E-1$ 求和，得到：

$$\mathbb{E}[\Phi_E - \Phi_0] \le -\sum_{t=0}^{T-1}\left(\frac{\rho\eta_t}{2\lambda}\right)\mathbb{E}\|\tilde{d}_t\|^2 - \sum_{t=0}^{T-1}\frac{\lambda\eta_t}{2\rho}\mathbb{E}\|\bar{e}_t\|^2 + \left(\frac{\rho\sigma^2 c^2}{8\lambda L^2 b_1} + \frac{\rho\zeta^2 c^2}{21\lambda L^2}\right)\sum_{t=0}^{T}\eta_t^3,$$

整理各项后，得到：

$$\begin{aligned}
\sum_{t=1}^{T}\left(\frac{\rho\eta_t}{2\lambda}\mathbb{E}\|\tilde{d}_t\|^2 + \frac{\lambda\eta_t}{2\rho}\mathbb{E}\|\bar{e}_t\|^2\right) &\le \mathbb{E}[\Phi_0 - \Phi_E] + \left(\frac{\rho\sigma^2 c^2}{8\lambda L^2 b_1} + \frac{\rho\zeta^2 c^2}{21\lambda L^2}\right)\sum_{t=0}^{T-1}\eta_t^3 \\
&\overset{(a)}{\le} f(x_0) - f^* + \frac{\rho K}{64\lambda L^2}\frac{\mathbb{E}\|e_0\|^2}{\eta_0} + \left(\frac{\rho\sigma^2 c^2}{8\lambda L^2 b_1} + \frac{\rho\zeta^2 c^2}{21\lambda L^2}\right)\sum_{t=0}^{T-1}\eta_t^3 \\
&\overset{(b)}{\le} f(x_0) - f^* + \frac{\sigma^2\rho}{64\lambda b L^2\eta_0} + \left(\frac{\rho\sigma^2 c^2}{8\lambda L^2 b_1} + \frac{\rho\zeta^2 c^2}{21\lambda L^2}\right)\sum_{t=0}^{T-1}\eta_t^3, \quad (26)
\end{aligned}$$

其中，(a) 源于以下事实：$f^* \le \Phi_E$；而 (b) 则是应用引理 5 得出的结果，b 是第一次迭代时的小批量大小。

接下来，对于上面式 (26) 中的最后一项，我们有：

$$\sum_{t=0}^{T-1}\eta_t^3=\sum_{t=0}^{T-1}\frac{\kappa^3}{w_t+t}\overset{(a)}{\leq}\sum_{t=0}^{T-1}\frac{\kappa^3}{1+t}=\kappa^3\sum_{t=0}^{T-1}\frac{1}{1+t}\overset{(b)}{\leq}\kappa^3\ln(T+1). \tag{27}$$

其中，上述不等式（a）由我们有 $w_t>1$ 这一事实得出，而不等式（b）则由应用命题3得出。

将（27）代入（26），两边乘以 $2\lambda/(\rho\eta_T T)$，并利用 η_t 关于 t 单调不增这一事实，可得

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\|\tilde{d}_t\|^2+\frac{\lambda^2}{\rho^2}\mathbb{E}\|\bar{e}_t\|^2\right)\leq\frac{2\lambda(f(x_0)-f^*)}{\rho\eta_T T}+\frac{1}{\eta_T T}\frac{\sigma^2}{32bL^2\eta_0}+\frac{\kappa^3}{\eta_T T}\left(\frac{\sigma^2c^2}{4b_1L^2}+\frac{2\zeta^2c^2}{21L^2}\right)\ln(T+1). \tag{28}$$

现在分别考察上述（28）中的各项。对于第一项：

$$\frac{1}{\eta_T T}=\frac{(w_T+T)^{1/3}}{\kappa T}\overset{(a)}{\leq}\frac{w_T^{1/3}}{\kappa T}+\frac{1}{\kappa T^{2/3}}=\frac{48\lambda LI^2}{\rho T}+\frac{\lambda L}{\rho K^{2/3}T^{2/3}}. \tag{29}$$

其中，不等式（a）由恒等式（$x+y$）$^{1/3}\leq x^{1/3}+y^{1/3}$ 得出，而不等式（b）则由 κ 的定义以及 w_T 得出。

$$w_T=\max\left\{(I+1),48^3I^6K^2-T,2*320^{1.5}K^{0.5}\right\}\leq 48^3I^6K^2,$$

类似地，对于（28）的第二项，根据 η_0 和 η_T 的定义，可得

$$\begin{aligned}\frac{1}{\eta_T T}\frac{\sigma^2}{32bL^2\eta_0}&\leq\left(\frac{48\lambda LI^2}{\rho T}+\frac{\lambda L}{\rho K^{2/3}T^{2/3}}\right)\times\frac{\sigma^2}{32bL^2}\times\frac{w_0^{1/3}}{\kappa}\\&\leq\left(\frac{48\lambda LI^2}{\rho T}+\frac{\lambda L}{\rho K^{2/3}T^{2/3}}\right)\times\frac{\sigma^2}{32bL^2}\times\frac{48\lambda LI^2}{\rho}\\&\leq\frac{72\lambda^2I^4}{b\rho^2T}\sigma^2+\frac{3\lambda^2I^2}{2b\rho^2K^{2/3}T^{2/3}}\sigma^2.\end{aligned} \tag{30}$$

最后，对于上面式 (28) 中的最后一项，根据步长的定义，我们有 η $_t$，

$$\begin{aligned}&\frac{\kappa^3c^2}{\eta_T TL^2}\left(\frac{\sigma^2}{4b_1}+\frac{2\zeta^2}{21}\right)\ln(T+1)\\&\leq\left(\frac{48\lambda LI^2}{\rho T}+\frac{\lambda L}{\rho K^{2/3}T^{2/3}}\right)\times\frac{192^2\lambda}{L\rho}\times\left(\frac{\sigma^2}{4b_1}+\frac{2\zeta^2}{21}\right)\log(T+1)\\&\leq\frac{192^2\lambda^2}{\rho^2}\times\left(\frac{48I^2}{T}+\frac{1}{K^{2/3}T^{2/3}}\right)\times\left(\frac{\sigma^2}{4b_1}+\frac{2\zeta^2}{21}\right)\log(T+1).\end{aligned} \tag{31}$$

最后，将式 (29)、(30) 和 (31) 中得到的界代入式 (28)，可得

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\|\tilde{d}_t\|^2+\frac{\lambda^2}{\rho^2}\mathbb{E}\|\bar{e}_t\|^2\right)\\
&\leq\left[\frac{96\lambda^2LI^2}{\rho^2T}+\frac{2\lambda^2L}{\rho^2K^{2/3}T^{2/3}}\right](f(x_0)-f^*)+\left[\frac{72\lambda^2I^4}{b\rho^2T}+\frac{3\lambda^2I^2}{2b\rho^2K^{2/3}T^{2/3}}\right]\sigma^2\\
&\quad+\frac{192^2\lambda^2}{\rho^2}\times\left(\frac{48I^2}{T}+\frac{1}{K^{2/3}T^{2/3}}\right)\times\left(\frac{\sigma^2}{4b_1}+\frac{2\zeta^2}{21}\right)\log(T+1).
\end{aligned}$$

至此，定理得证。 □

现在我们可以证明定理 1。首先注意到：

$$\frac{\lambda^2\mathcal{G}_t}{\rho^2}=\frac{1}{\eta_t^2}||\tilde{x}_t-\tilde{x}_{t+1}||^2+\frac{\lambda^2}{\rho^2}||\bar{\nu}_t-\nabla f(\tilde{x}_t)||^2=||\tilde{d}_t||^2+\frac{\lambda^2}{\rho^2}||\bar{e}_t||^2$$

结合定理 2，可得：

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\mathcal{G}_t]\leq&\left[\frac{96LI^2}{T}+\frac{2L}{K^{2/3}T^{2/3}}\right](f(x_0)-f^*)+\left[\frac{72I^4}{bT}+\frac{3I^2}{2bK^{2/3}T^{2/3}}\right]\sigma^2\\
&+192^2\times\left(\frac{48I^2}{T}+\frac{1}{K^{2/3}T^{2/3}}\right)\times\left(\frac{\sigma^2}{4b_1}+\frac{2\zeta^2}{21}\right)\log(T+1).
\end{aligned}$$

注7。对于测度 $\mathcal{G}_t$，我们分别讨论其在无约束情形和有约束情形下的直观含义。首先，对于无约束情形，即当 $\mathcal{X}=R^d$ 时，我们有：

$$\begin{aligned}
&||\nabla f(\tilde{x}_{\tau,i})||/||H_\tau||=||H_\tau\times H_\tau^{-1}\nabla f(\tilde{x}_{\tau,i})||/||H_\tau||\leq||H_\tau^{-1}\nabla f(\tilde{x}_{\tau,i})||\\
&=||H_\tau^{-1}\nabla f(\tilde{x}_{\tau,i})-H_\tau^{-1}\bar{\nu}_{\tau,i}+H_\tau^{-1}\bar{\nu}_{\tau,i}||\leq||H_\tau^{-1}\nabla f(\tilde{x}_{\tau,i})-H_\tau^{-1}\bar{\nu}_{\tau,i}||+||H_\tau^{-1}\bar{\nu}_{\tau,i}||\\
&\leq\frac{1}{\rho}||\bar{\nu}_{\tau,i}-\nabla f(\tilde{x}_{\tau,i})||+\frac{1}{\lambda\eta_{\tau,i}}||\tilde{x}_{\tau,i}-\tilde{x}_{\tau,i+1}||\leq\sqrt{2}\sqrt{\mathcal{G}_{\tau,i}}/\rho
\end{aligned}$$

在最后一个不等式中，我们使用了詹森不等式；在倒数第二个不等式中，我们使用了假设 *4* 以及以下事实：$\tilde{x}_{\tau,\ i+1}=x_{\tau,\ 0}+\lambda H_\tau^{-1}\bar{z}_{\tau,\ i+1}$ 和 $\tilde{x}_{\tau,\ i}=x_{\tau,\ 0}+\lambda H_\tau^{-1}\bar{z}_{\tau,\ i}$ 和 $\eta_{\tau,\ i}\bar{\nu}_{\tau,\ i}=\bar{z}_{\tau,\ i+1}-\bar{z}_{\tau,\ i}$在无约束情形下。换言之，我们有 $||\nabla f(\tilde{x}_t)||^2\leq\frac{2||H_\tau||^2}{\rho^2}\mathcal{G}_\tau$。注意，右侧的系数是 H_τ 的平方条件数的上界。在自适应梯度方法的分析中，通常假设 H_t 的条件数 *[15]* 有限。总之，我们的测度 $\mathcal{G}_t$ 的收敛意味着在无约束情形下收敛到一阶驻点。

接下来，对于约束情形，我们的测度上界了梯度映射 $\frac{1}{\eta_{\tau+1,\ i}}||x_\tau-x^*_{\tau+1,\ i}||$，$x_t^*$ 定义如下：

$$x^*_{\tau+1,i}=\arg\min_{x\in\mathcal{X}}\{-\langle x,z^*_{\tau+1,i})\rangle+\frac{1}{2\lambda}(x-x_\tau)^TH_\tau(x-x_\tau)\}$$

其中，$z^*_{\tau+1,\ i} = \sum_{\ell=0}^{i} -\eta_\ell \nabla f(\tilde{x}_{\tau+1,\ i})$ 是真实梯度的累积。接下来，根据引理 *1*，我们有：

$$\begin{aligned}\|x^*_{\tau+1,i} - \tilde{x}_{\tau+1,i}\| &\le \frac{\lambda}{\rho}\|z^*_{\tau+1,i} - \bar{z}_{\tau+1,i}\| \\ &= \frac{\lambda}{\rho}\|\sum_{l=0}^{i-1} -\eta_{\tau+1,\ell}(\nabla f(\tilde{x}_{\tau+1,\ell}) - \bar{\nu}_{\tau+1,\ell}))\| \overset{(a)}{\le} \sum_{l=0}^{i-1}\frac{\lambda\eta_{\tau+1,\ell}}{\rho}\|\nabla f(\tilde{x}_{\tau+1,\ell}) - \bar{\nu}_{\tau+1,\ell}\|\end{aligned}$$

其中，不等式 (a) 来自三角不等式。接下来有：

$$\begin{aligned}\|x_\tau - x^*_{\tau+1,i}\| &= \|x_\tau - \tilde{x}_{\tau+1,i} + \tilde{x}_{\tau+1,i} - x^*_{\tau+1,i}\| \le \|x_\tau - \tilde{x}_{\tau+1,i}\| + \|\tilde{x}_{\tau+1,i} - x^*_{\tau+1,i}\| \\ &\le \|\sum_{l=0}^{i-1}\tilde{d}_{\tau+1,i}\| + \|\tilde{x}_{\tau+1,i} - x^*_{\tau+1,i}\| \le \sum_{l=0}^{i-1}\left(\|\tilde{d}_{\tau+1,\ell}\| + \frac{\lambda\eta_{\tau+1,\ell}}{\rho}\|\nabla f(\tilde{x}_{\tau+1,\ell}) - \bar{\nu}_{\tau+1,\ell}\|\right)\end{aligned}$$

由詹森不等式以及测度 (9) 的定义可得

$$\|\tilde{d}_t\| + \frac{\lambda\eta_t}{\rho}\|\nabla f(\tilde{x}_t) - \bar{\nu}_t\| \le \frac{\sqrt{2}\lambda\eta_t}{\rho}\sqrt{\mathcal{G}_t},$$

因此我们有

$$\frac{1}{\eta_{\tau+1,i}}\|x_\tau - x^*_{\tau+1,i}\| \le \frac{\sqrt{2}\lambda}{\rho}\sum_{l=0}^{i-1}\frac{\eta_{\tau+1,l}}{\eta_{\tau+1,i}}\sqrt{\mathcal{G}_{\tau+1,l}} \le \frac{2\sqrt{2}\lambda}{\rho}\sum_{l=0}^{i-1}\sqrt{\mathcal{G}_{\tau+1,\ell}},$$

最后一个不等式由公式 (17) 得出。总之，当测度为 $\mathcal{G}_{\tau+1,\ \ell} \to 0$ 时，梯度映射 $\frac{1}{\eta_{\tau+1,\ i}}\|x_\tau - x^*_{\tau+1,\ i}\|$ 收敛到 *0*。

推论 2. 超参数按定理 *2* 中的方式选取。假设我们设置 $I = O((T/K^2)^{1/6})$，并在第一步中使用大小为 $O(I^2)$ 的样本小批量，则有：

$$\mathbb{E}[\mathcal{G}_t] = O\left(\frac{f(x_0) - f^*}{K^{2/3}T^{2/3}}\right) + \tilde{O}\left(\frac{\sigma^2}{K^{2/3}T^{2/3}}\right) + \tilde{O}\left(\frac{\zeta^2}{K^{2/3}T^{2/3}}\right).$$

而要达到一个 ϵ-驻点，我们需要执行 $\tilde{O}(\epsilon^{-1.5}/K)$ 步，并需要 $\tilde{O}(\epsilon^{-1})$轮通信。

证明。 通过应用定理 2 并将 I 和 b 取为相应的值，可以直接验证推论中 $\mathbb{E}[\mathcal{G}_t]$ 的表达式。关于算法的梯度复杂度和通信复杂度，我们有以下结果：达到一个 ϵ-驻点所需的总步数 T 为 $\mathcal{O}\left(\frac{1}{K\epsilon^{3/2}}\right)$，即 $\tilde{O}\left(\frac{1}{K^{2/3}T^{2/3}}\right) = \epsilon$，也就是梯度复杂度。达到一个 ϵ-驻点所需的通信总轮数为 $E = T/I$，因为我们有 $I = \mathsf{O}((T/K^2)^{1/6})$，因此 $T/I = \tilde{O}\left(K^{1/3}T^{5/6}\right)$。假设客户端数量相对较大，更具体地说，假设 $K \ge \sqrt{T}$。那么我们有 $T/I = \tilde{O}\left(K^{1/3}T^{5/6}\right) = \tilde{O}\left(K^{2/3}T^{2/3}\right)$，换言之，我们有 $E = \tilde{O}\left(\epsilon^{-1}\right)$。推论的证明至此完成。

□