arXiv:2302.03109v1 [cs.LG] 6 Feb 2023

循环客户端参与下联邦平均的收敛性

Yae JeeCho卡内基梅隆大学
yaejeec@andrew.cmu.edu

Pranay Sharma卡内基梅隆大学
pranaysh@andrew.cmu.edu

Gauri Joshi卡内基梅隆大学
gaurij@andrew.cmu.edu

ZhengXu SatyenKale Tong ZhangGoogle Research Google Research Google Research 和香港科技大学 xuzheng@google.com satyenkale@google.com tozhang@google.com

摘要

联邦平均（FedAvg）及其变体是联邦学习（FL）中最受欢迎的优化算法。以往对 FedAvg 的收敛性分析，要么假设全客户端参与，要么假设部分客户端参与且客户端可以均匀采样。然而，在实际的跨设备联邦学习系统中，某一时刻只有满足本地条件（例如电池状态、网络连接和最大参与频率要求〔以确保隐私〕）的一部分客户端可用于训练。因此，客户端可用性呈现出自然的循环模式。据我们所知，我们首次提出了一个理论框架，用于分析采用循环客户端参与的 FedAvg 在使用多种不同客户端优化器（如 GD、SGD 和打乱式 SGD）时的收敛性。我们的分析发现，在适当条件下，与采用均匀客户端参与的标准 FedAvg 相比，循环客户端参与能够实现更快的渐近收敛速率，为客户端采样协议的设计提供了宝贵见解。

1 引言

联邦学习（FL）是一种分布式学习框架，使边缘客户端（例如手机、平板电脑）能够在不共享本地数据的情况下协同训练机器学习（ML）模型 [32]。在跨设备联邦学习中 [21], , 数百万台移动设备由中央服务器协调进行训练；由于连接间歇性中断和资源限制，每个通信轮次中只有一部分客户端设备会参与 [6]。

联邦平均（FedAvg） [32] 及其变体 [41, 45, 49] 是联邦学习中最受欢迎的算法。在广义 FedAvg 框架 [41, 49]的每个通信轮次中：1）服务器将当前模型广播给一部分客户端；2）客户端使用本地数据更新模型，并将本地模型更新发送回服务器；3）服务器聚合客户端的模型更新，并计算新的全局模型。由于多种原因，该算法在实践中广受欢迎，其中包括与联邦学习系统实现 [6] 的兼容性，以及差分隐私 [33] 和安全聚合 [5] 等额外隐私技术。

鉴于（广义）FedAvg（也称为本地随机梯度下降）在实践中广受欢迎，近期许多研究都对其收敛性进行了探讨 [25, 28, 50, 53] 。尽管这些分析应对了数据

异质性这一理论挑战，但它们假设客户端要么全员参与，即所有客户端每轮都会参与；要么部分参与，即从所有客户端中均匀随机地选取客户端。然而，在实际的跨设备联邦学习系统中，只有满足本地条件（例如处于闲置状态、接通电源充电以及连接到不限流量的网络）时，客户端才能参与训练 [6, 17, 19, 37]。Eichner 等人 [10], 、Yang 等人 [56], 和 Zhu 等人 [59] 观察到，客户端参与呈现昼夜模式；Balle 等人 [3],、Kairouz 等人 [21],和 Wang 等人 [49] 则讨论了控制参与采样的难度。受差分隐私启发 [22], , McMahan 和 Thakurta [31] 致力于通过限制每个客户端在较长时间窗口内最多参与一次，来限制其贡献。基于这些原因，在实际的联邦学习系统中，客户端通常以循环模式参与训练。

据我们所知，本文首次对循环客户端参与下的联邦平均进行了收敛分析。我们假设客户端被隐式划分为若干组，且这些组按循环顺序向服务器提供数据。我们证明，对于全局 PL 目标函数 [16], ，在适当条件下，循环客户端参与可实现更快的 $\widetilde{\mathcal{O}}\left(1/T^2\right)$ 收敛；而 FedAvg 达到的误差收敛速率为标准的 $\mathcal{O}(1/T)$，其中 T为通信轮次的数量，$\widetilde{\mathcal{O}}(\cdot)$涵盖了所有对数项和常数。这一关键洞见与近期一项关于小批量和本地更新洗牌随机梯度下降收敛性的研究[58] 类似；该研究表明，在客户端完全参与（而非循环且部分参与）的设置下，客户端本地数据洗牌能够实现快速收敛（更多详情见第 2.2 节）。

我们的分析框架涵盖了多种循环参与情形和不同的客户端优化器：1）它包括从每个组中对子集客户端进行子采样，这些客户端会按循环方式变为可用；2）它刻画了一个周期内的组数或客户端组的数据异质性特征如何影响收敛；3）它涵盖了不同的客户端本地过程，包括梯度下降（GD）、随机梯度下降（SGD）和打乱式 SGD（SSGD）。得益于这种通用性，若干经过深入研究的 FedAvg 变体，例如采用部分客户端参与的标准 FedAvg [20, 28] , 小批量随机重排和本地随机重排 [58], 都可以成为我们框架的特例。我们证明，在这些特例下，我们的界与先前工作中的界相符，从而验证了我们结果的有效性。我们还给出了初步实验结果，表明与标准 FedAvg 相比，循环客户端参与在测试准确率和训练损失收敛方面确实表现更佳。

2 相关工作

2.1 联邦学习中的客户端参与。

由于跨设备联邦学习中的客户端总数庞大，每个训练轮次只选择一部分客户端是不可避免的。因此，已有大量关于联邦学习中客户端参与的研究 [21, 27]。大多数研究聚焦于分析采用无偏部分客户端参与的 FedAvg [20, 55], 并证明其收敛速率为 $\mathcal{O}(1/T)$。虽然联邦学习领域的一些研究也考虑了有偏部分客户端参与，以实现灵活客户端参与 [42] 或依赖损失的客户端参与 [7, 11], ，但这些先前的研究并未考虑循环参与模式。

另一项相关研究方向是近期工作中对任意客户端参与情况的分析 [2, 51]。Wang & Ji [51] 提出了不同客户端参与模式的类别，其中循环客户端参与属于正则化参与类别。然而，由于该表述具有一般性，他们的分析未能捕捉一些重要特征，例如客户端的排序或一个周期内客户端组的数量如何影响收敛。Avdyukhin & Kasiviswanathan [2] 分析了客户端以异步方式发送本地更新的 FedAvg，其中每个客户端都有自己定义的周期区间，用于发送更新

。然而，这种框架无法模拟现实联邦学习系统中观察到的循环模式：客户端组依次变为服务器可用。

直到最近，循环客户端参与才从隐私 [8, 23] 和通信效率 [60]的角度在联邦学习中得到关注。Kairouz 等人 [23] 表明，循环客户端参与可以提升联邦学习中的隐私保障，但并未考察其收敛性质。Zhu 等人 [60] 表明，根据客户端的参与频率选择客户端可以加快收敛，收敛速率为 $\mathcal{O}(1/TV)$，其中 V是一个常数，取决于部分客户端参与时由数据异质性引起的方差。然而，由于缺少对变量 V的界限，收敛加速的确切速率尚不明确。与此前的研究不同，我们给出了联邦学习中循环客户端参与下的收敛结果，其加速速率明确（速率为 $\widetilde{\mathcal{O}}\left(1/T^2\right)$），并确定了实现该加速所需的条件。这种加速是通过从基于洗牌的方法这一视角分析采用循环参与的 FedAvg 得出的，下面将对此进行更详细的解释。

2.2 基于洗牌的方法。

关于强凸二次函数的基于洗牌的方法，最初的进展由 [13, 14] 取得。这些工作及后续研究的总体思路是：由于基于洗牌的方法在每个轮次中都会对每个分量函数恰好使用一次，因此这类方法在一个轮次内取得的进展近似于全批量梯度下降的进展。

关于基于洗牌的方法的文献主要关注三种轮次：(i) 随机重排（RR），即每个轮次结束后都会打乱数据；(ii) 单次洗牌（SO），即仅在开始时打乱一次数据；以及 (iii) 增量梯度（IG）方法，即数据完全不打乱，并在每个轮次中遵循预先确定的顺序。我们将在第 3 节（尤其是定理 1）中看到，循环客户端参与在服务器层面上本质上近似于增量梯度法，其中每个客户端都被视为一个样本。

近期研究已针对这些洗牌方案下的洗牌随机梯度下降建立了上界和下界。对于 RR（以及 SO）， [43] 证明了强凸二次函数的下界 $\mathcal{O}(\frac{1}{n^2K^2}+\frac{1}{K^3})$（$K$ 为轮次，n 为样本数），而 [40] 证明了一般强凸F 且具有光滑 $\{f_i\}$时的 $\mathcal{O}(\frac{1}{nK^2})$ 下界。在轮次较大的情形下， [1, 34] 针对光滑 PL 函数、 [35] 针对光滑强凸函数分别获得了匹配的上界。对于 IG， [36] 证明了具有光滑 $\{f_i\}$的强凸 F 的 $\mathcal{O}(1/K^2)$速率。上述所有工作中对 K 依赖性的改进都要求 K 大于 $\mathcal{O}(\kappa^a)$，其中 κ 是问题的条件数， $a\in[1,2]$。研究已表明，这一较大的轮次要求在 [44] 中是必不可少的。

3 问题表述

系统模型与目标。考虑一种跨设备联邦学习场景，其中共有 M 个客户端。每个客户端 $m\in[M]$ 都拥有自己的本地训练数据集 $\mathcal{B}_m$ 及相应的本地经验损失函数 $F_m(\mathbf{w})=\frac{1}{|\mathcal{B}_m|}\sum_{\xi\in\mathcal{B}_m}\ell(\mathbf{w};\xi)$，其中 $\ell(\mathbf{w};\xi)$ 表示模型 $\mathbf{w}\in\mathbb{R}^d$ 在数据样本 ξ 上的损失值。优化任务与标准联邦学习 [21, 32] 相同，其中全局目标函数为 $F(\mathbf{w})=\frac{1}{M}\sum_{m=1}^{M}F_m(\mathbf{w})$，服务器旨在找到使 $\min_{\mathbf{w}}F(\mathbf{w})$ 达到最小值的模型。本文中，所有向量范数和矩阵范数分别指欧几里得范数和谱范数。

循环客户端参与（CyCP）。我们考虑将这 M 个客户端划分为 K 个互不重叠的

客户端组，使得每个组包含 M/K 个客户端，如图 1 所示。客户端组记为 $\sigma(i)$ $i \in [K]$，其中每个 $\sigma(i)$ 包含相关客户端的索引。这些组以及服务器遍历它们的顺序（例如， $\sigma(1)$ ……， $\sigma(K)$）均预先确定，并在整个训练过程中保持不变，以模拟客户端参与的循环结构。在每个通信轮次中，一旦客户端组 $\sigma(i)$ 可用，服务器便从 $\sigma(i)$ 中均匀随机地不放回抽样一个包含 N 个客户端的子集。因此，由于循环结构以及在每个可用客户端组内进行的子采样，客户端一旦被选中，至少在接下来的 $K-1$ 个轮次中无法参与训练。为简洁起见，本文将这一循环客户端参与框架称为 CyCP。

请注意，CyCP 框架体现了第1节中提到的若干实际联邦学习场景。在 CyCP 中，每个客户端在连续的 K 个通信轮次内最多只能参与一次，这满足隐私要求 [8, 23]。可以在每个客户端上设置计时器，确保客户端只有在经过一段时间后才能再次参与训练，而这段时间对应于 K。即使不强制执行基于计时器的标准，CyCP 也能反映客户端处于不同时区，或偏好在设备充电时参与训练所产生的自然参与模式 [19, 37, 56]。



图1：循环客户端参与（CyCP）的示意图，其中 $M=12$ 个客户端被划分为 $K=3$ 个组。在每个通信轮次中，从当时可用的客户端组中选出 $N=2$ 个客户端进行训练。所有组会在由 K 个通信轮次组成的一个周期轮次中依次遍历一次。

我们引入“周期轮次”这一术语，指服务器依次遍历所有客户端组 $\{\sigma(i) \quad i \in [K]\}$ 一次所经历的时间段。换言之，每个周期轮次由 K 个通信轮次组成。形式化地，我们将 k 设为周期轮次的索引，将 $i \in [K]$ 设为一个周期轮次内当前可用客户端组的索引。服务器将全局模型$\mathbf{w}^{(k,i-1)}$ 发送给集合 $\mathcal{S}^{(k,i)}$ 中的 N个客户端，这些客户端从客户端组 $\sigma(i)$ 中选出，以执行本地训练。我们考虑三种不同类型的客户端本地更新，并将在下文详细说明。

客户端本地更新。 每个客户端 $m \in \mathcal{S}^{(k,i)}$ 将其本地模型初始化为 $\mathbf{w}_m^{(k,i-1,0)} = \mathbf{w}^{(k,i-1)}$，并执行本地更新。全局模型更新如下：

$$\mathbf{w}^{(k,i)} = \mathbf{w}^{(k,i-1)} + \Delta^{(k,i-1)} \tag{1}$$

其中， $\Delta^{(k,i-1)}$ 是来自 $\mathcal{S}^{(k,i)}$ 中客户端的本地更新的聚合结果。对于 CyCP 框架，我们考虑三种不同的客户端本地更新过程。

(i) 本地梯度下降（GD）：选定客户端执行单个 GD 步骤来更新其本地模型。因此，

$$\Delta^{(k,i-1)} = -\frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,i-1)})$$

所得的全局算法在 [32] 中称为 FedSGD。

(ii) 本地随机梯度下降（本地 SGD）：为避免计算完整梯度的开销，每个客户端使用随机梯度 $\nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)})$ 对其模型执行 τ 次本地更新；这些随机梯度通过一个小批量 $\xi_m^{(k,i-1,l)}$从客户端 m 的本地数据集 $\mathcal{B}_m$ 中均匀随机抽取得到。因此，客户端的模型更新为

$$\Delta^{(k,i-1)} = -\frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{\tau-1} \nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)})$$

其中 $\mathbf{w}_m^{(k,i-1,l+1)} = \mathbf{w}_m^{(k,i-1,l)} - \eta \nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)})$。

(iii) 本地随机重排随机梯度下降（SSGD）：近期联邦学习研究 [30, 58] 提出采用本地 SSGD，即客户端将本地数据集划分为 B 个互不相交的部分，也就是说，客户端 m 的本地损失可表示为 $F_m(\mathbf{w}) = \frac{1}{B}\sum_{l=0}^{B-1} F_{m,l}(\mathbf{w})$。我们将 $\mathcal{P}_B$ 定义为 $\{0, ..., B-1\}$的所有排列构成的集合。在每一轮中，客户端按照随机排列 $\pi_m^k \sim \text{Unif}(\mathcal{P}_B)$所决定的顺序遍历所有部分并执行本地更新。由此得到的模型更新为

$$\Delta^{(k,i-1)} = -\frac{\eta}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}_m^{(k,i-1,l)})$$

其中 $\mathbf{w}_m^{(k,i-1,l+1)} = \mathbf{w}_m^{(k,i-1,l)} - \eta\nabla F_{m,\ \pi_m^k\ (l)}\ (\mathbf{w}_m^{(k,i-1,l)})$.

算法 1 展示了我们采用 CyCP 的联邦学习框架的更多细节。

CyCP的特殊情况。CyCP 框架涵盖了多种算法，例如采用部分客户端参与的标准 FedAvg，以及 [58] 中介绍的小批量随机重排和本地随机重排。当 $K = 1$ 时，CyCP 设置变为采用部分客户端参与的标准 *FedAvg*，其中每一轮都会从整个客户端群体中抽取 N 个客户端。当 $K = 1$ 且 $N = M$ 时，采用局部梯度下降的 CyCP 等同于小批量随机重排 [58]，其中有 M 个客户端，且每个客户端仅包含一个分量。两者都以指数速度收敛到最优解。另一种特殊情况是采用 $K = 1$ 和 $N = M$，但使用本地 SSGD；此时得到的是本地随机重排 [58]，每个客户端包含 B 个分量，并且每处理 B 个分量就同步一次更新。我们在第4节中证明，对于这些特殊情况，我们的理论结果相应地与已有界限一致。

算法 1 CyCP 在联邦学习中的框架

1：输入：全局模型 $\mathbf{w}^{(1,0)}$，客户端组 $\sigma(i)$ $i \in [K]$2：输出：全局模型 $\mathbf{w}^{K+1}$，0 3：对于 $k \in [K]$ 个周期轮次：# 循环参与 4：对于 $i \in [K]$：# $T = KK$ 个通信轮次 5：从客户端集合 $\sigma(i)$ 中均匀随机且不放回地抽取 $^{(k,i-1)}$ 发送给 $\mathcal{S}^{(k,i)}$中的客户端。 7：客户端 m $^{i-1)}, case)$ 9：$\mathbf{w}^{(k,i)} = \frac{1}{N}\sum_{m\in\mathcal{S}(k,i)}\mathbf{w}_m^{(k+1,0)}$10：$\mathbf{w}^{(k+1,0)} = \mathbf{w}^{(k,K)}$LocalUpdate($m, \mathbf{w}$ $case$)：11：设置本地模型 $\mathbf{w}_m = \mathbf{w}$ 12：如果 $case == LocalGD$：13：更新 $\mathbf{w}_m \leftarrow \mathbf{w}_m - \eta\nabla F_m(\mathbf{w}_m)$ 14：否则如果 $case == LocalSGD$：15：对于 $j \in [\tau]$：16：从本地数据集 $\mathcal{B}_m$中抽取小批量 ξ 17：更新 $\mathbf{w}_m \leftarrow \mathbf{w}_m - \eta\nabla F_m(\mathbf{w}_m, \xi_m)$ 18：否则如果 $case == ShuffledSGD$：19：抽取 $\pi_m^k \sim$Unif($\mathcal{P}_B$) 20：对于 $j \in [B]$：21：更新 $\mathbf{w}_m \leftarrow \mathbf{w}^m - \eta\nabla F_{m,\pi_m^k(j-1)}(\mathbf{w}_m)$

4 收敛分析

本节给出并比较上述三种客户端本地更新方法下，联邦学习中 CyCP 的收敛界，并说明 CyCP ($K > 1$) 所达到的复杂度与标准 FedAvg ($K = 1$) 相比如何。所有证明均见附录 C-E。

4.1 假设

首先，我们给出本文收敛保证所采用的假设。

假设 1 ($F_m(\mathbf{w})$ 的光滑性 $\forall\, m$)。 客户端的局部目标函数 $F_1(\mathbf{w})$ ...， $F_M(\mathbf{w})$ 均为 L-光滑，即 $\|\nabla F_m(\mathbf{w}) - \nabla F_m(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|$ 对所有 m、$\mathbf{w}$ 和 $\mathbf{w}'$ 均成立。

假设 2 (μ-Polyak-Łojasiewicz $F(\mathbf{w})$)。 对于某个 $\mu > 0$，全局目标函数满足 $\frac{1}{2}\|\nabla F(\mathbf{w})\|^2 \geq \mu(F(\mathbf{w}) - \min_{\mathbf{w}'} F(\mathbf{w}'))$，对所有 $\mathbf{w}$ 均成立。

假设 1、2 在优化和联邦学习文献中很常见 [12, 15, 16, 24]。尽管为简洁起见，并便于与先前工作进行比较 [58], ，我们将研究范围限制在 PL 函数，但使用 [29]中提出的技术，可以将这些分析推广到一般的非凸函数。

接下来，我们给出客户端组 $\sigma(1), \ldots, \sigma(K)$ 上的假设。

假设 3 （组内与组间数据异质性）。 存在常数 γ， $\alpha \geq 0$，使得对于所有 $\mathbf{w}$、所有 $i \in [K]$ 以及所有 $m \in \sigma(i)$， $\|\nabla F_m(\mathbf{w}) - \frac{1}{|\sigma(i)|}\sum_{m\in\sigma(i)} \nabla F_m(\mathbf{w})\| \leq \gamma$，并且 $\|\frac{1}{|\sigma(i)|}\sum_{m\in\sigma(i)} \nabla F_m(\mathbf{w}) - \nabla F(\mathbf{w})\| \leq \alpha$。

假设 3 将组内 客户端之间的数据异质性限制为 γ，并将组间 数据异质性限制为 α。假设 3 还意味着先前联邦学习文献中常用的数据异质性假设 [26, 41, 47, 48, 57] 如下：

引理 4.1. 若假设 *3* 成立，则存在 $\nu = \gamma + \alpha \geq 0$，使得对于所有客户端 $m \in [M]$ 以及所有 $\mathbf{w}$，均有 $\left\|\nabla F_m(\mathbf{w}) - \frac{1}{M}\sum_{i=1}^{M} \nabla F_i(\mathbf{w})\right\| \leq \nu$。

与标准假设相比，使用假设 3 能够根据 γ 和 α 推导出更紧的收敛界，并将两种异质性的影响区分开来，正如我们将在后续章节中讨论的那样。

4.2 局部梯度下降下 CyCP 的收敛

首先，我们从给出局部梯度下降下 CyCP 中全局模型的收敛性开始。

定理 1 (CyCP 下的收敛+梯度下降)。 在假设 *1*、*2*、*3* 成立的条件下，步长取为 $\eta = \log(MT^2/K^2)/\mu NT$，通信轮次取为 $T \geq 7\kappa K \log(MT^2/K^2)$，其中 $\kappa = L/\mu$:

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{\overline{K}^2(F(\mathbf{w}^{(0,0)}) - F^*)}{MT^2} + \tilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu T^2}\right) + \tilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\gamma^2}{\mu NT}\left(\frac{M/\overline{K} - N}{M/\overline{K} - 1}\right)\right), \quad (2)$$

其中 $\tilde{\mathcal{O}}(\cdot)$ 涵盖了所有对数项和常数。

尽管乍看之下，由于 K 出现在各项的分子中，界似乎会随着其增大而变差，但由于 $T = KK$ （见算法 1），较大的 K 不会对收敛产生不利影响。

收敛对 γ^2 和 α 的依赖。定理 1 表明，CyCP+梯度下降以 $\tilde{\mathcal{O}}\left(\frac{K\kappa\gamma^2}{\mu NT}\left(\frac{M/K-N}{M/K-1}\right)\right)$ 的速率收敛，该速率取决于 γ^2，即组内数据异质性。因此，较大的组内数据异质性 γ 会导致收敛表现变差。相反，如果 $\gamma \simeq 0$，CyCP+梯度下降则能够实现 $\tilde{\mathcal{O}}\left(1/T^2\right)$ 收敛，因为 $\tilde{\mathcal{O}}$（$1/T$）主导项会变为零。因此，在同一组内客户端的数据分布相似（即 γ 接近 0）的 CyCP 设置中，与标准 FedAvg（$K = 1$）相比，CyCP+梯度下降能够实现更快的收敛速率。在现实的联邦学习场景中，循环模式遵循昼夜模式时，便可能自然出现这种情况，如 [59] 所示。还值得注意的是，定理 1 中涉及组间数据异质性 α 的项以 $\widetilde{\mathcal{O}}\left(1/T^2\right)$ 的速率衰减。因此，在 CyCP 设置中，组内数据异质性 γ 对收敛误差的影响比组间数据异质性 α 更显著。

收敛对.....的依赖K。定理 1 还表明，即使对于 $\gamma \neq 0$，CyCP+梯度下降在 $K = M/N$ 时也能获得 $\tilde{\mathcal{O}}\left(1/T^2\right)$ 收敛速率。这一速率快于标准 FedAvg（设置为 $K = 1$），其速率为 $\mathcal{O}$（$1/T$）。虽然定理 1 已清楚给出了 $K = 1$ 和 $K = M/N$情形下的收敛速率，但对于中间区域 $1 < K < M/N$，情况仍不明朗。为此，我们比较 CyCP+梯度下降与标准 FedAvg 达到 ϵ 误差所需的总成本。我们将梯度下降每个通信轮次中的通信与计算成本（包括计算 N个梯度并向服务器传输 N 个向量）定义为 c_{GD}。然后，仅考虑式 (2) 中的主导项，达到 ϵ 误差所需的总成本 C_{GD} 为

$$C_{\mathrm{GD}}(\epsilon) = \tilde{\mathcal{O}}\left(\frac{c_{\mathrm{GD}}K\gamma^2}{\epsilon N}\left(\frac{M/K-N}{M/K-1}\right)\right) \tag{3}$$

我们将 $C_{\mathrm{GD}}(\epsilon)$ 与 $K > 1$ 和 $K = 1$ 分别记作 $C_{\mathrm{GD}|K>1}(\epsilon)$ 和 $C_{\mathrm{GD}|K=1}(\epsilon)$，并得出以下结果：

推论 1. 对于定义为 (3) 的总成本，当 $K < M/N$ 时，有 $C_{GD|K>1}(\epsilon) > C_{GD|K=1}(\epsilon)$。

推论 1 表明，当组数设为中间范围，即 $1 < K < M/N$ 时，CyCP 的成本并不低于标准 FedAvg（$K = 1$）。因此，要使 CyCP+梯度下降的成本低于标准 FedAvg，必要条件是满足 $K = M/N$。在某些情况下，K 可能是服务器无法控制的自然产生的量。值得注意的是，在这些情况下，服务器可以相应地选择 N（即每轮选定客户端的数量），以使成本低于标准 FedAvg。

与小批量随机重排相匹配的界 [58]。回想一下，当采用 $K = 1$ 和 $N = M$时，算法 1 中的 CyCP+梯度下降与小批量随机重排算法类似，其中有 M 个客户端，且每个客户端只有一个分量。在这种情况下，定理 1 中的界只剩下第一项，其衰减速率为 $\widetilde{\mathcal{O}}\left(1/MT^2\right)$，这恰好与 [定理 1] [58] 中的小批量随机重排界相匹配；该定理表明算法具有指数收敛性。

4.3 CyCP 与本地随机梯度下降的收敛性

接下来，我们给出采用本地随机梯度下降的 CyCP 的收敛结果。与梯度下降相比，本地随机梯度下降在推导收敛分析时带来了额外的技术挑战，因此需要以下额外的

随机梯度假设:

假设 4 （有界方差）。 对于局部目标函数 $F_m(\mathbf{w})$，使用从 $\mathcal{B}_m$ 中均匀随机抽取的小批量 ξ_m 计算得到的局部随机梯度 $\nabla F_m(\mathbf{w}\ \xi_m)$ 的方差有界，即$\mathbb{E}[\|\nabla F_m(\mathbf{w}\ \xi_m) - \nabla F_m(\mathbf{w})\|^2] \leq \sigma^2$，对所有 $m \in [M]$ 均成立。

假设 4 在随机优化文献中十分常用 [4, 28, 42, 46]。现在我们给出本地随机梯度下降的收敛界。

定理 2 （CyCP 下的收敛+随机梯度下降）。 在假设*1*、*2*、*3*和*4*以及步长 $\eta = \log(MT^2/K^2)/\tau\mu NT$ 的条件下，对于 $T \geq 10\kappa K \log(MT^2/K^2)$ 个通信轮次，其中 $\kappa = L/\mu$，收敛误差有如下界:

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{\overline{K}^2(F(\mathbf{w}^{(0,0)}) - F^*)}{MT^2} + \widetilde{\mathcal{O}}\left(\frac{\kappa^2\overline{K}(\overline{K}-1)\alpha^2}{\mu T^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\gamma^2}{\mu NT}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right) + \widetilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\sigma^2}{\mu\tau NT}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\tau-1)\nu^2}{\mu\tau N^2T^2}\right) \tag{4}$$

其中 $\widetilde{\mathcal{O}}(\cdot)$ 涵盖了所有对数项和常数项。

同样，尽管随着 K增大，界看起来可能会变差，因为算法 1 中有 $T = KK$ ，但较大的 $\overline{K}$ 不会对收敛产生不利影响。

收敛对 γ^2 和 σ^2的依赖。 在定理 2 中，主导的 $\widetilde{\mathcal{O}}(1/T)$ 项取决于两个因素：组内数据异质性 γ^2 （局部梯度下降中也出现了这一因素）和随机梯度方差 σ^2。因此，对于 $\sigma > 0$，即使采用 $K = M/N$或 $\gamma^2 \simeq 0$，我们也无法像局部梯度下降的情形那样达到 $\widetilde{\mathcal{O}}\left(1/T^2\right)$收敛速率。因此，即使使用 CyCP，采用本地随机梯度下降时所能达到的最佳收敛速率也只是 $\mathcal{O}(1/T)$。看到这一结果后，人们可能会想，使用 CyCP 执行本地随机梯度下降究竟是否有任何优势。下面，我们将 CyCP 与本地随机梯度下降的成本和标准 FedAvg 的成本进行比较，以回答这个问题。

CyCP（$K > 1$）搭配本地随机梯度下降对联邦学习有帮助吗? 乍看定理 2，人们可能会认为，在本地随机梯度下降的情况下，CyCP 并不能提高收敛速率，因为随机梯度方差出现在一个主导项 $\widetilde{\mathcal{O}}\left(\frac{K\kappa\sigma^2}{\mu\tau NT}\right) + \widetilde{\mathcal{O}}\left(\frac{K\kappa\gamma^2}{\mu NT}\left(\frac{M/K-N}{M/K-1}\right)\right)$ 中。然而，我们在推论 2 中证明，情况并非总是如此。与上一节中对 c_{GD} 的定义类似，我们将采用本地随机梯度下降时每个通信轮次的总通信与计算成本定义为 c_{SGD}。具体而言，仅考虑式 (4) 中的主导项时，对于本地随机梯度下降，要达到 ϵ 误差所需的总成本为:

$$C_{\mathrm{SGD}}(\epsilon) = \widetilde{\mathcal{O}}\left(\frac{c_{\mathrm{SGD}}\overline{K}\gamma^2}{\epsilon N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right) + \widetilde{\mathcal{O}}\left(\frac{c_{\mathrm{SGD}}\sigma^2\overline{K}}{\epsilon N\tau}\right) \tag{5}$$

我们分别将 CyCP 和标准 FedAvg 的成本记为 $C_{\mathrm{SGD}|K=1}$， $C_{\mathrm{SGD}|K>1}$。现在，我们给出使 $C_{\mathrm{SGD}|K>1} < C_{\mathrm{SGD}|K=1}$成立的条件，即 CyCP 的成本低于标准 FedAvg。

推论 2. 假设我们有 $N = M/K$，且组内数据异质性 γ满足 $\gamma^2 \geq M\sigma^2/N\tau$。那么，我们得到 $C_{SGD|K} \leq C_{SGD|K=1}$。

推论 2 表明，当组内数据异质性显著大于随机梯度方差除以本地迭代次数时，与标准 FedAvg ($K = 1$) 相比，CyCP +随机梯度下降确实能以更低的成本达到 ϵ 误差。请注意，可以通过增大小批量大小 b（从而降低方差 σ^2）或增加本地迭代次数 τ，来满足推论 2 中的条件 $\gamma^2 \geq M\sigma^2/N\tau$。

与标准 FedAvg 匹配的界。 对于 $K=1$，CyCP+随机梯度下降可还原采用本地随机梯度下降的标准 FedAvg；在全客户端参与的情况下，我们在 (4) 中得到的界具有 $\widetilde{\mathcal{O}}\left(\kappa^2\nu^2/\mu MT^2\right)+\widetilde{\mathcal{O}}\left(\kappa\sigma^2/\mu\tau MT\right)$ 的阶。我们证明，该界与 [38] 中的最后一次迭代界相匹配；该文假设全客户端参与，且随机梯度具有参数为 G 的有界范数，其形式为 $\widetilde{\mathcal{O}}\left(\kappa^2G^2/\mu T^2\right)+\widetilde{\mathcal{O}}\left(\kappa\sigma^2/\mu\tau T\right)$。其中，他们的学习率不像我们的情况那样随 M 衰减，因此其界中缺少 $1/M$。第一项的差异源于：他们的工作假设随机梯度具有有界范数 G，而我们只假设随机梯度具有有界方差。

4.4 使用本地 SSGD 的 CyCP 收敛性

对于 CyCP 的最后一种情形，我们给出采用本地 SSGD 时全局模型收敛性质的结果。我们对每个客户端的本地损失定义稍作修改，将其定义为 $F_m\mathbf{w}=\frac{1}{B}\sum_{l=0}^{B-1}F_{m,\ l}\mathbf{w}$，使得每个客户端有 B 个损失分量。对于本地 SSGD，客户端依次执行本地更新，遍历 $F_{m,\ \pi_m(l)}(\mathbf{w})$ $l\in[0,...,B-1]$，其中 $\pi_m\sim\text{Unif}(\mathcal{P}_B)$ 是对 B 个分量进行的随机排列，$\pi_m(l)$ 表示该排列中的第 l 个元素。对于 SSGD，我们需要用以下客户端内分量异质性假设取代假设 4；该假设在打乱式 SGD 文献中很常见 [30, 58]:

假设 5（客户端内分量异质性）。 存在常数 $\nu\geq0$，使得对于每个客户端 $m\in[M]$ 及其本地数据集中的每个分量 $l\in[0,...,B-1]$，$\|\nabla F_{m,\ l}(\mathbf{w})-\nabla F_m(\mathbf{w})\|\leq\nu$，对所有$\mathbf{w}$ 均成立。

现在我们给出 SSGD 情况下的收敛结果。

定理 3（CyCP 下的收敛+SSGD）。 在假设 *1*、*2*、*3* 和 *5* 下，且 $\eta=\log(MBT^2/K^2)/\mu BT$以及周期轮次 $T\geq10\kappa K\log\left(MBT^2/K^2\right)$其中 $\kappa=L/\mu$，以至少 $1-\delta$ 的概率，收敛误差满足如下界：

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}^{(K,0)})]-F^*\leq&\frac{K^2(F(\mathbf{w}^{(0,0)})-F^*)}{MBT^2}+\widetilde{\mathcal{O}}\left(\frac{\kappa^2(B-1)^2\nu^2}{\mu B^2T^2}\right)+\widetilde{\mathcal{O}}\left(\frac{\kappa^2(K-1)^2\alpha^2}{\mu T^2}\right)\\&+\widetilde{\mathcal{O}}\left(\frac{\kappa^2\nu^2}{\mu T^2}\left(\frac{(B^{3/2}-1)^2}{B^4}+\frac{(B-1)^2}{B^3}\right)\right)+\widetilde{\mathcal{O}}\left(\frac{\kappa K\gamma^2}{\mu NT}\left(\frac{M/K-N}{M/K-1}\right)\right)\end{aligned}\tag{6}$$

其中 $\widetilde{\mathcal{O}}(\cdot)$ 涵盖了所有对数项和常数。

同样，由于 $T=KK$，增大 K 不会对上述界产生不利影响。

对 K 的依赖性与 CyCP+梯度下降相同。 定理 3 表明，主导项为 $\widetilde{\mathcal{O}}\left(\frac{\kappa K\gamma^2}{\mu NT}\left(\frac{M/K-N}{M/K-1}\right)\right)$，这与定理 1 中局部梯度下降收敛速率的主导项相同。只要 $\gamma\neq0$ 或 $K\neq M/N$，该主导项就存在。因此，对于 $1<K<M/N$，与局部梯度下降的情况相同（见推论 1），CyCP+SSGD 并不会比标准 FedAvg 降低成本。同样，与 CyCP+梯度下降一样，$K=M/N$ 是 CyCP+SSGD 相较于标准 FedAvg 实现任何成本降低的必要条件。下面我们将更详细地比较这两种不同本地更新过程的成本。

CyCP+SSGD 是否优于 CyCP+梯度下降? 如上所述，由于主导项相同，且对于 $\gamma^2\neq0$ 和 $K<M/N$，该项非零，因此 CyCP+SSGD 和 CyCP+梯度下降以相同的速率收敛。然而，对于 $K=M/N$，主导项趋于 0，此时 CyCP+SSGD 和 CyCP+梯度下降的表现相当。

同样，我们将采用本地 SSGD 时每个通信轮次的总通信与计算成本定义为 c_{SSGD}。那么，对于 $K = M/N$，CyCP+梯度下降和 CyCP+SSGD 达到 ϵ 误差（分别记为 $C_{\mathrm{GD}|K=M/N}(\epsilon)$ $C_{\mathrm{SSGD}|K=M/N}(\epsilon)$）所需的总成本为

$$C_{\mathrm{GD}|\overline{K}=M/N}(\epsilon) = \widetilde{\mathcal{O}}\left(\frac{c_{\mathrm{SSGD}}}{\sqrt{\epsilon}}\left(\frac{\overline{K}}{\sqrt{M}} + \overline{K}\alpha\right)\right) \tag{7}$$

$$C_{\mathrm{SSGD}|\overline{K}=M/N}(\epsilon) = \widetilde{\mathcal{O}}\left(\frac{c_{\mathrm{SSGD}}}{\sqrt{\epsilon}}\left(\frac{\overline{K}}{\sqrt{MB}} + \nu + \overline{K}\alpha + \frac{\overline{\nu}}{\sqrt{B}}\right)\right) \tag{8}$$

在这些成本下，只有当 $K = M/N$时，CyCP+SSGD的成本才低于CyCP+GD的情况

$$1 - \frac{1}{\sqrt{B}} - \nu - \frac{\overline{\nu}}{\sqrt{B}} > 0,\ B > 1 \tag{9}$$

能够确定满足（9）的唯一条件是 $\nu \simeq \nu \simeq 0$。因此，对于 $K = M/N$, 只有当客户端之间的数据异质性接近于 0，且每个客户端内部的组件异质性也接近于 0 时，CyCP+SSGD的成本才可能低于CyCP+GD。这也与 [52, 58]中给出的理论结果高度一致。

具有CyCP特殊情况的匹配界+SSGD。CyCP+SSGD的特殊情况类似于不同的算法，例如CyCP+GD或本地随机重排 [58]。对于 $B = 1$，由于 $\nu = 0$ ，我们可以得到CyCP+GD（定理 1）。另一种特殊情况是，当 $K = 1,\quad N = M$ 对于CyCP+SSGD时。在这种情况下，对于每个通信轮次，我们都采用全客户端参与，每个客户端都对其本地组件执行SSGD，这类似于[58]中提出并进行理论分析的本地随机重排算法。请注意，在这种情况下，定理 3 中的界与 [58]中给出的本地随机重排界一致。下一段将对二者进行更详细的比较。

CyCP+SSGD 是否优于本地随机重排? 由于本地随机重排假设全客户端参与，为了公平比较，我们将本地随机重排与满足 $K = M/N$ 的 CyCP+SSGD 进行比较。定理 3 中 CyCP+SSGD 的收敛界里的主导项变为零，因此只剩下收敛速率为 $\widetilde{\mathcal{O}}\left(1/T^2\right)$ 的项。我们比较这两种算法达到 ϵ 误差所需的总成本，其中本地随机重排的总成本为：

$$C_{\mathrm{LocalRR}}(\epsilon) = \widetilde{\mathcal{O}}\left(\frac{\overline{K}c_{\mathrm{SSGD}}}{\sqrt{\epsilon}}\left(\frac{1}{\sqrt{MB}} + \nu + \frac{\overline{\nu}}{\sqrt{B}}\right)\right) \tag{10}$$

请注意，对于本地随机重排，其计算成本和通信成本均为 CyCP+SSGD 的 K 倍，因为在本地随机重排中，所有客户端都参与每个通信轮次，而在 CyCP+SSGD 中，每个通信轮次只有 M/K 个客户端参与。结合 CyCP+SSGD 在 $K = M/N$ 下的成本（见式 (8)）以及本地随机重排的成本（见式 (10)），我们得到以下比较这两种方法的理论结果：

推论 3. 对于足够大的 M，满足

$$M > N\left(1 + \frac{\alpha}{\gamma + \frac{\overline{\nu}}{\sqrt{B}}}\right) \tag{11}$$

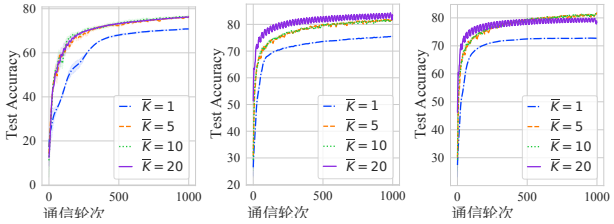
其中，对于 *CyCP+SSGD*，*CyCP+SSGD* 在达到 *epsilon* 误差所需的总成本方面始终 优于本地随机重排。

推论 3 表明，当 $K = M/N$ 且 M 足够大时，为了降低成本，CyCP+SSGD 始终优于本地随机重排。由于 $K = M/N$，较大的 M 表示较大的 K。观察下界

推论 3 中关于 M 的下界，在 α 较小、 γ 较大时会更小。这表明，为了使 CyCP+SSGD 优于本地随机重排，客户端组的组间数据异质性应较小，而组内数据异质性应较大。

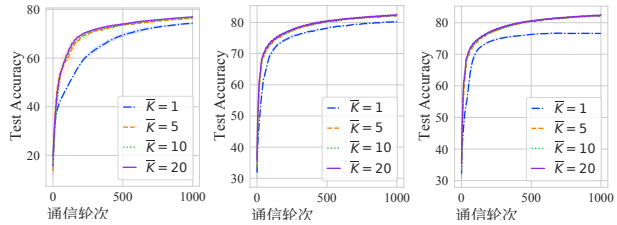
5 实验结果

设置。 我们使用 FedAvg 和 CyCP，在标准数据集上训练机器学习模型，并针对不同的客户端本地更新过程，考察循环性如何影响联邦学习的性能。我们使用多层感知机进行图像分类实验，数据集为 FMNIST [54] 和 EMNIST [9]；其中，EMNIST 数据集包含 62 个标签。实验中，客户端总数分别为 100 和 500，每个通信轮次分别选择 5 个和 10 个客户端。我们使用狄利克雷分布 $\mathrm{Dir}_K(\alpha)$ [18] 在客户端之间划分数据，其中 α 决定客户端之间数据异质性的程度。 α 越小，数据异质性越大。我们使用三个不同的随机种子对客户端之间的数据集划分进行随机化，并报告平均结果。受篇幅限制，实验的更多细节以及展示训练损失的结果见附录 A。



(a) 梯度下降 (b) 随机梯度下降 (c) SSGD

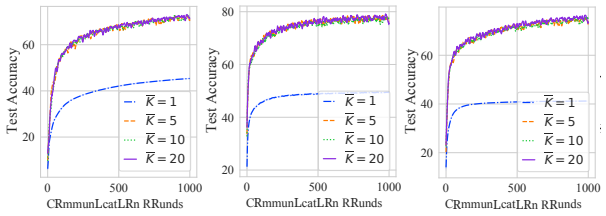
图2： 高数据异质性（$\alpha = 0.5$）下 FMNIST 的测试准确率。对于所有不同的客户端本地过程，CyCP（$K > 1$）的测试准确率均高于 $K = 1$ （标准 FedAvg），提升了 5–10%。



(a) 梯度下降 (b) 随机梯度下降 (c) SSGD

图3： FMNIST 在低数据异质性条件下的测试准确率（$\alpha = 2.0$）。与图 2 中高数据异质性的情况一致，对于各种客户端本地过程，CyCP（$K > 1$）的测试准确率均高于 $K = 1$ （标准 FedAvg），提升幅度为 2%–8%。不过，CyCP 与标准 FedAvg 之间的性能差距小于数据异质性较高时的差距。

.....的影响 K 与数据异质性。我们在图 2 中展示了 FMNIST数据集在不同 K 值和客户端本地过程下、数据异质性较高时（$\alpha = 0.5$）的测试准确率。回想一下， $K = 1$ 与标准 FedAvg 类似，而 $1 < K \leq M/N$ （在 FMNIST 情况下， $M/N = 20$ ）意味着循环客户端参与。较高的 K 表示服务器在单个周期内访问更多客户端组。我们表明，对于较高的数据异质性，无论采用哪种客户端本地过程，较高的 K 都能使测试准确率提升约 5-10%。尽管理论结果表明，CyCP 仅在 下表现出



(a) 梯度下降　(b) 本地随机梯度下降　(c) 打乱式 SGD

图 4：高数据异质性（$\alpha = 0.05$）下 EMNIST 的测试准确率。对于所有不同的客户端本地过程，CyCP 的测试准确率均高于标准 FedAvg。其提升幅度大于 FMNIST，这主要是因为 EMNIST 的数据集划分具有更高的数据异质性。

收敛性提升，但我们发现即使在 $M/\overline{N} > K > 1$下也有提升。这可能是因为我们的理论结果针对的是 PL 目标函数，而深度神经网络的损失景观未必属于这一类别 [39]。对于图 3 所示的数据异质性较低的情况，与标准 FedAvg 相比，$K > 1$的提升约为 2-8%。这一提升低于数据异质性较高时的提升，这与第4节的理论结果一致：该结果表明，增大 K会降低依赖于组内数据异质性的主导项。在图 4 中，我们展示了 CyCP 与标准 FedAvg 之间的性能差距更大，因为其数据异质性甚至高于 FMNIST 情况。

客户端本地过程之间的差异。 在图 2(b)- (c) 和图 4(b)-(c) 中可以观察到一个显著特点，它们展示了高数据异质性下本地随机梯度下降和 SSGD 的结果：对于最高的 K，测试准确率曲线会出现振荡。这是由于 CyCP 所致；随着 K 增大，组间数据异质性也会随之增大，导致服务器依次访问各组进行训练时出现振荡。以往研究也观察到了这种现象：服务器以循环方式依次在不同组上进行训练 [59]。附录 A 中图 5 和图 6 所示的训练损失曲线也呈现出类似的振荡现象。

6 总结

在实际的联邦学习系统中，循环客户端参与十分常见 [23, 59]，，但其对 FedAvg 收敛性的影响尚未得到充分理解。本文提出了一个新框架，用于分析 PL 目标函数下循环客户端参与时 FedAvg 的收敛性。我们的分析涵盖了不同的客户端本地过程，例如梯度下降、随机梯度下降和打乱式 SGD。通过该分析，我们可以了解 FedAvg 的收敛性如何受到系统不同特征的影响，例如本地更新过程、循环上线的客户端组内部及组间的数据异质性，以及周期长度K。我们发现，在某些条件下，循环客户端参与的收敛速度快于标准 FedAvg。我们还比较了不同客户端本地过程，以及作为本框架特例的算法，并分析了达到给定 ϵ 误差所需的代价。值得关注的未来工作包括将分析扩展至非 PL 目标函数和更一般的优化器（例如包含动量的优化器），以及客户端组并非互斥且在训练期间并非固定不变的情形。

参考文献

[1] Ahn, K., Yun, C., and Sra, S. 随机重排随机梯度下降：无需分量凸性且无需大量轮次要求的最优速率。神经信息处理系统进展，33:17526–17535，2020年。[2] Avdyukhin, D. and Kasiviswanathan, S. 任意通信模式下的联邦学习。载于第*38*届国际机器学习会议论文集，2021年。[3] Balle, B., Kairouz, P., McMahan, B., Thakkar, O., and Guha Thakurta, A. 通过随机签到实现隐私放大。神经信息处理系统进展，33:4623–4634，2020年。[4] Basu, D., Data, D., Karakus, C., and Diggavi, S. Qsparse-local-sgd：采用量化、稀疏化和本地计算的分布式随机梯度下降。载于神经信息处理系统进展，第14695–14706页，2019年。[5] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. 面向用户持有数据联邦学习的实用安全聚合。载于*NIPS*隐私多方机器学习研讨会，2016年。[6] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konecny, J., Mazzocchi, S., McMahan, H. B., Overveldt, T. V., Petrou, D., Ramage, D., and Roselander, J. 迈向大规模联邦学习：系统设计。*SysML*，2019年4月。网址：https://www.sysml.cc/doc/ 2019/193.pdf。[7] Cho, Y. J., Wang, J., and Joshi, G. 联邦学习中的客户端选择：收敛分析与选择权选择策略。*arXiv:2010.01243*，abs/2010.01243，2020年。网址：http://arxiv.org/ abs/2010.01243。[8] Choquette-Choo, C. A., McMahan, H. B., Rush, K., and Thakurta, A. 用于隐私机器学习的多轮矩阵分解机制。*arXiv*预印本 *arXiv:2211.06530*，2022年。[9] Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. EMNIST：扩展至手写字母的 MNIST。*arXiv*预印本 *arXiv:1702.05373*，2017 年。[10] Eichner, H., Koren, T., McMahan, B., Srebro, N., and Talwar, K. 半循环随机梯度下降。载于国际机器学习会议，第1764–1773页。PMLR，2019年。[11] Goetz, J., Malik, K., Bui, D., Moon, S., Liu, H., and Kumar, A. 主动联邦学习。*arXiv*，2019年。[12] Gower, R. M., Sebbouh, O., and Loizou, N. 结构化非凸函数的随机梯度下降：学习率、小批量训练与插值。国际人工智能与统计会议（*AISTATS*），2021年。[13] Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P. A. 增量梯度与增量牛顿方法的收敛速率。*SIAM*优化杂志，29(4):2542–2565，2019年。[14] Gürbüzbalaban, M., Ozdaglar, A., and Parrilo, P. A. 随机重排为何优于随机梯度下降。数学规划，186(1):49–84，2021年。[15] Haddadpour, F. and Mahdavi, M. 联邦学习中本地下降方法的收敛性。*arXiv*预印本 *arXiv:1910.14425*，2019年。[16] Haddadpour, F., Kamani, M. M., Mahdavi, M., and Cadambe, V. 周期平均的本地随机梯度下降：更紧致的分析与自适应同步。载于神经信息处理系统进展，第11080–11092页，2019年。

[17] Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. 面向移动键盘预测的联邦学习。*arXiv*预印本 *arXiv:1811.03604*，2018年。[18] Hsu, T.-M. H., Qi, H., and Brown, M. 衡量非同分布数据对联邦视觉分类的影响。载于*NeurIPS 2019*联合举办的用户隐私与数据机密性联邦学习国际研讨会（*FL-NeurIPS'19*），2019年12月。[19] Huba, D., Nguyen, J., Malik, K., Zhu, R., Rabbat, M., Yousefpour, A., Wu, C.-J., Zhan, H., Ustinov, P., Srinivas, H., 等。Papaya：实用、私密且可扩展的联邦学习。《机器学习与系统会议论文集》，4:814–832，2022年。[20] Jhunjhunwala, D., Sharma, P., Nagarkatti, A., and Joshi, G. Fedvarp：应对联邦学习中部分客户端参与造成的方差。$arXiv$，2022年。[21] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascon, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecny, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Ozgur, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramer, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. 联邦学习的进展与开放问题。*arXiv*预印本 *arXiv:1912.04977*，2019年。[22] Kairouz, P., McMahan, B., Song, S., Thakkar, O., Thakurta, A., and Xu, Z. 无需采样或洗牌的实用型隐私保护（深度）学习。载于国际机器学习会议，第5213–5225页。PMLR，2021年。[23] Kairouz, P., McMahan, B., Song, S., Thakkar, O., Thakurta, A., and Xu, Z. 无需采样或洗牌的实用型隐私保护（深度）学习。*arXiv*预印本 *arXiv:2103.00039*，2021年12月。[24] Karimi, H., Nutini, J., and Schmidt, M. 在Polyak-Łojasiewicz条件下梯度方法和近端梯度方法的线性收敛。$CoRR$，abs/1608.04636，2020年。URL http://arxiv.org/abs/ 1608.04636。[25] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD：设备端联邦学习的随机控制平均。*arXiv*预印本 *arXiv:1910.06378*，2019年。[26] Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. U. 变化拓扑与本地更新下去中心化随机梯度下降的统一理论。载于第*37*届国际机器学习会议论文集，2020年。[27] Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. 联邦学习：挑战、方法与未来方向。《*IEEE*信号处理杂志》，37(3):50–60，2020年。[28] Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. 关于FedAvg在非独立同分布数据上的收敛。载于国际学习表征会议（*ICLR*），2020年7月。URL https: //arxiv.org/abs/1907.02189。[29] Li, X., Milzarek, A., and Qiu, J. 随机重排在Kurdyka-{\L} ojasiewicz不等式下的收敛。*arXiv*预印本 *arXiv:2110.04926*，2021年。[30] Malinovsky, G., Sailanbayev, A., and Richtárik, P. 带方差缩减的随机重排：新分析与更优收敛率。*arXiv*预印本 *arXiv:2104.09342*，2021年。[31] McMahan, B. and Thakurta, A. 具有形式化差分隐私保障的联邦学习。$GoogleAI\ Blog$，2022年。

[32] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. 去中心化数据上的深度网络通信高效学习。国际人工智能与统计会议（*AISTATS*），2017年4月。URL https://arxiv.org/abs/1602.05629。[33] McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. 差分隐私循环语言模型的学习。国际学习表征会议，2018年。[34] Mishchenko, K., Khaled, A., and Richtárik, P. 随机重排：简明分析与大幅改进。神经信息处理系统进展，33:17309–17320，2020年。[35] Nagaraj, D., Jain, P., and Netrapalli, P. 无放回随机梯度下降：一般光滑凸函数的更优收敛率。载于国际机器学习会议，第4703–4711页。PMLR，2019年。[36] Nguyen, L. M., Tran-Dinh, Q., Phan, D. T., Nguyen, P. H., and Van Dijk, M. 洗牌型梯度方法的统一收敛分析。机器学习研究杂志，22(1):9397–9440，2021年。[37] Paulik, M., Seigel, M., Mason, H., Telaar, D., Kluivers, J., van Dalen, R., Lau, C. W., Carlson, L., Granqvist, F., Vandevelde, C., 等。面向设备端个性化的联邦评估与调优：系统设计与应用。*arXiv*预印本 *arXiv:2102.08503*，2021年。[38] Qu, Z., Lin, K., Kalagnanam, J., Li, Z., Zhou, J., and Zhou, Z. 联邦学习的优势：FedAvg实现线性加速。*https://arxiv.org/abs/2007.05690*，2020年。[39] Qu, Z., Lin, K., Kalagnanam, J., Li, Z., Zhou, J., and Zhou, Z. 过参数化非线性系统和神经网络的损失景观与优化。*https://arxiv.org/abs/2003.00307*，2021年。[40] Rajput, S., Gupta, A., and Papailiopoulos, D. 缩小无放回随机梯度下降的收敛差距。载于国际机器学习会议，第7964–7973页。PMLR，2020年。[41] Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečn`y, J., Kumar, S., and McMahan, H. B. 自适应联邦优化。载于国际学习表征会议（*ICLR*），2021年。[42] Ruan, Y., Zhang, X., Liang, S.-C., and Joe-Wong, C. 面向非独立同分布数据的联邦学习：迈向灵活的设备参与。*ArXiv*，2020年。[43] Safran, I. and Shamir, O. 带随机洗牌的随机梯度下降效果如何？载于学习理论会议，第3250–3284页。PMLR，2020年。[44] Safran, I. and Shamir, O. 在病态问题中，随机洗牌只有经过许多轮次才优于随机梯度下降。神经信息处理系统进展，34:15151–15161，2021年。[45] Sahu, A. K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., and Smith, V. 面向异构网络的联邦优化。载于第三届机器学习系统会议论文集，2020年1月。[46] Stich, S. U. 本地随机梯度下降收敛迅速且通信开销低。载于国际学习表征会议（*ICLR*），2019年。[47] Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. 应对异构联邦优化中的目标不一致问题。预印本，2020年5月。URL https://arxiv.org/abs/2007.07481。[48] Wang, J., Tantia, V., Ballas, N., and Rabbat, M. SlowMo：利用慢动量改进通信高效的分布式随机梯度下降。载于国际学习表征会议，2020年。URL https://openreview.net/forum?id=SkxJ8REYPH。[49] Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., 等。联邦优化指南。*arXiv*预印本 *arXiv:2107.06917*，2021年。

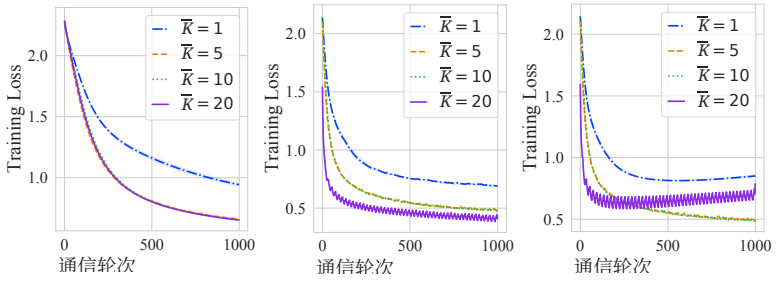
[50] Wang, J., Das, R., Joshi, G., Kale, S., Xu, Z., and Zhang, T. 异质数据下联邦平均的惊人有效性。*arXiv* 预印本 *arXiv:2206.04723*，2022年。[51] Wang, S. and Ji, M. 对任意客户端参与情况下联邦学习的统一分析。载于神经信息处理系统进展，2022年。[52] Woodworth, B., Patel, K. K., Stich, S. U., Dai, Z., Bullins, B., McMahan, H. B., Shamir, O., and Srebro, N. 本地随机梯度下降优于小批量随机梯度下降吗? 载于第*37*届国际机器学习会议论文集，2020年。[53] Woodworth, B. E., Patel, K. K., and Srebro, N. 异质分布式学习中的小批量随机梯度下降与本地随机梯度下降。神经信息处理系统进展，33：6281–6292，2020年。[54] Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist：一个用于机器学习算法基准测试的新型图像数据集。*https://arxiv.org/abs/1708.07747*，2017年8月。[55] Yang, H., Fang, M., and Liu, J. 在非独立同分布联邦学习中通过部分工作节点参与实现线性加速。载于国际学习表征会议，2021年。[56] Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., Ramage, D., and Beaufays, F. 联邦学习实践：改进 Google 键盘查询建议。*arXiv* 预印本 *arXiv:1812.02903*，2018年。[57] Yu, H., Jin, R., and Yang, S. 分布式非凸优化中通信高效动量随机梯度下降的线性加速分析。载于国际机器学习会议（*ICML*）论文集，2019年6月。[58] Yun, C., Rajput, S., and Sra, S. 洗牌情况下的小批量与本地随机梯度下降：紧致收敛界及其他结果。国际学习表征会议（*ICLR*），2022年。[59] Zhu, C., Xu, Z., Chen, M., Konečn`y, J., Hard, A., and Goldstein, T. 白昼还是夜晚? 基于周期性变化分布训练多分支网络的联邦学习。载于国际学习表征会议，2021年。[60] Zhu, F., Zhang, J., and Wang, X. 采用基于年龄的工作节点选择的通信高效本地随机梯度下降。*arXiv* 预印本 *arXiv:2210.17073*，2022年12月。

A 其他实验细节和结果

补充实验设置详情。所有实验均在配备一块 NVIDIA TitanX GPU 的集群上进行。算法使用 PyTorch 1.11. 0 实现。所有实验所用代码均包含在补充材料中。对于所有实验，我们对所需超参数进行网格搜索，以找出表现最佳的参数；随后固定这些超参数，仅改变 K。具体而言，我们对学习率：$\eta \in \{0.05, 0.01, 0.005, 0.001\}$、批量大小：$b \in \{32, 64, 128\}$和本地迭代次数： $\tau \in \{5, 10, 30, 50\}$ 进行网格搜索，为每个基准测试找出测试准确率最高的超参数。实验所用的深度多层感知机采用包含 2 个隐藏层的网络，每层有 $[64, 30]$ 个单元，并在第一个隐藏层后使用丢弃；输入为归一化展平图像，输出为标签空间。所有实验中，数据按 80%/10%/10% 的比例划分为训练集/验证集/测试集，其中训练数据随后再次异质地划分到各客户端。

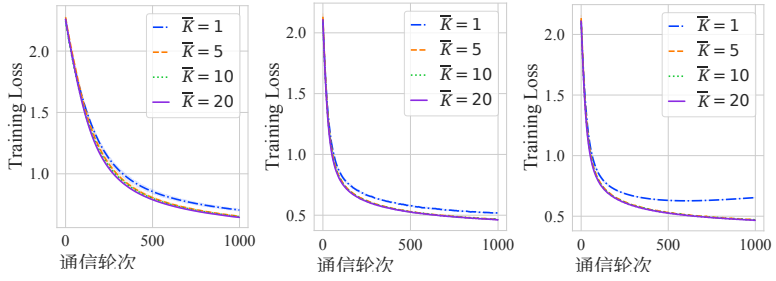
图 2 和图 3 中结果的训练损失。我们分别在图 5 和图 6 中展示了图 2 和图 3 所示测试准确率结果对应的训练损失曲线。可以看到，这些结果与我们在测试准确率中观察到的情况一致：较高的 $K > 1$ 会带来更快的收敛。与图 6 所示较低数据异质性下的改进幅度相比，图 5 所示较高数据异质性下的收敛改进幅度更大。此外，对于高数据异质性下的 SGD 和 SSGD 客户端本地过程（图 5(b)-(c)），在最高的 $K = 20$ 下，我们观察到了测试准确率曲线中也出现过的振荡。这是因为客户端循环参与，而客户端组的数据具有异质性。在数据异质性较低且 K 值较低时，则不会观察到这种振荡。

—



(a) 梯度下降　(b) 随机梯度下降　(c) SSGD

图5： 高数据异质性（$\alpha = 0.5$）下 FMNIST 的训练损失。



(a) 梯度下降（$\alpha = 2.0$）　(b) 随机梯度下降 ($\alpha = 2.0$)　(c) SSGD ($\alpha = 2.0$)

图6： 低数据异质性下 FMNIST 的训练损失 ($\alpha = 2.0$)。

B 常用不等式

引理 B.1 (杨氏不等式**)**。给定两个维数相同的向量$\mathbf{u}$、$\mathbf{v}$ $\in \mathbb{R}^d$，欧几里得内积可按如下方式进行界定：

$$\langle \mathbf{u}, \mathbf{v} \rangle \leq \frac{\|\mathbf{u}\|^2}{2\gamma} + \frac{\gamma \|\mathbf{v}\|^2}{2}$$

对任意常数 $\gamma > 0$ 均成立。

引理 B.2 （詹森不等式）。 给定凸函数 f 和随机变量 X，有以下结论成立。

$$f\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[f(X)\right].$$

引理 B.3 （平方和）。 对于正整数 K 以及向量集合 $\mathbf{x}_1$、...、$\mathbf{x}_K$，有以下结论成立：

$$\left\| \sum_{k=1}^{K} \mathbf{x}_k \right\|^2 \leq K \sum_{k=1}^{K} \|\mathbf{x}_k\|^2.$$

引理 B.4 （无放回抽样方差）。 对于正整数 K 和一组向量$\mathbf{x}_1$、...、$\mathbf{x}_K$，设这些向量的均值为 $\mathbf{x} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{x}_k$，并且从 $[K]$ 中均匀随机地进行不放回抽样，得到大小为 N 的小批量 $\mathcal{N}$，则有

$$\mathbb{E}\left[\left\| \frac{1}{N} \sum_{k \in \mathcal{N}} \mathbf{x}_k - \overline{\mathbf{x}} \right\|\right]^2 = \frac{(K-N)}{NK(K-1)} \sum_{k=1}^{K} \|\mathbf{x}_k - \overline{\mathbf{x}}\|^2 \tag{12}$$

C CyCP+局部梯度下降情形的证明

本节给出定理 1 的证明。为简化起见，证明安排如下：首先，在附录 C.1 中证明模型更新步骤是带噪梯度下降步骤。接着，在附录 C.2 中给出分析所需的一些中间结果，随后在附录 C.3 中给出定理 1 的证明。最后，在附录 C.4 中给出这些中间结果的证明。

我们将算法中截至周期轮次 k 的随机性所生成的 σ-代数定义如下： $\mathcal{F}_k \triangleq \sigma\left\{\{\mathbf{w}^{(1,i)}\}_{i=1}^K,\ \{\mathbf{w}^{(2,i)}\}_{i=1}^K,\ \ldots,\ \{\mathbf{w}^{(k-1,i)}\}_{i=1}^K\right\}$。我们用 $\mathbb{E}_k[\cdot]$ 作为期望 $\mathbb{E}[\cdot|\mathcal{F}_k]$ 的简写。

C.1 全局模型更新作为带噪梯度下降步骤

采用 CyCP 且每个客户端执行一次梯度下降更新时（算法 1 中的情形 (i)），全局模型的更新规则（$k \in [K]$, $i \in [K]$）为

$$\mathbf{w}^{(k,i)} - \mathbf{w}^{(k,i-1)} = -\eta \sum_{m \in \mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,i-1)}) \tag{13}$$

回忆一下，$\mathbf{w}^{(k+1,0)} = \mathbf{w}^{(k,K)}$。因此，我们可以展开 (13)，得到以下结果 .

引理 C.1。

$$\mathbb{E}[\mathbf{w}^{(k+1,0)}|\mathcal{F}_k] - \mathbf{w}^{(k,0)} = -\eta\overline{K}N\nabla F(\mathbf{w}^{(k,0)}) + \eta^2\mathbb{E}[\bar{\mathbf{r}}^{(k,0)}|\mathcal{F}_k], \tag{14}$$

其中 $\mathbf{r}^{(k,0)} \triangleq \sum_{i=1}^{K-1}\left(\prod_{j=i+2}^{K}(\mathbf{I}_d - \eta\mathbf{S}^{(k,j)})\right)\mathbf{S}^{(k,i+1)}\sum_{j'=1}^{i}\mathbf{q}^{(k,j')}$，其中

$$\left.\begin{array}{c}\overline{\mathbf{S}}^{(k,i)} := \sum_{m\in\mathcal{S}^{(k,i)}}\int_0^1 \nabla^2 F_m(\mathbf{w}^{(k,0)} + t(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}))\,dt \\ \overline{\mathbf{q}}^{(k,i)} := \sum_{m\in\mathcal{S}^{(k,i)}}\nabla F_m(\mathbf{w}^{(k,0)})\end{array}\right\},\ \textit{for all } i \in [\overline{K}].$$

证明。 注意，(13) 是正文中更新规则的重新缩放版本：更新被放大了 N，但我们改为将学习率设为 $\eta = \log(MK^2)/\mu NKK$，从而缩小步长。我们可以将局部目标函数的梯度改写为：

$$\begin{aligned}&\nabla F_m(\mathbf{w}^{(k,i-1)}) = \nabla F_m(\mathbf{w}^{(k,0)}) + \nabla F_m(\mathbf{w}^{(k,i-1)}) - \nabla F_m(\mathbf{w}^{(k,0)})\\ \Rightarrow\ &\nabla F_m(\mathbf{w}^{(k,i-1)}) - \nabla F_m(\mathbf{w}^{(k,0)}) = \underbrace{\int_0^1 \nabla^2 F_m(\mathbf{w}^{(k,0)} + t(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}))\,dt}_{:=\mathbf{H}_m^{(k,i)}}(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)})\end{aligned} \tag{15}$$

注意，由于我们假设局部目标函数 $F_m(\cdot)$ $m \in [M]$ 可微且为 L-光滑（见 [58] 中的附录 D.2），因此 $\mathbf{H}_m^{(k,i)}$ 存在，进而可以证明 $\|\mathbf{H}_m^{(k,i)}\| \le L,\ \ \forall m,\ k,\ i$。利用式 (15) 中的表达式可得

$$\mathbf{w}^{(k,i)} - \mathbf{w}^{(k,i-1)} = -\eta \sum_{m\in\mathcal{S}^{(k,i)}}\left(\nabla F_m(\mathbf{w}^{(k,0)}) + \mathbf{H}_m^{(k,i)}(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)})\right). \tag{16}$$

定义 $\mathbf{q}^{(k,i)} := \sum_{m\in\mathcal{S}^{(k,i)}}\nabla F_m(\mathbf{w}^{(k,0)})$ 和 $\mathbf{S}^{(k,i)} := \sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{H}_m^{(k,i)}$，我们有

$$\mathbf{w}^{(k,i)} - \mathbf{w}^{(k,i-1)} = -\eta\overline{\mathbf{q}}^{(k,i)} - \eta\overline{\mathbf{S}}^{(k,i)}(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)})$$

$$\Rightarrow \mathbf{w}^{(k,i)} - \mathbf{w}^{(k,0)} = \left(\mathbf{I} - \eta\overline{\mathbf{S}}^{(k,i)}\right)\left(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}\right) - \eta\overline{\mathbf{q}}^{(k,i)}. \tag{17}$$

展开（17）式，可得

$$\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)} = -\eta\sum_{i=1}^{\overline{K}}\overline{\mathbf{q}}^{(k,i)} + \eta^2\underbrace{\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d - \eta\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}}_{:=\overline{\mathbf{r}}^{(k,0)}} \tag{18}$$

给定 $\mathcal{F}_k$，式 (18) 中唯一的随机性来自随机客户端集合 $\{\mathcal{S}^{(k,i)}\}_i$。因此，

$$\mathbb{E}[\mathbf{w}^{(k+1,0)}|\mathcal{F}_k] - \mathbf{w}^{(k,0)} = -\eta\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\nabla F_m(\mathbf{w}^{(k,0)})\right] + \eta^2\mathbb{E}_k\left[\overline{\mathbf{r}}^{(k,0)}\right],$$

计算期望即可完成证明。 □

将 ηKN 视为有效学习率，则式 (14) 中的更新是一个带噪梯度下降步骤。证明的主要部分在于界定噪声项 $\mathbf{r}^{(k,0)}$。

C.2 中间结果

引理 C.2（客户端组梯度和的界）。若客户端函数 F_m 满足假设 *3*，则对于任意 $i \in [K]$ 和任意 $\mathbf{w}$

$$\left\|\sum_{j=1}^{i}\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j)}\nabla F_m(\mathbf{w})\right\| \leq i\alpha + i\|\nabla F(\mathbf{w})\|.$$

引理 C.3（CyCP 导致的误差项界）。在假设 *1*、*3* 下，(14)中的误差可界定为

$$\left\|\mathbb{E}_k[\overline{\mathbf{r}}^{(k,0)}]\right\| \leq \frac{3N^2L\overline{K}(\overline{K}-1)}{5}\left(\alpha + \left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right), \tag{19}$$

$$\mathbb{E}_k\left[\left\|\overline{\mathbf{r}}^{(k,0)}\right\|^2\right] \leq \frac{(\overline{K}-1)^2\overline{K}^2N^3L^2}{2}\left[\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 + 2N\left(\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \alpha^2\right)\right]. \tag{20}$$

观察式 (20) 中的界可知，除了式 (19) 中界所包含的项外，由于在每个组内对子客户端进行子采样，我们还得到了对组内异质性 γ 的额外依赖。

C.3 定理 1 的证明

为便于参考，我们在此重述定理 1。

定理（采用 CyCP 的全局模型收敛+局部梯度下降（算法 1 中的情形 (i)）。假设局部客户端函数 $\{F_m\}$ 满足假设 *1* 和假设 *3*，而全局损失函数满足

假设 *2*。若学习率满足 $\eta \leq \frac{1}{7LNK}$，则算法 *1* 在客户端执行一步本地梯度下降后生成的迭代序列满足

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \left(1 - \eta\overline{K}N\mu\right)^K \left(F(\mathbf{w}^{(0,0)}) - F^*\right) + \mathcal{O}\left(\frac{\eta^2 L^2 N^2 (\overline{K}-1)^2\alpha^2}{\mu}\right) + \mathcal{O}\left(\kappa\eta\overline{K}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2\right),$$

其中 $\kappa = L/\mu$ 为条件数。选择步长 $\eta = \log(MK^2)/\mu NKK$，其中 $K \geq 7\kappa\log(M\overline{K}^2)$，则收敛误差有如下界:

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{F(\mathbf{w}^{(0,0)}) - F^*}{MK^2} + \tilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu\overline{K}^2K^2}\right) + \tilde{\mathcal{O}}\left(\frac{\kappa\gamma^2}{\mu NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right)$$

~其中 $\mathcal{O}(\cdot)$ 包含对数项和数值常数。

推论 1. 将算法 *1* 中的通信轮次总数记为 $T = KK$，用T 表示，上述界可写为

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{\overline{K}^2(F(\mathbf{w}^{(0,0)}) - F^*)}{MT^2} + \tilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu T^2}\right) + \tilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\gamma^2}{\mu NT}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right).$$

其中，由于 $K \geq 7\kappa\log(MK^2)$，我们有 $T \geq 7\overline{\kappa K}\log(M\overline{K}^2T^2)$；由于 T 的下界取决于 K，因此若不相应增大T，就无法增大 K。

证明。利用全局目标函数 $F(\mathbf{w})$ 的 L-光滑性性质（假设 1），我们有

$$\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})] - F(\mathbf{w}^{(k,0)}) \leq \langle\nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)}\rangle + \frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)}\right\|^2\right] \quad (21)$$

此处的期望是对所有 $i \in [K]$ 的选定客户端集合 $\mathcal{S}^{(k,i)}$ 取的。首先，我们对式 (21)中的内积项进行界定。利用引理 C.1 和引理 C.3，可得

$$\begin{aligned}
&\langle\nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)}\rangle \\
&= \langle\nabla F(\mathbf{w}^{(k,0)}), -\eta\overline{K}N\nabla F(\mathbf{w}^{(k,0)}) + \eta^2\mathbb{E}_k[\overline{\mathbf{r}}^{(k,0)}]\rangle && \text{(Using Lemma C.1)} \\
&\leq -\eta\overline{K}N\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \eta^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\frac{3N^2L\overline{K}(\overline{K}-1)}{5}\left(\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| + \alpha\right) && \text{(Using Lemma C.3)} \\
&= -\eta\overline{K}N\left(1 - \frac{3\eta(\overline{K}-1)NL}{5}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{3\eta^2N^2L\overline{K}(\overline{K}-1)\alpha}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| && (22)
\end{aligned}$$

我们可以将 (22) 中的最后一项界定为

$$\begin{aligned}
\frac{3\eta^2N^2L\overline{K}(\overline{K}-1)\alpha}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| &= \left(\frac{\eta^{1/2}N^{1/2}\overline{K}^{1/2}}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)\left(3\eta^{3/2}N^{3/2}L\overline{K}^{1/2}(\overline{K}-1)\alpha\right) \\
&\leq \frac{\eta N\overline{K}}{50}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{9\eta^3\alpha^2L^2N^3\overline{K}(\overline{K}-1)^2}{2} \quad (\because \textit{Lemma B.1}) \quad (23)
\end{aligned}$$

将 (23) 代回 (22)，得到

$$\langle\nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)}\rangle \leq -\eta\overline{K}N\left(\frac{49}{50} - \frac{3\eta(\overline{K}-1)NL}{5}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{9\eta^3\alpha^2L^2N^3\overline{K}(\overline{K}-1)^2}{2}. \quad (24)$$

现在我们要对 (21) 右端的最后一项进行界定。我们有

$$\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)}-\mathbf{w}^{(k,0)}\right\|^2\right]\overset{(18)}{=}\frac{L}{2}\mathbb{E}_k\left[\left\|-\eta\sum_{i=1}^{\overline{K}}\overline{\mathbf{q}}^{(k,i)}+\eta^2\overline{\mathbf{r}}^{(k,0)}\right\|^2\right]$$
$$\leq L\eta^2\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{q}}^{(k,i)}\right\|^2\right]+L\eta^4\mathbb{E}_k\left[\left\|\overline{\mathbf{r}}^{(k,0)}\right\|^2\right] \tag{25}$$

最后一项已在引理 C.3 的式 (20) 中得到界定。现在，我们来界定第一项,

$$\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{q}}^{(k,i)}\right\|^2\right]=\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]=N^2\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]$$
$$=N^2\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\left\{\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\nabla F_m(\mathbf{w}^{(k,0)})-\frac{1}{M/\overline{K}}\sum_{m\in\sigma(i)}\nabla F_m(\mathbf{w}^{(k,0)})+\frac{1}{M/\overline{K}}\sum_{m\in\sigma(i)}\nabla F_m(\mathbf{w}^{(k,0)})\right\}\right\|^2\right]$$
$$=N^2\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\left\{\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\nabla F_m(\mathbf{w}^{(k,0)})-\frac{1}{M/\overline{K}}\sum_{m\in\sigma(i)}\nabla F_m(\mathbf{w}^{(k,0)})\right\}\right\|^2\right]+N^2\left\|\frac{1}{M/\overline{K}}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2$$
$$\leq N\overline{K}^2\left[\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+N\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\right]. \tag{26}$$

按照与式 (37) 的证明类似的步骤。最后，将式 (24)、(25) 和 (26) 代入式 (21)，得到$\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})]-F(\mathbf{w}^{(k,0)})$

$$\leq-\eta\overline{K}N\left(\frac{49}{50}-\frac{3\eta(\overline{K}-1)NL}{5}-L\eta N\overline{K}-\eta^3\overline{K}(\overline{K}-1)^2N^3L^3\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{9\eta^3\alpha^2L^2N^3\overline{K}(\overline{K}-1)^2}{2}$$
$$+\eta^4\overline{K}^3(\overline{K}-1)N^4L^3\alpha^2+\frac{1}{2}\eta^4\overline{K}^2(\overline{K}-1)^2N^3L^3\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+L\eta^2N\overline{K}^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2$$
$$\leq-\frac{\eta\overline{K}N}{2}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{47\eta^3L^2N^3\overline{K}(\overline{K}-1)^2}{10}\alpha^2+\frac{11L\eta^2N\overline{K}^2}{5}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 \tag{27}$$

$$\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})]-F^*\leq(1-\eta\overline{K}N\mu)(F(\mathbf{w}^{(k,0)})-F^*)+\frac{47\eta^3L^2N^3\overline{K}(\overline{K}-1)^2}{10}\alpha^2$$
$$+\frac{11L\eta^2N\overline{K}^2}{5}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 \tag{28}$$

其中，最后一个不等式由假设 2 得出。展开式 (28)，并利用 $\eta=\frac{\log(MK^2)}{\mu KKN}$，可得

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})]-F^*$$
$$\leq\left(1-\frac{\log(MK^2)}{K}\right)^K(F(\mathbf{w}^{(0,0)})-F^*)+\frac{47\eta^3L^2N^3\overline{K}(\overline{K}-1)^2}{10\eta\overline{K}N\mu}\alpha^2+\frac{11L\eta^2N\overline{K}^2}{5\eta\overline{K}N\mu}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2$$
$$\leq\frac{F(\mathbf{w}^{(0,0)})-F^*}{MK^2}+\frac{47\eta^2L^2N^2(\overline{K}-1)^2}{10\mu}\alpha^2+\frac{11L\eta\overline{K}}{5\mu}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2$$
$$=\frac{F(\mathbf{w}^{(0,0)})-F^*}{MK^2}+\frac{47\log^2(MK^2)\kappa^2(\overline{K}-1)^2}{10\mu\overline{K}^2K^2}\alpha^2+\frac{11L\log(MK^2)}{5\mu^2NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2$$

$$= \frac{F(\mathbf{w}^{(0,0)}) - F^*}{MK^2} + \tilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu\overline{K}^2K^2}\right) + \tilde{\mathcal{O}}\left(\frac{\kappa}{\mu NK}\left(\frac{M-N\overline{K}}{M-\overline{K}}\right)\gamma^2\right).$$

此外，由于我们总共有 $T = K\overline{K}$ 个通信轮次，也可以将界表示为关于 T 的形式，如下所示。

$$\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq \frac{\overline{K}^2(F(\mathbf{w}^{(0,0)}) - F^*)}{MT^2} + \tilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu T^2}\right) + \tilde{\mathcal{O}}\left(\frac{\overline{K}\kappa}{\mu NT}\left(\frac{M-N\overline{K}}{M-\overline{K}}\right)\gamma^2\right).$$

□

推论 1 的证明 我们重述式 (3)，达到 ϵ 误差所需的总成本为 C_{GD}：

$$C_{\mathrm{GD}}(\epsilon) = \frac{c_{\mathrm{GD}}\overline{K}\gamma^2}{\epsilon N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right).$$

为了得到 $C_{\overline{\mathrm{GD}}|K>1}(\epsilon) < C_{\overline{\mathrm{GD}}|K=1}(\epsilon)$，不等式

$$\overline{K}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) - \left(\frac{M-N}{M-1}\right) < 0$$

对 $K > 1$ 应成立。重新整理各项，得到

$$N(M-1)\overline{K}^2 + (N-M^2)\overline{K} + M^2 - MN > 0 \tag{29}$$

由于 $(M(M-2N)+N)^2 = (N-M^2)^2 - 4N(M^2-MN)(M-1) \geq 0$，因此对于 $K > \overline{M}(M-N)/N(M-1)$，有 $K\left(\frac{M/K-N}{M/K-1}\right) > \left(\frac{M-N}{M-1}\right)$。稍作整理，得到

$$M < N + N\overline{K}\left(1 - \frac{1}{M}\right).$$

回忆一下，$\overline{K} \in [\frac{M}{N}]$。对于 $K = M/N$，得到

$$N + N\frac{M}{N}\left(1 - \frac{1}{M}\right) = M - 1 + N > M$$

对于任意 $N > 1$。然而，如果 $\overline{K} = \frac{M}{N} = 1$，则得到

$$N + N\left(\frac{M}{N} - 1\right)\left(1 - \frac{1}{M}\right) = M - 1 + N - N + \frac{N}{M} = M - 1 + \frac{N}{M} < M$$

对于任意 $N > 1$。因此，只有在 $K = M/N$ 时，$K > 1$ 才比 $K = 1$ 更有优势，这意味着每个客户端组都实现了全客户端参与。

C.4 中间引理的证明

引理 *C.2* 的证明。

$$\left\|\sum_{j=1}^{i}\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j)}\nabla F_m(\mathbf{w})\right\| \leq \sum_{j=1}^{i}\left\|\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j)}\nabla F_m(\mathbf{w}) - \nabla F(\mathbf{w}) + \nabla F(\mathbf{w})\right\|$$

$$\leq \sum_{j=1}^{i} \left\| \frac{1}{M/\overline{K}} \sum_{m\in\sigma(j)} \nabla F_m(\mathbf{w}) - \nabla F(\mathbf{w}) \right\| + i\|\nabla F(\mathbf{w})\| \leq i\alpha + i\|\nabla F(\mathbf{w})\|. \quad \text{(using Assumption 3)}$$

□

引理 *C.3* 的证明。首先，我们对其进行界定 $\left\|\mathbb{E}_k[\mathbf{r}^{(k,0)}]\right\|$。

$$\begin{aligned}
\left\|\mathbb{E}_k[\overline{\mathbf{r}}^{(k,0)}]\right\| &= \left\| \sum_{i=1}^{\overline{K}-1} \mathbb{E}_k \left[\left(\prod_{j=i+2}^{\overline{K}} (\mathbf{I}_d - \eta \overline{\mathbf{S}}^{(k,j)}) \right) \overline{\mathbf{S}}^{(k,i+1)} \sum_{j'=1}^{i} \overline{\mathbf{q}}^{(k,j')} \right] \right\| \\
&\leq \sum_{i=1}^{\overline{K}-1} \left\| \mathbb{E}_k \left[\left(\prod_{j=i+2}^{\overline{K}} (\mathbf{I}_d - \eta \overline{\mathbf{S}}^{(k,j)}) \right) \overline{\mathbf{S}}^{(k,i+1)} \sum_{j'=1}^{i} \overline{\mathbf{q}}^{(k,j')} \right] \right\| \\
&\leq \sum_{i=1}^{\overline{K}-1} \mathbb{E}_k \left[\underbrace{\left\| \prod_{j=i+2}^{\overline{K}} (\mathbf{I}_d - \eta \overline{\mathbf{S}}^{(k,j)}) \right\|}_{A_1} \underbrace{\left\| \overline{\mathbf{S}}^{(k,i+1)} \right\|}_{A_2} \underbrace{\left\| \sum_{j'=1}^{i} \overline{\mathbf{q}}^{(k,j')} \right\|}_{A3} \right] \qquad (\because \text{范数的次乘性})
\end{aligned} \tag{30}$$

接下来，我们分别对 A_1，A_2，A_3 进行界定，如下所示。

$$\begin{aligned}
A_1 = \left\| \prod_{j=i+2}^{\overline{K}} (\mathbf{I}_d - \eta \overline{\mathbf{S}}^{(k,j)}) \right\| &\leq \prod_{j=i+2}^{\overline{K}} \left\| (\mathbf{I}_d - \eta \overline{\mathbf{S}}^{(k,j)}) \right\| \\
&= \prod_{j=i+2}^{\overline{K}} \left\| I_d - \eta \sum_{m\in\mathcal{S}^{(k,j)}} \mathbf{H}_m^{(k,j)} \right\| \leq (1+\eta NL)^{\overline{K}} \leq e^{1/7} \leq 6/5,
\end{aligned} \tag{31}$$

其中，在 (31) 中我们使用了假设 1，且 $\eta \leq \frac{1}{7LNK}$。接下来，我们对 A_2进行界定。

$$A_2 = \left\| \overline{\mathbf{S}}^{(k,i+1)} \right\| = \left\| \sum_{m\in\mathcal{S}^{(k,i+1)}} \mathbf{H}_m^{(k,i+1)} \right\| \leq NL \tag{32}$$

接下来，我们对 A_3 进行如下界定:

$$\begin{aligned}
A_3 = \left\| \sum_{j'=1}^{i} \mathbb{E}_k \left[\overline{\mathbf{q}}^{(k,j')} \right] \right\| = \left\| \sum_{j'=1}^{i} \mathbb{E}_k \left[\sum_{m\in\mathcal{S}^{(k,j')}} \nabla F_m(\mathbf{w}^{(k,0)}) \right] \right\| &= N \left\| \sum_{j'=1}^{i} \frac{1}{M/\overline{K}} \sum_{m\in\sigma(j')} \nabla F_m(\mathbf{w}^{(k,0)}) \right\| \\
&\leq iN\alpha + iN\|\nabla F(\mathbf{w})\|.
\end{aligned} \tag{33}$$

最后一个不等式由引理 C.2 得出。将 (31)-(33) 中的界代入 (30)，即可得到 (19) 中的界。

接下来，我们推导式 (20) 中的界。

$$\mathbb{E}_k \left[\left\| \overline{\mathbf{r}}^{(k,0)} \right\|^2 \right] = \mathbb{E}_k \left[\left\| \sum_{i=1}^{\overline{K}-1} \left(\prod_{j=i+2}^{\overline{K}} (\mathbf{I}_d - \eta \overline{\mathbf{S}}^{(k,j)}) \right) \overline{\mathbf{S}}^{(k,i+1)} \sum_{j'=1}^{i} \overline{\mathbf{q}}^{(k,j')} \right\|^2 \right]$$

$$\leq (\overline{K}-1)\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]$$

$$\leq (\overline{K}-1)\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right\|^2\left\|\overline{\mathbf{S}}^{(k,i+1)}\right\|^2\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]. \tag{34}$$

注意,

$$\left\|(\mathbf{I}_d-\eta\overline{\mathbf{S}}^{(k,j)})\right\|\leq 1+\eta\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{H}_m^{(k,i)}\right\|\leq 1+\eta\sum_{m\in\mathcal{S}^{(k,i)}}\left\|\mathbf{H}_m^{(k,i)}\right\|\leq 1+\eta NL,\ \forall j\in[\overline{K}]. \tag{35}$$

将 (31)、(32) 和 (35) 代入 (34)，得到

$$\mathbb{E}_k\left[\left\|\overline{\mathbf{r}}^{(k,0)}\right\|^2\right]\leq\frac{36(\overline{K}-1)N^2L^2}{25}\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right] \tag{36}$$

最后，我们对式（36）中的期望值进行界定:

$$\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]=N^2\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\sum_{j'=1}^{i}\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,j')}}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]$$

$$=N^2\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left\|\sum_{j'=1}^{i}\left\{\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,j')}}\nabla F_m(\mathbf{w}^{(k,0)})-\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})+\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\}\right\|^2$$

$$=N^2\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\sum_{j'=1}^{i}\left\{\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,j')}}\nabla F_m(\mathbf{w}^{(k,0)})-\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\}\right\|^2\right]$$

$$+N^2\sum_{i=1}^{\overline{K}-1}\left\|\frac{1}{M/\overline{K}}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2 \qquad \text{(Cross-terms are zero)}$$

$$\leq N^2\sum_{i=1}^{\overline{K}-1}i\sum_{j'=1}^{i}\mathbb{E}_k\left[\left\|\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,j')}}\nabla F_m(\mathbf{w}^{(k,0)})-\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right] \qquad \text{(Using Lemma B.3)}$$

$$+N^2\sum_{i=1}^{\overline{K}-1}\left\|\frac{1}{M/\overline{K}}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2$$

$$=N^2\sum_{i=1}^{\overline{K}-1}i\sum_{j'=1}^{i}\frac{1}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\frac{1}{M/\overline{K}}\sum_{m'\in\sigma(j')}\left\|\nabla F_{m'}(\mathbf{w})-\frac{1}{M/\overline{K}}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w})\right\|^2$$

(Using without replacement sampling)

$$+N^2\sum_{i=1}^{\overline{K}-1}\left\|\frac{1}{M/\overline{K}}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2$$

$$\leq N^2 \sum_{i=1}^{\overline{K}-1} i \sum_{j'=1}^{i} \frac{1}{N} \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \gamma^2 + N^2 \sum_{i=1}^{\overline{K}-1} \left\| \frac{1}{M/\overline{K}} \sum_{j'=1}^{i} \sum_{m \in \sigma(j')} \nabla F_m(\mathbf{w}^{(k,0)}) \right\|^2 \quad \text{(Using Assumption 3)}$$

$$\leq \sum_{i=1}^{\overline{K}-1} i^2 N \left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \gamma^2 + N^2 \sum_{i=1}^{\overline{K}-1} 2i^2 \left(\left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \alpha^2 \right) \quad \text{(using Lemma C.2)}$$

$$\leq \frac{1}{3}(\overline{K} - 1)\overline{K}^2 N \left[\left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \gamma^2 + 2N \left(\left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \alpha^2 \right) \right]. \tag{37}$$

将式（37）代入式（36），可得

$$\mathbb{E}_k \left[\left\| \overline{\mathbf{r}}^{(k,0)} \right\|^2 \right] \leq \frac{12(\overline{K} - 1)^2 \overline{K}^2 N^3 L^2}{25} \left[\left(\frac{M/\overline{K} - N}{M/\overline{K} - 1} \right) \gamma^2 + 2N \left(\left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|^2 + \alpha^2 \right) \right].$$

至此，证明完成。 □

D CyCP+本地随机梯度下降情形的证明

本节给出定理 2 的证明。为简洁起见，证明安排如下：首先，在附录 D.1 中，我们证明模型更新步骤是带噪梯度下降步骤。接着，在附录 D.2 中，我们给出分析中将用到的一些中间结果，然后在附录 D.3 中证明定理 2。最后，在附录 D.5 中，我们给出这些中间结果的证明。

类似于我们在附录 C.3 中所做的，我们将截至周期轮次 k 时由算法中的随机性生成的 σ-代数定义如下：$\mathcal{F}_k \triangleq \sigma\left\{\{\mathbf{w}^{(1,i)}\}_{i=1}^K,\ \{\mathbf{w}^{(2,i)}\}_{i=1}^K,\ \ldots\ldots,\ \{\mathbf{w}^{(k-1,i)}\}_{i=1}^K\right\}$。我们使用 $\mathbb{E}_k[\cdot]$ 作为期望 $\mathbb{E}[\cdot|\mathcal{F}_k]$的简写。

D.1 将全局模型更新视为带噪声的梯度下降步骤

回顾全局模型更新规则为

$$\mathbf{w}^{(k,i)} - \mathbf{w}^{(k,i-1)} = -\eta \sum_{m\in\mathcal{S}^{(k,i)}} \sum_{l=0}^{\tau-1} \nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)}) \tag{38}$$

其中，$\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\ \xi_m^{(k,i-1,l)}) = \frac{1}{b}\sum_{\xi\in\xi_m^{(k,i-1,l)}} \nabla f(\mathbf{w}_m^{(k,i-1,l)},\ \xi)$ 是使用从客户端 m的本地数据集 $\mathcal{B}_m$ 中随机采样的、大小为 b 的小批量 $\xi_m^{(k,i-1,l)}$计算得到的随机梯度。回顾一下，k 是一个周期轮次索引，表示每个通信轮次；在该通信轮次中，我们通过从每个客户端组 σ中均匀地随机进行不放回抽样 N 个客户端 i，遍历了 K 个客户端组。索引 $i\in[K]$ 表示每个内部通信轮次。回顾一下，对于所有 $k\in[K]$，有 $\mathbf{w}^{(k+1,0)} = \mathbf{w}^{(k,K)}$。结合（38）可得

$$\begin{aligned}
\mathbf{w}^{(k,i)} - \mathbf{w}^{(k,i-1)} &= -\eta \sum_{m\in\mathcal{S}^{(k,i)}} \sum_{l=0}^{\tau-1} \Big(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)}) - \nabla F_m(\mathbf{w}^{(k,i-1)}) + \nabla F_m(\mathbf{w}^{(k,i-1)})\Big) \\
&= -\eta \sum_{m\in\mathcal{S}^{(k,i)}} \underbrace{\sum_{l=0}^{\tau-1} (\nabla F_m(\mathbf{w}_m^{(k,i-1,l)}, \xi_m^{(k,i-1,l)}) - \nabla F_m(\mathbf{w}^{(k,i-1)}))}_{:=\mathbf{d}_m^{(k,i)}} - \eta\tau \sum_{m\in\mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,i-1)}) \qquad (39) \\
&= -\eta \underbrace{\sum_{m\in\mathcal{S}^{(k,i)}} \mathbf{d}_m^{(k,i)}}_{:=\overline{\mathbf{d}}^{(k,i)}} - \eta\tau \sum_{m\in\mathcal{S}^{(k,i)}} \Big(\nabla F_m(\mathbf{w}^{(k,0)}) + \mathbf{H}_m^{(k,i)}(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}\Big) \qquad (\mathbf{H}_m^{(k,i)} \text{ defined in (15)}) \\
&= -\widetilde{\eta}\underbrace{(\overline{\mathbf{d}}^{(k,i)}/\tau + \overline{\mathbf{q}}^{(k,i)})}_{:=\widetilde{\mathbf{q}}^{(k,i)}} - \widetilde{\eta} \underbrace{\sum_{m\in\mathcal{S}^{(k,i)}} \mathbf{H}_m^{(k,i)}}_{:=\overline{\mathbf{S}}^{(k,i)}} (\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}) \qquad (40)
\end{aligned}$$

其中 $\widetilde{\eta} = \tau\eta$ 和 $\mathbf{q}^{(k,i)} := \sum_{m\in\mathcal{S}^{(k,i)}} \nabla F_m(\mathbf{w}^{(k,0)})$。展开式 (40)，可得

$$\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)} = -\widetilde{\eta}\sum_{i=1}^{K} \widetilde{\mathbf{q}}^{(k,i)} + \widetilde{\eta}^2 \underbrace{\sum_{i=1}^{K-1} \left(\prod_{j=i+2}^{K} (\mathbf{I}_d - \widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right) \overline{\mathbf{S}}^{(k,i+1)} \sum_{j'=1}^{i} \widetilde{\mathbf{q}}^{(k,j')}}_{:=\widetilde{\mathbf{r}}^{(k,0)}} \tag{41}$$

以 $\mathcal{F}_k$ 为条件，可得

$$\begin{aligned}\mathbb{E}_k[\mathbf{w}^{(k+1,0)}]-\mathbf{w}^{(k,0)} &= -\frac{\widetilde{\eta}}{\tau}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\overline{\mathbf{d}}^{(k,i)}\right]-\widetilde{\eta}\mathbb{E}_k\left[\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\nabla F_m(\mathbf{w}^{(k,0)})\right]+\widetilde{\eta}^2\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]\\ &= -\frac{\widetilde{\eta}}{\tau}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\overline{\mathbf{d}}^{(k,i)}\right]-\widetilde{\eta}\overline{K}N\nabla F(\mathbf{w}^{(k,0)})+\widetilde{\eta}^2\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]\\ &= -\frac{\widetilde{\eta}}{\tau}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right]-\widetilde{\eta}\overline{K}N\nabla F(\mathbf{w}^{(k,0)})+\widetilde{\eta}^2\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]. \qquad (42)\end{aligned}$$

D.2 中间结果

引理 D.1（本地 SGD 步骤所导致的误差项范数上界）。 我们可以如下界定本地随机梯度下降步骤所导致的误差项（$\mathbf{d}_m^{(k,i)}$在(39)中定义）的范数：

$$\left\|\sum_{i=1}^{\overline{K}}\mathbb{E}\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right]\right\|\leq\frac{3\eta LN\overline{K}(\tau-1)\tau}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{14\eta LN\overline{K}(\tau-1)\tau\nu}{25}+\frac{\eta LN(\overline{K}-1)(\tau-1)\tau\alpha}{36}.$$

引理 D.2（本地随机梯度下降步骤所导致的误差项范数平方上界）。 与引理 *D.1* 中的上界类似，我们可以如下界定本地随机梯度下降步骤所导致的误差项的平方：

$$\begin{aligned}\frac{3L\widetilde{\eta}^2}{2\tau^2}\mathbb{E}\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{d}}^{(k,i)}\right\|^2\right] &\leq \frac{\widetilde{\eta}N\overline{K}}{200}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{81L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{50\tau}+\frac{209\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{50}\\ &\quad+\frac{L^3\widetilde{\eta}^4\overline{K}^4N^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{9L^3\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4\alpha^2}{100}.\end{aligned}$$

引理 D.1 和引理 D.2 均对因采取本地随机梯度下降步骤而非完整梯度下降步骤所产生的误差项给出了界。该上界主要取决于本地步骤的数量 τ 以及组内和组间数据异质性 γ和 α。回想一下， $\nu=\gamma+\alpha$ 和 $\widetilde{\eta}=\eta\tau$。

引理 D.3（CyCP 所导致的误差项范数上界）。

$$\widetilde{\eta}^2\left\|\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]\right\|\leq\frac{17\widetilde{\eta}N\overline{K}}{250}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{84\widetilde{\eta}^2\eta N^2L^2\overline{K}(\overline{K}-1)(\tau-1)\nu}{125}+\frac{16\widetilde{\eta}^2N^2L\overline{K}(\overline{K}-1)\alpha}{25}.$$

引理 D.4（CyCP 所导致的误差项范数平方上界）。

$$\begin{aligned}\frac{3L\widetilde{\eta}^4}{2}\mathbb{E}\left[\left\|\widetilde{\mathbf{r}}^{(k,0)}\right\|^2\right] &\leq \frac{9\widetilde{\eta}N\overline{K}}{500}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{3\widetilde{\eta}^2L\overline{K}^2N\sigma^2}{200\tau}+\frac{173\widetilde{\eta}^4\overline{K}^4N^3L^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\\ &\quad+\frac{3\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{50}+\frac{59\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4L^3\alpha^2}{10}.\end{aligned}$$

引理 D.3 和引理 D.4 都对循环客户端参与所产生的误差项进行了界定；在证明概要中，我们将其建模为一个较大的全梯度步骤加上这些误差项。平方范数误差取决于由随机梯度产生的额外方差项，该方差项依赖于 σ^2。

引理 D.5 （每个周期轮次的初始全局模型与其在该周期轮次内轨迹之间的距离界）。

$$\begin{aligned}
&\sum_{r=1}^{\overline{K}} \mathbb{E}\left[\left\|\mathbf{w}^{(k,r-1)} - \mathbf{w}^{(k,0)}\right\|^2\right] \\
&\leq \frac{83\widetilde{\eta}^2 N^2 \overline{K}^2(\overline{K}-1)}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 + \frac{51\widetilde{\eta}^2 N\overline{K}^2(\overline{K}-1)\sigma^2}{50\tau} + \frac{3\eta^2\overline{K}\tau(\tau-1)\nu^2}{100} \\
&+ \frac{51\widetilde{\eta}^2 N\overline{K}^2(\overline{K}-1)\gamma^2}{25}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right) + \frac{41\widetilde{\eta}^2 N^2\overline{K}^2(\overline{K}-1)\alpha^2}{10}.
\end{aligned}$$

引理 D.5 界定了每个周期轮次开始时的初始全局模型与其在该周期轮次内的 K 个通信轮次中的轨迹之间的距离。

D.3 定理2的证明

定理（CyCP 下的收敛+随机梯度下降）。在假设*1*、*2*、*3*和*4*成立且步长为 $\eta = \log(MK^2)/\tau\mu NKK$ 时，对于满足 $\kappa = L/\mu$ 的 $K \geq 10\kappa\log(MK^2)$ 个通信轮次，收敛误差有如下界：

$$\begin{aligned}
\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq{}& \frac{F(\mathbf{w}^{(0,0)}) - F^*}{MK^2} + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)\alpha^2}{\mu\overline{K}K^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa\gamma^2}{\mu NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right) \\
&+ \widetilde{\mathcal{O}}\left(\frac{\kappa\sigma^2}{\mu\tau NK}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\tau-1)\nu^2}{\mu\tau N^2\overline{K}^2K^2}\right)
\end{aligned}$$

其中 $\mathcal{O}(\cdot)$ 涵盖所有对数项和常数。根据 $T = KK$，用通信轮次表示可得

$$\begin{aligned}
\mathbb{E}[F(\mathbf{w}^{(K,0)})] - F^* \leq{}& \frac{\overline{K}^2(F(\mathbf{w}^{(0,0)}) - F^*)}{MT^2} + \widetilde{\mathcal{O}}\left(\frac{\kappa^2\overline{K}(\overline{K}-1)\alpha^2}{\mu T^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\gamma^2}{\mu NT}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right) \\
&+ \widetilde{\mathcal{O}}\left(\frac{\overline{K}\kappa\sigma^2}{\mu\tau NT}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\tau-1)\nu^2}{\mu\tau N^2T^2}\right)
\end{aligned}$$

在整个证明中，我们使用 $\eta = \log(MK^2)/\mu\tau NKK$ 和 $\overline{K} \geq 10\kappa\log(MK^2)$，由此得到 $\eta \leq 1/(10\tau NLK)$。同样，利用全局目标函数的 L-光滑性，可得

$$\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})] - F(\mathbf{w}^{(k,0)}) \leq \langle\nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)}\rangle + \frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)} - \mathbf{w}^{(k,0)}\right\|^2\right] \quad (43)$$

这里的期望是对选定的客户端集合 $\mathcal{S}^{(k,i)},\ i \in [K]$ 和随机梯度取的。首先，我们对右端 (43) 中的内积项进行界定。利用 (42)，

$$\begin{aligned}
&\langle\nabla F(\mathbf{w}^{(k,0)}), \mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)}\rangle \\
&= \langle\nabla F(\mathbf{w}^{(k,0)}), -\frac{\widetilde{\eta}}{\tau}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right] - \widetilde{\eta}\overline{K}N\nabla F(\mathbf{w}^{(k,0)}) + \widetilde{\eta}^2\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]\rangle \\
&\leq \frac{\widetilde{\eta}}{\tau}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\left\|\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right]\right\| - \widetilde{\eta}\overline{K}N\left\|\nabla F(\mathbf{w}^{(k,0)}\right\|^2 + \widetilde{\eta}^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\left\|\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]\right\|. \quad (44)
\end{aligned}$$

$\left\|\sum_{i=1}^{K}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right]\right\|$ 和 $\|\mathbb{E}_k[\tilde{\mathbf{r}}^{(k,0)}]\|$ 已分别在引理 D.1 和引理 D.3 中得到界定。将这些界代入 (44)，可得

$$
\begin{aligned}
&\langle\nabla F(\mathbf{w}^{(k,0)}),\mathbb{E}_k[\mathbf{w}^{(k+1,0)}]-\mathbf{w}^{(k,0)}\rangle\\
&\le\frac{\widetilde{\eta}}{\tau}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\left(\frac{3\eta LN\overline{K}(\tau-1)\tau}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{14\eta LN\overline{K}(\tau-1)\tau\nu}{25}+\frac{\eta LN(\overline{K}-1)(\tau-1)\tau\alpha}{36}\right)\\
&\quad-\widetilde{\eta}\overline{K}N\left\|\nabla F(\mathbf{w}^{(k,0)}\right\|^2+\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\left(\frac{17\widetilde{\eta}N\overline{K}}{250}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{84\widetilde{\eta}^2\eta N^2L^2\overline{K}(\overline{K}-1)(\tau-1)\nu}{125}\right.\\
&\qquad\left.+\frac{16\widetilde{\eta}^2N^2L\overline{K}(\overline{K}-1)\alpha}{25}\right)\\
&\le-\widetilde{\eta}\overline{K}N\left(1-\frac{3}{50}-\frac{17}{250}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\left(\frac{\widetilde{\eta}^{1/2}N^{1/2}\overline{K}^{1/2}}{25}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)(14\widetilde{\eta}^{1/2}\eta LN^{1/2}\overline{K}^{1/2}(\tau-1)\nu)\\
&\quad+\left(\frac{\widetilde{\eta}^{1/2}N^{1/2}\overline{K}^{1/2}}{40}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)(28\widetilde{\eta}^{3/2}\eta L^2N^{3/2}\overline{K}^{3/2}(\tau-1)\nu)\\
&\quad+\left(\frac{\widetilde{\eta}^{1/2}N^{1/2}\overline{K}^{1/2}}{36}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)\left(\frac{\widetilde{\eta}^{1/2}\eta LN^{1/2}(\overline{K}-1)(\tau-1)\alpha}{\overline{K}^{1/2}}\right)\\
&\quad+\left(\frac{\widetilde{\eta}^{1/2}N^{1/2}\overline{K}^{1/2}}{25}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)\left(16\widetilde{\eta}^{3/2}LN^{3/2}\overline{K}^{1/2}(\overline{K}-1)\alpha\right)\qquad(\because\ \eta\le1/(10\tau NL\overline{K}))\\
&\le-\widetilde{\eta}\overline{K}N\left(1-\frac{3}{50}-\frac{17}{250}-\frac{1}{2\times40^2}-\frac{1}{2\times36^2}-\frac{1}{2\times25^2}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+98\widetilde{\eta}\eta^2L^2N\overline{K}(\tau-1)^2\nu^2\\
&\quad+392\widetilde{\eta}^3\eta^2L^4N^3\overline{K}^3(\tau-1)^2\nu^2+\frac{\widetilde{\eta}\eta^2L^2N(\overline{K}-1)^2(\tau-1)^2\alpha^2}{2\overline{K}}+128\widetilde{\eta}^3L^2N^3\overline{K}(\overline{K}-1)^2\alpha^2\qquad(\because\ Lemma\ B.1)\\
&\le-\frac{107\widetilde{\eta}\overline{K}N}{125}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+102\widetilde{\eta}\eta^2L^2N\overline{K}(\tau-1)^2\nu^2+\frac{\widetilde{\eta}\eta^2L^2N(\overline{K}-1)^2(\tau-1)^2\alpha^2}{2\overline{K}}+128\widetilde{\eta}^3L^2N^3\overline{K}(\overline{K}-1)^2\alpha^2.
\end{aligned}
\tag{45}
$$

现在我们如下界定（43）式右端的第二项

$$
\begin{aligned}
\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)}-\mathbf{w}^{(k,0)}\right\|^2\right]&=\frac{L}{2}\mathbb{E}_k\left[\left\|-\widetilde{\eta}\sum_{i=1}^{\overline{K}}\widetilde{\mathbf{q}}^{(k,i)}+\widetilde{\eta}^2\widetilde{\mathbf{r}}^{(k,0)}\right\|^2\right]\\
&=\frac{L}{2}\mathbb{E}_k\left[\left\|-\widetilde{\eta}\sum_{i=1}^{\overline{K}}\overline{\mathbf{d}}^{(k,i)}/\tau-\widetilde{\eta}\sum_{i=1}^{\overline{K}}\overline{\mathbf{q}}^{(k,i)}+\widetilde{\eta}^2\widetilde{\mathbf{r}}^{(k,0)}\right\|^2\right]\qquad(\text{Using (41)})\\
&\le\frac{3L\widetilde{\eta}^2}{2\tau^2}\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{d}}^{(k,i)}\right\|^2\right]+\frac{3L\widetilde{\eta}^2}{2}\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{q}}^{(k,i)}\right\|^2\right]+\frac{3L\widetilde{\eta}^4}{2}\mathbb{E}_k\left[\left\|\tilde{\mathbf{r}}^{(k,0)}\right\|^2\right].
\end{aligned}
\tag{46}
$$

其中，在式 (46) 中我们使用了 $\|a+b+c\|^2\le3(\|a\|^2+\|b\|^2+\|c\|^2)$。代入式 (26) 和引理 D.2 中的界

将引理 D.4 代入式 (46)，得到

$$
\begin{aligned}
&\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)}-\mathbf{w}^{(k,0)}\right\|^2\right] \leq \frac{\widetilde{\eta}N\overline{K}}{200}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{81L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{50\tau}+\frac{209\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{50}\\
&+\frac{L^3\widetilde{\eta}^4\overline{K}^4N^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{9L^3\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4\alpha^2}{100}+\frac{3L\widetilde{\eta}^2N\gamma^2\overline{K}^2}{2}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\\
&+\frac{3\widetilde{\eta}N\overline{K}}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{9\widetilde{\eta}N\overline{K}}{500}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{3\widetilde{\eta}^2L\overline{K}^2N\sigma^2}{200\tau}+\frac{173\widetilde{\eta}^4\overline{K}^4N^3L^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\\
&+\frac{3\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{50}+\frac{59\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4L^3\alpha^2}{10}\\
&\leq \widetilde{\eta}N\overline{K}\left(\frac{1}{200}+\frac{3}{20}+\frac{9}{500}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{\tau}\left(\frac{81}{50}+\frac{3}{200}\right)\\
&+(\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2)\left(\frac{209}{50}+\frac{3}{50}\right)+L\widetilde{\eta}^2\overline{K}^2N\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\left(\frac{174}{2000}+\frac{3}{2}\right) \qquad (\because\ \eta\leq 1/(10\tau NL\overline{K}))\\
&+L^3\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4\alpha^2\left(\frac{9}{100}+\frac{59}{10}\right)\\
&\leq \frac{173\widetilde{\eta}N\overline{K}}{1000}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{41L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{25\tau}+\frac{41\widetilde{\eta}\eta^2L^2\overline{K}N\tau(\tau-1)\nu^2}{250}+\frac{8L\widetilde{\eta}^2\overline{K}^2N\gamma^2}{5}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\\
&+6L^3\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4\alpha^2
\end{aligned}
\tag{47}
$$

将式 (47) 和式 (45) 推导出的界代入式 (43) 的上界，可得

$$
\begin{aligned}
&\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})]-F(\mathbf{w}^{(k,0)})\\
&\leq -\frac{\widetilde{\eta}\overline{K}N}{2}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+103\widetilde{\eta}\eta^2L^2N\overline{K}(\tau-1)\tau\nu^2+\frac{\widetilde{\eta}\eta^2L^2N(\overline{K}-1)^2(\tau-1)^2\alpha^2}{2\overline{K}}\\
&\quad+129\widetilde{\eta}^3L^2N^3\overline{K}^2(\overline{K}-1)\alpha^2+\frac{41L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{25\tau}+\frac{8L\widetilde{\eta}^2\overline{K}^2N\gamma^2}{5}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)
\end{aligned}
\tag{48}
$$

根据假设 2，我们有

$$
\begin{aligned}
&\mathbb{E}[F(\mathbf{w}^{(k+1,0)})]-F^*\\
&\leq (1-\widetilde{\eta}\overline{K}N\mu)(\mathbb{E}[F(\mathbf{w}^{(k,0)})]-F^*)+103\widetilde{\eta}\eta^2L^2N\overline{K}(\tau-1)\tau\nu^2+\frac{\widetilde{\eta}\eta^2L^2N(\overline{K}-1)^2(\tau-1)^2\alpha^2}{2\overline{K}}\\
&\quad+129\widetilde{\eta}^3L^2N^3\overline{K}^2(\overline{K}-1)\alpha^2+\frac{41L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{25\tau}+\frac{8L\widetilde{\eta}^2\overline{K}^2N\gamma^2}{5}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right),
\end{aligned}
\tag{49}
$$

展开（49）后，我们得到

$$
\begin{aligned}
&\mathbb{E}[F(\mathbf{w}^{(K,0)})]-F^*\\
&\leq\left(1-\frac{\log(MK^2)}{K}\right)^K(F(\mathbf{w}^{(0,0)})-F^*)+\frac{103\widetilde{\eta}\eta^2L^2N\overline{K}(\tau-1)\tau\nu^2}{\widetilde{\eta}\overline{K}N\mu}+\frac{\widetilde{\eta}\eta^2L^2N(\overline{K}-1)(\tau-1)^2\alpha^2}{2\widetilde{\eta}\overline{K}N\mu}\\
&\quad+\frac{129\widetilde{\eta}^3L^2N^3\overline{K}^2(\overline{K}-1)\alpha^2}{\widetilde{\eta}\overline{K}N\mu}+\frac{41L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{25\tau\widetilde{\eta}\overline{K}N\mu}+\frac{8L\widetilde{\eta}^2\overline{K}^2N\gamma^2}{5\widetilde{\eta}\overline{K}N\mu}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{F(\mathbf{w}^{(0,0)})-F^*}{MK^2}+\frac{103\eta^2L^2(\tau-1)\tau\nu^2}{\mu}+\frac{\eta^2L^2(\overline{K}-1)(\tau-1)^2\alpha^2}{2\overline{K}\mu}+\frac{129\widetilde{\eta}^2L^2N^2\overline{K}(\overline{K}-1)\alpha^2}{\mu}\\
&\quad+\frac{41L\widetilde{\eta}\overline{K}\sigma^2}{25\tau\mu}+\frac{8L\widetilde{\eta}\overline{K}\gamma^2}{5\mu}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\\
&=\frac{F(\mathbf{w}^{(0,0)})-F^*}{MK^2}+\frac{103\log^2(MK^2)\kappa^2(\tau-1)\nu^2}{\mu\tau N^2\overline{K}^2K^2}+\frac{\log^2(MK^2)\kappa^2(\overline{K}-1)(\tau-1)\alpha^2}{2\mu\tau N^2\overline{K}^3K^2}\\
&\quad+\frac{196\log^2(MK^2)\kappa^2(\overline{K}-1)\alpha^2}{\mu\overline{K}K^2}+\frac{41\log(MK^2)\kappa\sigma^2}{25\mu\tau NK}+\frac{8\log(MK^2)\kappa\gamma^2}{5\mu NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\\
&=\frac{F(\mathbf{w}^{(0,0)})-F^*}{MK^2}+\widetilde{\mathcal{O}}\left(\frac{\kappa^2(\tau-1)\nu^2}{\mu\tau N^2\overline{K}^2K^2}\right)+\widetilde{\mathcal{O}}\left(\frac{\kappa^2(\tau-1)(\overline{K}-1)\alpha^2}{\mu\tau N^2\overline{K}^3K^2}\right)+\widetilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)\alpha^2}{\mu\overline{K}K^2}\right)+\widetilde{\mathcal{O}}\left(\frac{\kappa\sigma^2}{\mu\tau NK}\right)\\
&\quad+\widetilde{\mathcal{O}}\left(\frac{\kappa\gamma^2}{\mu NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right)
\end{aligned}
\tag{50}
$$

D.4 推论 2 的证明

我们重申 根据 (5)，算法 1 中本地随机梯度下降情形下达到 ϵ 误差的总成本 as

$$
C_{\mathrm{SGD}}=\frac{c_{\mathrm{SGD}}\overline{K}}{\mu\epsilon N}\left[\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{\sigma^2}{\tau}\right]. \tag{51}
$$

要得到 $C_{\mathrm{SGD}|\overline{K}>1}<C_{\mathrm{SGD}|K=1}$，我们需要

$$
\begin{aligned}
\frac{c_{\mathrm{SGD}}\overline{K}}{\mu\epsilon N}\left[\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{\sigma^2}{\tau}\right]&<\frac{c_{\mathrm{SGD}}}{\mu\epsilon N}\left[\gamma^2\left(\frac{M-N}{M-1}\right)+\frac{\sigma^2}{\tau}\right]\\
\Rightarrow(\overline{K}-1)\frac{\sigma^2}{\tau}&<\gamma^2\left(\frac{M-N}{M-1}-\overline{K}\frac{M-N\overline{K}}{M-\overline{K}}\right)\\
&=\gamma^2\frac{M(\overline{K}-1)\left(N\overline{K}-M+N-\frac{N\overline{K}}{M}\right)}{(M-1)(M-\overline{K})}.
\end{aligned}
$$

对于 $K=M/N$，右侧为正。然而，即使对于 $K=M/N-1$，我们也注意到右侧有

$$
N\overline{K}-M+N-\frac{N\overline{K}}{M}=M-N-M+N-\frac{N}{M}\left(\frac{M}{N}-1\right)<0.
$$

因此，除了 $K=\overline{M}/N$ 外，我们可以看到，要使 $\gamma\geq A_1\sigma^2/\tau$ 达到 $C_{\mathrm{SGD}|K>1}\nRightarrow C_{\mathrm{SGD}|K=1}$，需要满足

$$
\left(\frac{M}{N}-1\right)\frac{\sigma^2}{\tau}<\frac{A_1\sigma^2}{\tau}\left(\frac{M-N}{M-1}\right) \tag{52}
$$

$$
\Rightarrow\frac{M-N}{N}<A_1\left(\frac{M-N}{M-1}\right) \tag{53}
$$

$$
\therefore\frac{M-1}{N}<A_1 \tag{54}
$$

完成证明。

D.5 中间引理的证明

引理 *D.1* 的证明。

$$\begin{aligned}
&\left\|\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right]\right\| \leq \frac{N\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\left\|\mathbb{E}_k[\mathbf{d}_m^{(k,i)}]\right\| \\
&= \frac{N\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\left\|\mathbb{E}_k\left[\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right]\right\| \\
&\leq \frac{N\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\|. \qquad (55)
\end{aligned}$$

接下来，利用假设 **1**，

$$\begin{aligned}
&\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\| \leq L\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\mathbf{w}_m^{(k,i-1,l)}-\mathbf{w}^{(k,i-1)}\right]\right\| \\
&= L\eta\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\sum_{l'=0}^{l-1}\nabla F_m(\mathbf{w}_m^{(k,i-1,l')},\xi_m^{(k,i-1,l')})\right]\right\| \\
&\leq L\eta\sum_{l=0}^{\tau-1}\sum_{l'=0}^{l-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l')})-\nabla F_m(\mathbf{w}^{(k,i-1)})+\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\| \\
&\leq L\eta\tau\sum_{l'=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l')})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\|+L\eta\sum_{l=0}^{\tau-1}\sum_{l'=0}^{l-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\| \\
&\Rightarrow(1-L\eta\tau)\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\| \\
&\leq \frac{L\eta\tau(\tau-1)}{2}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}^{(k,i-1)})-\nabla F_m(\mathbf{w}^{(k,0)})+\nabla F_m(\mathbf{w}^{(k,0)})\right]\right\| \\
&\leq \frac{L^2\eta\tau(\tau-1)}{2}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right]\right\|+\frac{L\eta\tau(\tau-1)}{2}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|. \qquad \text{(Using Assumption 1)}
\end{aligned}$$

利用满足 $\eta\leq\frac{1}{10L\tau}$ 的 ~~η，~~我们得到

$$\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right]\right\| \leq \frac{5L\eta\tau(\tau-1)}{9}\left[L\left\|\mathbb{E}_k[\mathbf{w}^{(k,i-1)}]-\mathbf{w}^{(k,0)}\right\|+\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|\right]. \qquad (56)$$

界定 $\left\|\mathbb{E}_k[\mathbf{w}^{(k,i-1)}]-\mathbf{w}^{(k,0)}\right\|$。 与（**41**）类似，我们有

$$\left\|\mathbf{w}^{(k,r)}-\mathbf{w}^{(k,0)}\right\| = \left\|-\frac{\widetilde{\eta}}{\tau}\sum_{i=1}^{r}\overline{\mathbf{d}}^{(k,i)}-\widetilde{\eta}\sum_{i=1}^{r}\overline{\mathbf{q}}^{(k,i)}+\widetilde{\eta}^2\widetilde{\mathbf{r}}^{(k,r)}\right\|,$$

其中 $\widetilde{\ }\mathbf{r}^{(k,r)}:=\sum_{i=1}^{r-1}\left(\prod_{j=i+2}^{r}(\mathbf{I}_d-\widetilde{\eta}\mathbf{S}_-^{(k,j)})\right)\mathbf{S}_-^{(k,i+1)}\sum_{j'=1}^{i}\widetilde{\ }\mathbf{q}^{(k,j')}$。因此，我们有

$$\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,r)}-\mathbf{w}^{(k,0)}\right]\right\| \leq \frac{\widetilde{\eta}}{\tau}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\overline{\mathbf{d}}^{(k,i)}\right]\right\|+\widetilde{\eta}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\overline{\mathbf{q}}^{(k,i)}\right]\right\|+\widetilde{\eta}^2\left\|\mathbb{E}_k\left[\widetilde{\mathbf{r}}^{(k,r)}\right]\right\| \qquad (57)$$

首先，我们将 (57) 式右端的第一项界定如下:

$$\begin{aligned}
&\frac{\widetilde{\eta}}{\tau}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\overline{\mathbf{d}}^{(k,i)}\right]\right\| = \frac{\widetilde{\eta}}{\tau}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right]\right\| \\
&= \frac{\widetilde{\eta}}{\tau}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right]\right\| \\
&\leq \frac{\widetilde{\eta}N\overline{K}}{\tau M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\sum_{l=0}^{\tau-1}\left\|\mathbb{E}_k\left[\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right]\right\| \\
&\leq \frac{\widetilde{\eta}N\overline{K}}{\tau M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\left(\frac{5\eta L(\tau-1)\tau}{9}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|+\frac{5\eta L^2(\tau-1)\tau}{9}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right]\right\|\right) \quad \text{(Using (56))} \\
&= \frac{5\widetilde{\eta}\eta LN\overline{K}(\tau-1)}{9M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|+\frac{5\widetilde{\eta}\eta NL^2(\tau-1)}{9}\sum_{i=1}^{r}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right]\right\|. \qquad (58)
\end{aligned}$$

利用引理 C.2，我们可以对式 (57) 右端的第二项进行界定，得到

$$\widetilde{\eta}\left\|\sum_{i=1}^{r}\mathbb{E}_k\left[\overline{\mathbf{q}}^{(k,i)}\right]\right\| \leq r\widetilde{\eta}N\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\alpha\widetilde{\eta}Nr. \qquad (59)$$

第三，我们可以如下界定式 (57) 右端的第三项:

$$\begin{aligned}
&\widetilde{\eta}^2\left\|\mathbb{E}_k\left[\widetilde{\mathbf{r}}^{(k,r)}\right]\right\| = \widetilde{\eta}^2\left\|\mathbb{E}_k\left[\sum_{i=1}^{r-1}\left(\prod_{j=i+2}^{r}(\mathbf{I}_d-\widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right]\right\| \\
&\leq \widetilde{\eta}^2\sum_{i=1}^{r-1}\left\|\mathbb{E}_k\left[\prod_{j=i+2}^{r}(\mathbf{I}_d-\widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right]\right\|\left\|\mathbb{E}_k\left[\overline{\mathbf{S}}^{(k,i+1)}\right]\right\|\left\|\mathbb{E}_k\left[\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right]\right\| \quad (\because \text{Submultiplicativity of Norms}) \\
&\leq \frac{6\widetilde{\eta}^2NL}{5}\sum_{i=1}^{r-1}\left\|\sum_{j'=1}^{i}\mathbb{E}_k\left[\overline{\mathbf{d}}^{(k,j')}/\tau+\overline{\mathbf{q}}^{(k,j')}\right]\right\| \quad \text{(Using (31), (32) and } \widetilde{\eta}\leq\tfrac{1}{7LN\overline{K}}) \\
&\leq \frac{6\widetilde{\eta}^2NL}{5}\sum_{i=1}^{r-1}\left(\frac{5\eta LN\overline{K}(\tau-1)}{9M}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|+\frac{5\eta NL^2(\tau-1)}{9}\sum_{j'=1}^{i}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right]\right\|\right) \\
&\quad+\frac{6\widetilde{\eta}^2NL}{5}\sum_{i=1}^{r-1}\left(iN\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+i\alpha N\right) \quad \text{(Using (58), (59))} \\
&\leq \frac{2\widetilde{\eta}^2\eta N^2L^2\overline{K}(\tau-1)(r-1)}{3M}\sum_{j'=1}^{r}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|+\frac{2\widetilde{\eta}^2\eta N^2L^3(\tau-1)(r-1)}{3}\sum_{j'=1}^{r-1}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right]\right\| \\
&\quad+\frac{3\widetilde{\eta}^2N^2Lr(r-1)}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{3\widetilde{\eta}^2N^2Lr(r-1)\alpha}{5}. \qquad (60)
\end{aligned}$$

将（58）、（59）和（60）代入（57），并对 r 求和，可得

$$\begin{aligned}
&\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,r)}-\mathbf{w}^{(k,0)}\right]\right\| \\
&\leq\left(\frac{5\widetilde{\eta}\eta LN\overline{K}(\tau-1)}{9M}+\frac{2\widetilde{\eta}^2\eta N^2L^2\overline{K}(\tau-1)(r-1)}{3M}\right)\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|
\end{aligned}$$

$$+\left(\frac{5\widetilde{\eta}\eta NL^2(\tau-1)}{9}+\frac{2\widetilde{\eta}^2\eta N^2L^3(\tau-1)(r-1)}{3}\right)\sum_{i=1}^{r}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right]\right\|$$
$$+\left(r\widetilde{\eta}N+\frac{3\widetilde{\eta}^2N^2Lr(r-1)}{5}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\alpha\widetilde{\eta}Nr+\frac{3\widetilde{\eta}^2N^2Lr(r-1)\alpha}{5}$$
$$\therefore\sum_{r=1}^{\overline{K}}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,r-1)}-\mathbf{w}^{(k,0)}\right]\right\|$$
$$\leq\left(\frac{5\widetilde{\eta}\eta LN\overline{K}(\overline{K}-1)(\tau-1)}{9M}+\frac{\widetilde{\eta}^2\eta N^2L^2\overline{K}^2(\tau-1)(\overline{K}-1)}{3M}\right)\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|$$
$$+\left(\frac{5\widetilde{\eta}\eta NL^2\overline{K}(\tau-1)}{9}+\frac{\widetilde{\eta}^2\eta N^2L^3\overline{K}(\tau-1)(\overline{K}-1)}{3}\right)\sum_{i=1}^{\overline{K}}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right]\right\|$$
$$+\left(\frac{\overline{K}(\overline{K}-1)\widetilde{\eta}N}{2}+\frac{\widetilde{\eta}^2N^2L\overline{K}^2(\overline{K}-1)}{5}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{49\overline{K}(\overline{K}-1)\widetilde{\eta}N\alpha}{100}$$
$$\leq\frac{31\widetilde{\eta}N\overline{K}(\overline{K}-1)}{500M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|+\frac{27\widetilde{\eta}N\overline{K}^2}{50}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|$$
$$+\frac{\overline{K}(\overline{K}-1)\widetilde{\eta}N\alpha}{2}\qquad\left(\because\frac{5\widetilde{\eta}\eta NL^2\overline{K}(\tau-1)}{9}+\frac{\widetilde{\eta}^2\eta N^2L^3\overline{K}(\tau-1)(\overline{K}-1)}{3}\leq 3/500\right)$$
$$\leq\frac{31\widetilde{\eta}N\overline{K}^2}{50}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{31\widetilde{\eta}N\overline{K}(\overline{K}-1)\nu}{500}+\frac{\overline{K}(\overline{K}-1)\widetilde{\eta}N\alpha}{2}.\qquad(\because\ \eta\leq 1/(10\tau NL\overline{K}))\tag{61}$$

最后，因此可得

$$\left\|\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right]\right\|$$
$$\leq\frac{N\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\left(\frac{5\eta L(\tau-1)\tau}{9}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|+\frac{5\eta L^2(\tau-1)\tau}{9}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right]\right\|\right)\quad\text{(Using (56))}$$
$$\leq\frac{5\eta LN\overline{K}(\tau-1)\tau}{9}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{5\eta LN\overline{K}(\tau-1)\tau\nu}{9}$$
$$+\frac{5\eta NL^2(\tau-1)\tau}{9}\left(\frac{31\widetilde{\eta}N\overline{K}^2}{50}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{31\widetilde{\eta}N\overline{K}(\overline{K}-1)\nu}{500}+\frac{\overline{K}(\overline{K}-1)\widetilde{\eta}N\alpha}{2}\right)\qquad\text{(Using (61))}$$
$$\leq\frac{3\eta LN\overline{K}(\tau-1)\tau}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{14\eta LN\overline{K}(\tau-1)\tau\nu}{25}+\frac{\eta LN(\overline{K}-1)(\tau-1)\tau\alpha}{36}.$$

□

引理 *D.3* 的证明。由式（60）可得

$$\widetilde{\eta}^2\left\|\mathbb{E}_k[\widetilde{\mathbf{r}}^{(k,0)}]\right\|\leq\frac{2\widetilde{\eta}^2\eta N^2L^2\overline{K}(\overline{K}-1)(\tau-1)}{3M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|+\frac{3\widetilde{\eta}^2N^2L\overline{K}(\overline{K}-1)}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|$$
$$+\frac{2\widetilde{\eta}^2\eta N^2L^3(\tau-1)(\overline{K}-1)}{3}\sum_{j'=1}^{\overline{K}-1}\left\|\mathbb{E}_k\left[\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right]\right\|+\frac{3\widetilde{\eta}^2N^2L\overline{K}(\overline{K}-1)\alpha}{5}$$

$$
\begin{aligned}
&\leq \frac{2\widetilde{\eta}^2\eta N^2L^2\overline{K}(\overline{K}-1)(\tau-1)}{3M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|+\frac{3\widetilde{\eta}^2N^2L\overline{K}(\overline{K}-1)}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| \\
&\quad+\frac{2\widetilde{\eta}^2\eta N^2L^3(\tau-1)(\overline{K}-1)}{3}\left(\frac{31\widetilde{\eta}N\overline{K}^2}{50}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{31\widetilde{\eta}N\overline{K}^2\nu}{500}+\frac{\overline{K}(\overline{K}-1)\widetilde{\eta}N\alpha}{2}\right) \qquad \text{(Using (61))} \\
&\quad+\frac{3\widetilde{\eta}^2N^2L\overline{K}(\overline{K}-1)\alpha}{5} \\
&\leq \frac{2\widetilde{\eta}^2\eta N^2L^2\overline{K}(\overline{K}-1)(\tau-1)}{3}\left[\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\nu\right]+\frac{3\widetilde{\eta}^2N^2L\overline{K}(\overline{K}-1)}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\| \\
&\qquad\qquad\qquad

这由 (31)、(32) 推出。注意，(63) 中的两项与 (62) 中的前两项形式相似。我们将在界定 (62) 中的前两项后，再回来界定 (63)。

$$\mathbb{E}_k\left[\left\|\sum_{i=1}^{r}\overline{\mathbf{d}}^{(k,i)}\right\|^2\right]=\mathbb{E}_k\left[\left\|\sum_{i=1}^{r}\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right\|^2\right]\leq r\sum_{i=1}^{r}\mathbb{E}_k\left[\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right\|^2\right]\qquad\text{(Using Lemma B.3)}$$

$$=r\sum_{i=1}^{r}\mathbb{E}_k\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}\left(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})+\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right)\right\|^2$$

$$\leq\frac{rN^2\tau\overline{K}}{M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\left(\frac{103\eta^2L^2(\tau-1)\tau\sigma^2}{200}+\frac{69\eta^2L^4\tau^2(\tau-1)}{50}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\right.$$

$$\left.+\frac{69\eta^2L^2\tau^2(\tau-1)}{50}\mathbb{E}_k\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]\right)+r^2N\tau\sigma^2\qquad\text{(Using the bound in (69))}$$

$$=\frac{69rN^2\tau^3(\tau-1)\overline{K}\eta^2L^2}{50M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2+\frac{69rN^2\tau^3(\tau-1)\overline{K}\eta^2L^4}{50M}\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]$$

$$+r^2N\tau\sigma^2\left(1+\frac{103N\tau(\tau-1)L^2\eta^2}{200}\right)\tag{64}$$

对于式 (62) 中的第二项，类似于推导式 (37) 的方法，我们有

$$\mathbb{E}_k\left[\left\|\sum_{i=1}^{r}\overline{\mathbf{q}}^{(k,i)}\right\|^2\right]\leq\frac{1}{3}Nr^2\left[2N\left(\alpha^2+\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\right)+\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right].\tag{65}$$

我们可以利用式 (64) 和 (65) 对式 (63) 作如下界定:

$$\frac{3L\widetilde{\eta}^4}{2}\mathbb{E}_k\left[\left\|\widetilde{\mathbf{r}}^{(k,r)}\right\|^2\right]$$

$$\leq\frac{54N^2L^3\widetilde{\eta}^4(r-1)}{25}\sum_{i=1}^{r-1}\left(\frac{69iN^2\tau(\tau-1)\overline{K}\eta^2L^2}{25M}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right.$$

$$+\frac{69iN^2\tau(\tau-1)\overline{K}\eta^2L^4}{25M}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]+\frac{2i^2N\sigma^2}{\tau}\left(1+\frac{103N\tau(\tau-1)L^2\eta^2}{200}\right)$$

$$\left.+8N^2i^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+4Ni^2\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+8N^2i^2\alpha^2\right)$$

$$\leq\frac{3N^4L^5\tau(\tau-1)\overline{K}\eta^2\widetilde{\eta}^4r(r-1)^2}{M}\sum_{j'=1}^{r}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2+\frac{29N^4L^3\widetilde{\eta}^4(r-1)^2r^2}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2$$

$$+\frac{3N^4L^7\tau(\tau-1)\overline{K}\eta^2\widetilde{\eta}^4(r-1)^2r}{M}\sum_{j'=1}^{r}\sum_{m\in\sigma(j')}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]+\frac{72N^3L^3\widetilde{\eta}^4(r-1)^2r^2\gamma^2}{25}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)$$

$$+\frac{36N^3L^3\widetilde{\eta}^4(r-1)^2r^2\sigma^2}{25\tau}\left(1+\frac{103N\tau(\tau-1)L^2\eta^2}{200}\right)+\frac{29N^4L^3\widetilde{\eta}^4(r-1)^2r^2\alpha^2}{5}.\tag{66}$$

最后，将 (64)、(65) 和 (66) 代入 (62)，得到

$$\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,r)}-\mathbf{w}^{(k,0)}\right\|^2\right]$$

$$
\begin{aligned}
&\leq \frac{2\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}r}{M}\left(2+3\widetilde{\eta}^2L^2N^2(r-1)^2\right)\sum_{j'=1}^{r}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\\
&\quad+\frac{2\widetilde{\eta}^2\eta^2L^4N^2\tau(\tau-1)\overline{K}r}{M}\left(2+3\widetilde{\eta}^2L^2N^2(r-1)^2\right)\sum_{i=1}^{r}\sum_{m\in\sigma(i)}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\\
&\quad+\frac{2\widetilde{\eta}^2Nr^2\sigma^2}{\tau}\left(1+\frac{103\eta^2N\tau(\tau-1)L^2}{200}\right)\left(\frac{3}{2}+\frac{36\widetilde{\eta}^2L^2N^2(r-1)^2}{25}\right)\\
&\quad+2\widetilde{\eta}^2Nr^2\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\left(3+\frac{72\widetilde{\eta}^2L^2N^2(r-1)^2}{25}\right)\\
&\quad+\widetilde{\eta}^2N^2r^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\left(12+\frac{58\widetilde{\eta}^2L^2N^2(r-1)^2}{5}\right)+\widetilde{\eta}^2N^2r^2\alpha^2\left(12+\frac{58\widetilde{\eta}^2L^2N^2(r-1)^2}{5}\right)
\end{aligned}
$$

重新排列各项后，得到

$$
\begin{aligned}
\Rightarrow\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,r-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]&\leq\frac{2\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}(r-1)}{M}\left(2+3\widetilde{\eta}^2L^2N^2\overline{K}^2\right)\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\\
&\quad+\frac{2\widetilde{\eta}^2\eta^2L^4N^2\tau(\tau-1)\overline{K}(r-1)}{M}\left(2+3\widetilde{\eta}^2L^2N^2\overline{K}^2\right)\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\\
&\quad+\frac{2\widetilde{\eta}^2N(r-1)^2\sigma^2}{\tau}\left(1+\frac{103\eta^2N\tau(\tau-1)L^2}{200}\right)\left(\frac{3}{2}+\frac{36\widetilde{\eta}^2L^2N^2\overline{K}^2}{25}\right)\\
&\quad+2\widetilde{\eta}^2N(r-1)^2\gamma^2\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\left(3+\frac{72\widetilde{\eta}^2L^2N^2\overline{K}^2}{25}\right)\\
&\quad+\widetilde{\eta}^2N^2r^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\left(12+\frac{58\widetilde{\eta}^2L^2N^2r^2}{5}\right)+\widetilde{\eta}^2N^2(r-1)^2\alpha^2\left(12+\frac{58\widetilde{\eta}^2L^2N^2r^2}{5}\right)\\
&\leq\frac{102\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}(r-1)}{25M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2+\frac{61\widetilde{\eta}^2N^2(r-1)^2}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&\quad+\frac{102\widetilde{\eta}^2\eta^2L^4N^2\tau(\tau-1)(r-1)}{25}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]+\frac{61\widetilde{\eta}^2N(r-1)^2\sigma^2}{20\tau}\\
&\quad+\frac{303\widetilde{\eta}^2N(r-1)^2\gamma^2}{50}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{61\widetilde{\eta}^2N^2(r-1)^2\alpha^2}{5}\qquad(\because\ \eta\leq1/(10\tau NL\overline{K}))
\end{aligned}
$$

对 $r=1$、.....、K 求和，得到

$$
\begin{aligned}
\Rightarrow\sum_{r=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,r-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]&\leq\frac{51\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}^2(\overline{K}-1}{50M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\\
&\quad+\frac{61\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{15}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{51\widetilde{\eta}^2\eta^2L^4N^2\tau(\tau-1)\overline{K}^2}{50}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\\
&\quad+\frac{61\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{60\tau}+\frac{101\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{50}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{61\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{15}\\
&\quad\therefore\left(1-\frac{51\widetilde{\eta}^2\eta^2L^4N^2\tau(\tau-1)\overline{K}^2}{50}\right)\sum_{r=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,r-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]
\end{aligned}
$$

$$\leq \frac{51\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}^2(\overline{K}-1)}{50M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2+\frac{61\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{15}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2$$
$$+\frac{61\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{60\tau}+\frac{101\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{50}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{61\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{15} \tag{67}$$

再次利用 $\eta\leq 1/(10\tau NLK))$, 得到

$$\sum_{r=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k,r-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]$$
$$\leq \frac{103\widetilde{\eta}^2\eta^2L^2N^2\tau(\tau-1)\overline{K}^2(\overline{K}-1)}{100M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2$$
$$+\frac{41\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{10}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{50\tau}+\frac{203\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{100}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)$$
$$+\frac{41\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{10}$$
$$\leq \frac{83\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{50\tau}+\frac{3\eta^2\overline{K}\tau(\tau-1)\nu^2}{100}+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{25}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)$$
$$+\frac{41\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{10}.$$

□

引理 *D.2* 的证明。

$$\frac{3L\widetilde{\eta}^2}{2\tau^2}\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{d}}^{(k,i)}\right\|^2\right]=\frac{3L\widetilde{\eta}^2}{2\tau^2}\mathbb{E}_k\left[\left\|\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right\|^2\right]\leq\frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\mathbf{d}_m^{(k,i)}\right\|^2\right]$$
$$=\frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})+\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right\|^2\right]$$
$$=\frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1,l)}))\right\|^2\right]\qquad\text{(cross terms are zero)}$$
$$+\frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\left\|\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right\|^2\right]$$
$$\leq\frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\left\|\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1,l)}))\right\|^2\right]$$
(independence of stochastic gradients across clients)
$$+\frac{3L\widetilde{\eta}^2\overline{K}N}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\left\|\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)}))\right\|^2\right]$$

$$\leq \frac{3L\widetilde{\eta}^2\overline{K}}{2\tau^2}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}\left\|\nabla F_m(\mathbf{w}_m^{(k,i-1,l)},\xi_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1,l)})\right\|^2\right]$$

(unbiasedness of stochastic gradients)

$$+\frac{3L\widetilde{\eta}^2\overline{K}N}{2\tau}\sum_{i=1}^{\overline{K}}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{\tau-1}\left\|\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right\|^2\right]$$

$$\leq \frac{3L\widetilde{\eta}^2\overline{K}^2N}{2M\tau^2}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{\tau-1}\sigma^2 \qquad \text{(Assumption 4)}$$

$$+\frac{3L\widetilde{\eta}^2\overline{K}^2N^2}{2M\tau}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{\tau-1}\mathbb{E}_k\left[\left\|\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right\|^2\right]$$

$$\leq \frac{3L\widetilde{\eta}^2\overline{K}^2N}{2\tau}\sigma^2+\frac{3L\widetilde{\eta}^2\overline{K}^2N^2}{2M\tau}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{\tau-1}\mathbb{E}_k\left[\left\|\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right\|^2\right] \qquad (68)$$

由（68）可得

$$\sum_{l=0}^{\tau-1}\mathbb{E}_k\left[\left\|\nabla F_m(\mathbf{w}_m^{(k,i-1,l)})-\nabla F_m(\mathbf{w}^{(k,i-1)})\right\|^2\right]$$

$$\leq L^2\sum_{l=0}^{\tau-1}\mathbb{E}_k\left[\left\|\mathbf{w}_m^{(k,i-1,l)}-\mathbf{w}^{(k,i-1)})\right\|^2\right] \qquad \text{(Assumption 1)}$$

$$\leq \frac{103\eta^2L^2(\tau-1)\tau\sigma^2}{200}+\frac{69\eta^2L^4\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]+\frac{69\eta^2L^2\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]. \qquad (69)$$

将 (69) 代入 (68)，得到

$$\frac{3L\widetilde{\eta}^2}{2\tau^2}\mathbb{E}\left[\left\|\sum_{i=1}^{\overline{K}}\overline{\mathbf{d}}^{(k,i)}\right\|^2\right]$$

$$\leq \frac{3L\widetilde{\eta}^2\overline{K}^2N}{2\tau}\sigma^2+\frac{3L\widetilde{\eta}^2\overline{K}^2N^2}{2M\tau}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\Bigg(\frac{103\eta^2L^2(\tau-1)\tau\sigma^2}{200}$$

$$\frac{69\eta^2L^4\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]+\frac{69\eta^2L^2\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]\Bigg)$$

$$=\frac{8L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{5\tau}+\frac{52L^5\widetilde{\eta}^2\eta^2\overline{K}N^2\tau(\tau-1)}{25}\sum_{i=1}^{\overline{K}}\mathbb{E}\left[\left\|\mathbf{w}^{(k,i-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]$$

$$+\frac{52}{25}\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)(2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+2\nu^2)$$

$$=\frac{8L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{5\tau}+\frac{52L^5\widetilde{\eta}^2\eta^2\overline{K}N^2\tau(\tau-1)}{25}\Bigg(\frac{83\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{50\tau}$$

$$+\frac{3\eta^2\overline{K}\tau(\tau-1)\nu^2}{100}+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{25}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{41\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{10}\Bigg)$$

$$
\begin{aligned}
&+\frac{52}{25}\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)(2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+2\nu^2)\\
&=\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\left(\frac{83\times52\widetilde{\eta}^2L^2N^2\overline{K}^2}{25\times20}+\frac{104}{25}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&+\frac{L\widetilde{\eta}^2N\overline{K}^2\sigma^2}{\tau}\left(\frac{8}{5}+\frac{26\times51L^4\widetilde{\eta}^2\eta^2\overline{K}^2N^2\tau(\tau-1)}{25^2}\right)+\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\left(\frac{104}{25}+\frac{156\eta^2L^2\tau(\tau-1)}{2500}\right)\nu^2\\
&+\frac{52\times51L^5\widetilde{\eta}^4\eta^2\overline{K}^4N^3\tau(\tau-1)\gamma^2}{25^2}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{52\times41L^5\widetilde{\eta}^4\eta^2\overline{K}^3(\overline{K}-1)N^4\tau(\tau-1)\alpha^2}{250}\\
&\leq\frac{\widetilde{\eta}N\overline{K}}{200}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{81L\widetilde{\eta}^2\overline{K}^2N\sigma^2}{50\tau}+\frac{209\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{50}+\frac{L^3\widetilde{\eta}^4\overline{K}^4N^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\\
&+\frac{9L^3\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4\alpha^2}{100}\quad(\because\eta=\log(MK^2)/\mu\tau N\overline{K}K,\ K\geq10\kappa\log(MK^2)).
\end{aligned}
$$

□

引理 *D.4* 的证明。

$$
\begin{aligned}
\frac{3L\widetilde{\eta}^4}{2}\mathbb{E}\left[\left\|\widetilde{\mathbf{r}}^{(k,0)}\right\|^2\right]&=\frac{3L\widetilde{\eta}^4}{2}\mathbb{E}\left[\left\|\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right\|^2\right]\\
&\leq\frac{3L\widetilde{\eta}^4\overline{K}}{2}\sum_{i=1}^{\overline{K}-1}\mathbb{E}\left[\left\|\left(\prod_{j=i+2}^{\overline{K}}(\mathbf{I}_d-\widetilde{\eta}\overline{\mathbf{S}}^{(k,j)})\right)\overline{\mathbf{S}}^{(k,i+1)}\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right\|^2\right]\quad\text{(Using Lemma B.3)}\\
&\leq\frac{54\widetilde{\eta}^4\overline{K}N^2L^3}{25}\sum_{i=1}^{\overline{K}-1}\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right\|^2\right],\qquad(70)
\end{aligned}
$$

其中，式（70）的推导方式与式（36）相同。现在我们对式（70）中的以下项进行界定

$$
\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\widetilde{\mathbf{q}}^{(k,j')}\right\|^2\right]&=\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{d}}^{(k,j')}/\tau+\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]\leq2\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{d}}^{(k,j')}/\tau\right\|^2\right]+2\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]\\
&=\frac{2}{\tau^2}\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\sum_{m\in\mathcal{S}^{(k,j')}}\mathbf{d}_m^{(k,j')}\right\|^2\right]+2\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right].\qquad(71)
\end{aligned}
$$

我们有

$$
\begin{aligned}
&\frac{2}{\tau^2}\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\sum_{m\in\mathcal{S}^{(k,j')}}\mathbf{d}_m^{(k,j')}\right\|^2\right]\leq\frac{2i}{\tau^2}\sum_{j'=1}^{i}\mathbb{E}\left[\left\|\sum_{m\in\mathcal{S}^{(k,j')}}\mathbf{d}_m^{(k,j')}\right\|^2\right]\\
&=\frac{2i}{\tau^2}\sum_{j'=1}^{i}\mathbb{E}\left[\left\|\sum_{m\in\mathcal{S}^{(k,j')}}\sum_{l=0}^{\tau-1}(\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)},\xi_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})+\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}^{(k,j'-1)}))\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{2i}{\tau^2}\sum_{j'=1}^{i}\mathbb{E}\left[\left\|\sum_{m\in\mathcal{S}^{(k,j')}}\sum_{l=0}^{\tau-1}\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)},\xi_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})\right\|^2\right]\\
&\quad+\frac{2i}{\tau^2}\sum_{j'=1}^{i}\mathbb{E}\left[\left\|\sum_{m\in\mathcal{S}^{(k,j')}}\sum_{l=0}^{\tau-1}\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}^{(k,j'-1)}))\right\|^2\right] && \text{(Cross-terms are zero)}\\
&\le \frac{2i}{\tau^2}\sum_{j'=1}^{i}\mathbb{E}\left[\sum_{m\in\mathcal{S}^{(k,j')}}\sum_{l=0}^{\tau-1}\left\|\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)},\xi_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})\right\|^2\right]\\
&\quad+\frac{2iN}{\tau}\sum_{j'=1}^{i}\mathbb{E}\left[\sum_{m\in\mathcal{S}^{(k,j')}}\sum_{l=0}^{\tau-1}\left\|\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}^{(k,j'-1)}))\right\|^2\right] && \text{(Using Lemma B.3)}\\
&\le \frac{2iN\overline{K}}{M\tau^2}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\sum_{l=0}^{\tau-1}\sigma^2 && \text{(Using Assumption 4)}\\
&\quad+\frac{2iN^2\overline{K}}{M\tau}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\sum_{l=0}^{\tau-1}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}_m^{(k,j'-1,l)})-\nabla F_m(\mathbf{w}^{(k,j'-1)}))\right\|^2\right]\\
&\le \frac{2i^2N\sigma^2}{\tau}+\frac{2iN^2\overline{K}}{M\tau}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\left(\frac{103\eta^2L^2(\tau-1)\tau\sigma^2}{200}+\frac{69\eta^2L^4\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\right.\\
&\quad\left.+\frac{69\eta^2L^2\tau^2(\tau-1)}{50}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]\right) && \text{(Using (69))}\\
&= \frac{2i^2N\sigma^2}{\tau}+\frac{103i^2N^2\eta^2L^2(\tau-1)\sigma^2}{100}+\frac{69iN^2\eta^2L^4\tau(\tau-1)}{25}\sum_{j'=1}^{i}\mathbb{E}\left[\left\|\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\\
&\quad+\frac{69iN^2\overline{K}\eta^2L^2\tau(\tau-1)}{25M}\sum_{j'=1}^{i}\sum_{m\in\sigma(j')}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]\\
&\le \frac{2i^2N\sigma^2}{\tau}+\frac{103i^2N^2\eta^2L^2(\tau-1)\sigma^2}{100}+\frac{69iN^2\overline{K}\eta^2L^2\tau(\tau-1)}{25M}\sum_{j'=1}^{\overline{K}}\sum_{m\in\sigma(j')}\mathbb{E}\left[\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]\\
&\quad+\frac{69iN^2\eta^2L^4\tau(\tau-1)}{25}\sum_{j'=1}^{\overline{K}}\mathbb{E}\left[\left\|\mathbf{w}^{(k,j'-1)}-\mathbf{w}^{(k,0)}\right\|^2\right]\\
&\le \frac{2i^2N\sigma^2}{\tau}+\frac{103i^2N^2\eta^2L^2(\tau-1)\sigma^2}{100}+\frac{69iN^2\overline{K}\eta^2L^2\tau(\tau-1)}{25}(2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+2\nu^2)\\
&\quad+\frac{69iN^2\eta^2L^4\tau(\tau-1)}{25}\left(\frac{83\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\sigma^2}{50\tau}+\frac{3\eta^2\overline{K}\tau(\tau-1)\nu^2}{100}\right.\\
&\quad\left.+\frac{51\widetilde{\eta}^2N\overline{K}^2(\overline{K}-1)\gamma^2}{25}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{41\widetilde{\eta}^2N^2\overline{K}^2(\overline{K}-1)\alpha^2}{10}\right) && \text{(Using Lemma 4.1 and Lemma D.5)}\\
&\le i^2N\sigma^2\left(\frac{2}{\tau}+\frac{103N\eta^2L^2(\tau-1)}{100}\right)+iN^2\eta^2L^2\tau(\tau-1)\overline{K}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\left(\frac{138}{25}+\frac{69\times83L^2\widetilde{\eta}^2N^2\overline{K}^2}{1250}\right)\\
&\quad+\frac{69\times51iN^3\eta^2L^4(\tau-1)\widetilde{\eta}^2\overline{K}^3\sigma^2}{1250}+iN^2\eta^2L^2\tau(\tau-1)\overline{K}\nu^2\left(\frac{138}{25}+\frac{207L^2\eta^2\tau(\tau-1)}{2500}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{69\times 51iN^3\eta^2L^4(\tau-1)\widetilde{\eta}^2\overline{K}^3\gamma^2}{25^2}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{69\times 41i\widetilde{\eta}^2\eta^2N^4L^4\tau(\tau-1)\overline{K}^2(\overline{K}-1)\alpha^2}{250}\\
&\le i^2N\sigma^2\left(\frac{2}{\tau}+\frac{103N\eta^2L^2(\tau-1)}{100}\right)+\frac{28iN^2\eta^2L^2\tau(\tau-1)\overline{K}}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{141iN^3\eta^2L^4(\tau-1)\widetilde{\eta}^2\overline{K}^3\sigma^2}{50}\\
&+\frac{111iN^2\eta^2L^2\tau(\tau-1)\overline{K}\nu^2}{20}+\frac{113iN^3\eta^2L^4(\tau-1)\widetilde{\eta}^2\overline{K}^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\\
&+\frac{23i\widetilde{\eta}^2\eta^2N^4L^4\tau(\tau-1)\overline{K}^2(\overline{K}-1)\alpha^2}{2}. \qquad (72)
\end{aligned}
$$

其中，在最后两个界中我们使用 $\eta\le 1/(10\tau NLK)$).将 (72) 代入 (70) 中的 (71)，得到

$$
\begin{aligned}
&\frac{3L\widetilde{\eta}^4}{2}\mathbb{E}\left[\left\|\widetilde{\mathbf{r}}^{(k,0)}\right\|^2\right]\\
&\le\frac{54\widetilde{\eta}^4\overline{K}N^2L^3}{25}\sum_{i=1}^{\overline{K}-1}\left(i^2N\sigma^2\left(\frac{2}{\tau}+\frac{103N\eta^2L^2(\tau-1)}{100}\right)+\frac{28iN^2\eta^2L^2\tau(\tau-1)\overline{K}}{5}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\right.\\
&+\frac{141iN^3\eta^2L^4(\tau-1)\widetilde{\eta}^2\overline{K}^3\sigma^2}{50}+\frac{111iN^2\eta^2L^2\tau(\tau-1)\overline{K}\nu^2}{20}+\frac{113iN^3\eta^2L^4(\tau-1)\widetilde{\eta}^2\overline{K}^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\\
&\left.+\frac{23i\widetilde{\eta}^2\eta^2N^4L^4\tau(\tau-1)\overline{K}^2(\overline{K}-1)\alpha^2}{2}+2\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]\right)\\
&\le\frac{18\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^3L^3\sigma^2}{25}\left(\frac{2}{\tau}+\frac{103N\eta^2L^2(\tau-1)}{100}\right)+\frac{54\times 14\widetilde{\eta}^4\eta^2\overline{K}^3(\overline{K}-1)N^4L^5(\tau-1)\tau}{125}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&+\frac{27\times 141\widetilde{\eta}^6\eta^2\overline{K}^5(\overline{K}-1)N^5L^7(\tau-1)\sigma^2}{1250}+\frac{27\times 111\widetilde{\eta}^4\eta^2\overline{K}^3(\overline{K}-1)N^4L^5(\tau-1)\tau\nu^2}{500}\\
&+\frac{27\times 113\widetilde{\eta}^6\eta^2\overline{K}^5(\overline{K}-1)N^5L^7(\tau-1)\gamma^2}{500}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\\
&+\frac{54\times 23\widetilde{\eta}^6\eta^2N^6L^7\tau(\tau-1)\overline{K}^4(\overline{K}-1)^2\alpha^2}{100}+\frac{108\widetilde{\eta}^4\overline{K}N^2L^3}{25}\sum_{i=1}^{\overline{K}-1}\mathbb{E}\left[\left\|\sum_{j'=1}^{i}\overline{\mathbf{q}}^{(k,j')}\right\|^2\right]\\
&\le\frac{18\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^3L^3\sigma^2}{25}\left(\frac{2}{\tau}+\frac{103N\eta^2L^2(\tau-1)}{100}\right)+\frac{54\times 14\widetilde{\eta}^4\eta^2\overline{K}^3(\overline{K}-1)N^4L^5(\tau-1)\tau}{125}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&+\frac{27\times 141\widetilde{\eta}^6\eta^2\overline{K}^5(\overline{K}-1)N^5L^7(\tau-1)\sigma^2}{1250}+\frac{27\times 111\widetilde{\eta}^4\eta^2\overline{K}^3(\overline{K}-1)N^4L^5(\tau-1)\tau\nu^2}{500}\\
&+\frac{27\times 113\widetilde{\eta}^6\eta^2\overline{K}^5(\overline{K}-1)N^5L^7(\tau-1)\gamma^2}{500}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{54\times 23\widetilde{\eta}^6\eta^2N^6L^7\tau(\tau-1)\overline{K}^4(\overline{K}-1)^2\alpha^2}{100}\\
&+\frac{108\widetilde{\eta}^4\overline{K}N^2L^3}{25}\left(2\overline{K}^3N\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+4N^2\overline{K}^3\left\|\nabla F(\mathbf{w}^{(k,0)}\right\|^2+\frac{4N^2\overline{K}^2(\overline{K}-1)\alpha^2}{3}\right)\quad\text{(Using (37))}\\
&\le\widetilde{\eta}N\overline{K}(\widetilde{\eta}^3\overline{K}^3N^3L^3)\left(\frac{108\times 4}{25}+\frac{54\times 14\widetilde{\eta}^2L^2}{125}\right)\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{27\times 141\widetilde{\eta}^2L\overline{K}^2N\sigma^2}{1250\tau}(\widetilde{\eta}^4\eta^2\overline{K}^4N^4L^6\tau(\tau-1))\\
&+\frac{18\widetilde{\eta}^2L\overline{K}^2N\sigma^2}{25\tau}\left(2\widetilde{\eta}^2\overline{K}^2N^2L^2+\frac{103\widetilde{\eta}^2\overline{K}^2N^3L^4\eta^2\tau(\tau-1)}{100}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\widetilde{\eta}^4\overline{K}^4N^3L^3\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2\left(\frac{216}{25}+\frac{27\times113\widetilde{\eta}^2\eta^2\overline{K}^2N^3L^4(\tau-1)}{500}\right)\\
&+\frac{27\times111\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{500}(\widetilde{\eta}^2\overline{K}^2N^2L^2)\\
&+\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4L^3\alpha^2\left(\frac{108\times4}{75}+\frac{54\times23\widetilde{\eta}^2\eta^2\overline{K}^2N^2L^4\tau(\tau-1)}{100}\right)\\
\leq&\frac{9\widetilde{\eta}N\overline{K}}{500}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{3\widetilde{\eta}^2L\overline{K}^2N\sigma^2}{200\tau}+\frac{173\widetilde{\eta}^4\overline{K}^4N^3L^3\gamma^2}{20}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)+\frac{3\widetilde{\eta}^2\eta^2L^3\overline{K}^2N^2\tau(\tau-1)\nu^2}{50}\\
&+\frac{59\widetilde{\eta}^4\overline{K}^3(\overline{K}-1)N^4L^3\alpha^2}{10}.
\end{aligned}
$$

其中，同样地，在最后两个界中我们使用 $\eta\leq1/(10\tau NLK))$. □

E CyCP+打乱式 SGD 情形的证明

现在，我们将分析扩展到在本地执行打乱式 SGD 的客户端。回顾一下，对于打乱式 SGD，如算法 1 所示，每个客户端在 $m \in \mathcal{S}^{(k,i)}$ 接收全局模型 $\mathbf{w}^{(k,i-1)}$，并将其本地模型初始化为全局模型，即 $\mathbf{w}_m^{(k,i-1,0)} = \mathbf{w}^{(k,i-1)}$。随后，客户端按顺序对其 B个分量执行打乱式 SGD，更新规则为 $\mathbf{w}_m^{(k,i-1,l+1)} = \mathbf{w}_m^{(k,i-1,l)} - \eta \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1,l)})$ $l \in [0, ..., B-1]$其中 $\pi_m^k \sim$Unif($\mathcal{P}_B$)，且 $F_{m,\pi_m^k(l)}(\mathbf{w})$ 是客户端 m 本地损失的第 $\pi_m^k(l)^{\text{th}}$ 个分量，并且每个客户端的所有分量之和等于该客户端的本地损失，即 $F_m(\mathbf{w}) = \frac{1}{B}\sum_{l=0}^{B-1} F_{m,l}(\mathbf{w})$。因此，内循环 $i \in [K]$ 中的更新规则如下：

$$\mathbf{w}^{(k,i)} = \mathbf{w}^{(k,i-1)} - \frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \nabla F_{m,\pi_m^k(l)}(\mathbf{w}_m^{(k,i-1,l)}).$$

使用

$$\begin{aligned}
&\nabla F_{m,\pi_m^k(l)}(\mathbf{w}_m^{(k,i-1,l)}) = \nabla F_{m,\pi_m^k(l)}(\mathbf{w}_m^{(k,i-1,l)}) - \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) + \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) \\
&= \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) + \underbrace{\int_0^1 \nabla^2 F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)} + t(\mathbf{w}_m^{(k,i-1,l)} - \mathbf{w}^{(k,i-1)}))dt}_{:=\widehat{\mathbf{H}}_{m,l}^{(k,i-1)}} (\mathbf{w}_m^{(k,i-1,l)} - \mathbf{w}^{(k,i-1)}),
\end{aligned}$$

我们有

$$\mathbf{w}^{(k,i)} = \mathbf{w}^{(k,i-1)} - \frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \left[\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) + \widehat{\mathbf{H}}_{m,l}^{(k,i-1)} (\mathbf{w}_m^{(k,i-1,l)} - \mathbf{w}^{(k,i-1)}) \right].$$

利用$\mathbf{w}_m^{(k,i-1,0)} = \mathbf{w}^{(k,i-1)}$ 这一事实，我们可以通过递推得到更新规则

$$\mathbf{w}^{(k,i)} = \mathbf{w}^{(k,i-1)} - \frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \left(\prod_{j=B-1}^{l+1} \left(\mathbf{I} - \eta \widehat{\mathbf{H}}_{m,j}^{(k,i-1)} \right) \right) \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) \tag{73}$$

同理，我们可以定义

$$\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,i-1)}) = \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) + \underbrace{\int_0^1 \nabla^2 F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)} + t(\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}))dt}_{:=\widetilde{\mathbf{H}}_{m,l}^{(k,i-1)}} (\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}),$$

得到

$$\begin{aligned}
\mathbf{w}^{(k,i)} &= \mathbf{w}^{(k,i-1)} - \frac{\eta}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \left(\prod_{j=B-1}^{l+1} \left(\mathbf{I} - \eta \widehat{\mathbf{H}}_{m,j}^{(k,i-1)} \right) \right) \left(\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) + \widetilde{\mathbf{H}}_{m,l}^{(k,i-1)} (\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)}) \right) \\
&= \mathbf{w}^{(k,i-1)} - \eta \times \underbrace{\frac{1}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \left(\prod_{j=B-1}^{l+1} \left(\mathbf{I} - \eta \widehat{\mathbf{H}}_{m,j}^{(k,i-1)} \right) \right) \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})}_{:=\mathbf{t}^{(k,i)}} \\
&\quad - \eta \times \underbrace{\frac{1}{N} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \left(\prod_{j=B-1}^{l+1} \left(\mathbf{I} - \eta \widehat{\mathbf{H}}_{m,j}^{(k,i-1)} \right) \right) \widetilde{\mathbf{H}}_{m,l}^{(k,i-1)} (\mathbf{w}^{(k,i-1)} - \mathbf{w}^{(k,0)})}_{:=\mathbf{T}^{(k,i)}}
\end{aligned} \tag{74}$$

展开式（74），即可得到外循环的更新规则如下:

$$\mathbf{w}^{(k+1,0)} = \mathbf{w}^{(k,0)} - \eta \sum_{i=1}^{\overline{K}} \left(\prod_{j=\overline{K}}^{i+1} (\mathbf{I}_d - \eta \mathbf{T}^{(k,j)}) \right) \mathbf{t}^{(k,i)} \tag{75}$$

对 $\prod_{j=K}^{i+1}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})$ 和 $\mathbf{t}^{(k,i)}$ 应用分部求和处理式 (75)，然后再次对 $\prod_{j=B-1}^{l+1}\left(\mathbf{I}-\eta\widehat{\ }\mathbf{H}_m^{(k,i-1)}{}_{,j}\right)$ 和 $\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})$ 在 $\mathbf{t}^{(k,i)}$ 上应用分部求和，可将式 (75) 改写为

$$\begin{aligned}
\mathbf{w}^{(k+1)} &= \mathbf{w}^{(k,0)} - \frac{\eta}{N} \sum_{i=1}^{\overline{K}} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) \\
&+ \eta^2 \times \underbrace{\frac{1}{N} \sum_{i=1}^{\overline{K}} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-2} \left(\prod_{t=B-1}^{l+2} (\mathbf{I}_d - \eta \widehat{\mathbf{H}}_{m,t}^{(k,i-1)}) \right) \widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)} \sum_{j=0}^{l} \nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})}_{:=\mathbf{r}_1^{(k,0)}} \\
&+ \eta^2 \times \underbrace{\frac{1}{N} \sum_{i=1}^{\overline{K}-1} \left(\prod_{j=\overline{K}}^{i+2} (\mathbf{I}_d - \eta \mathbf{T}^{(k,j)}) \right) \mathbf{T}^{(k,i+1)} \left(\sum_{j=1}^{i} \sum_{m \in \mathcal{S}^{(k,j)}} \sum_{l=0}^{B-1} \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) \right)}_{:=\mathbf{r}_2^{(k,0)}} \\
&- \eta^3 \times \underbrace{\frac{1}{N} \sum_{i=1}^{\overline{K}-1} \left(\prod_{j=\overline{K}}^{i+2} (\mathbf{I}_d - \eta \mathbf{T}^{(k,j)}) \right) \mathbf{T}^{(k,i+1)} \left(\sum_{j=1}^{i} \sum_{m \in \mathcal{S}^{(k,j)}} \sum_{l=0}^{B-2} \left(\prod_{t=B-1}^{l+2} (\mathbf{I}_d - \eta \widehat{\mathbf{H}}_{m,t}^{(k,i-1)}) \right) \widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)} \sum_{j'=0}^{l} \nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)}) \right)}_{:=\mathbf{r}_3^{(k,0)}} \\
&= \mathbf{w}^{(k,0)} - \frac{\eta}{N} \sum_{i=1}^{\overline{K}} \sum_{m \in \mathcal{S}^{(k,i)}} \sum_{l=0}^{B-1} \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) + \eta^2 \mathbf{r}_1^{(k,0)} + \eta^2 \mathbf{r}_2^{(k,0)} - \eta^3 \mathbf{r}_3^{(k,0)}.
\end{aligned}$$

对截至$\mathbf{w}^{(k,0)}$ 的所有过去信息取条件期望，可得

$$\mathbb{E}_k[\mathbf{w}^{(k+1,0)}] - \mathbf{w}^{(k,0)} = -\eta \overline{K} B \nabla F(\mathbf{w}^{(k,0)}) + \eta^2 \mathbb{E}_k[\underbrace{\mathbf{r}_1^{(k,0)} + \mathbf{r}_2^{(k,0)} - \eta \mathbf{r}_3^{(k,0)}}_{:=\mathbf{r}^{(k,0)}}],$$

这由对选定的客户端集合 $\mathcal{S}^{(k,i)},\ i \in [K]$ 取期望可得。 —

E.1 中间结果

引理 E.1（CyCP 所致误差项的范数上界）。

$$\begin{aligned}
\left\| \mathbb{E}_k[\mathbf{r}^{(k,0)}] \right\| &\leq \left(\frac{2e^{1/10}}{3} + \frac{2exp(e^{1/10}/10)e^{1/5}\eta BL(\overline{K}-1)}{3} \right) \sqrt{8 \log \left(\frac{4MBK}{\delta} \right)} \overline{K} L (B^{3/2}-1) \overline{\nu} \\
&+ \left(\frac{e^{1/10}}{2} + \frac{exp(e^{1/10}/10)e^{1/5}\eta BL(\overline{K}-1)}{2} \right) \overline{K} LB(B-1)\nu + \frac{exp(e^{1/10}/10)e^{1/10}B^2 L\overline{K}(\overline{K}-1)\alpha}{2} \\
&+ \left(\frac{e^{1/10}}{2} + \frac{exp(e^{1/10}/10)e^{1/10}(\overline{K}-1)}{2} + \frac{exp(e^{1/10}/10)e^{1/5}\eta(B-1)L(\overline{K}-1)}{2} \right) \overline{K} LB^2 \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\|.
\end{aligned}$$

引理 E.2 （CyCP 所致误差项的范数平方上界）。

$$
\begin{aligned}
&3L\eta^4\mathbb{E}_k\left[\left\|\mathbf{r}_1^{(k,0)}\right\|^2+\left\|\mathbf{r}_2^{(k,0)}\right\|^2+\left\|\mathbf{r}_3^{(k,0)}\right\|^2\right]\\
&\leq\frac{31\eta^3\overline{K}L^2(B-1)^2}{10}\log\left(4MBK/\delta\right)\overline{\nu}^2+\frac{\eta^3\overline{K}L^2B(B-1)^2\nu^2}{2}+\frac{3\eta\overline{K}B}{100}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&\quad+\frac{6(\overline{K}-1)\eta^4\overline{K}^3B^4L^3exp(e^{1/10}/5)e^{1/5}}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+4\overline{K}^2(\overline{K}-1)^2B^4L^3\eta^4\alpha^2exp(e^{1/10}/5)e^{1/5}.
\end{aligned}
$$

引理 E.1 和引理 E.2 给出了 CyCP 导致的误差界。指数常数源于学习率设为 $\eta=\log(MK^2)/\tau\mu NKK$，以及周期轮次的下界为 $K\geq 10\kappa\log\left(MK^2\right)$。这些界取决于组件内异质性 ν 以及组内和组间异质性 γ和 α，其中 $\nu=\gamma+\alpha$。

E.2 定理 3 的证明

定理。在假设 *1*-假设 *3* 和假设 *5* 下，对于步长 $\eta=\log(MBK^2)/\mu KBK$以及 $K\geq 10\kappa\log\left(MBK^2\right)$，其中 $\kappa=L/\mu$，收敛误差满足如下界：

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{w}^{(K,0)})]-F^*\leq&\left(1-\mu\eta\overline{K}B\right)^K\left(F(\mathbf{w}^{(0,0)})-F^*\right)+\mathcal{O}\left(\frac{\kappa\eta B\overline{K}}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2\right)\\
&+\tilde{\mathcal{O}}\left(\eta^2(B-1)\kappa L\overline{\nu}^2+\eta^2\kappa L(B-1)^2\nu^2\right)+\mathcal{O}\left((\overline{K}-1)^2B^2L\kappa\eta^2\alpha^2\right).
\end{aligned}
$$

取 $\eta=\log(MBK^2)/\mu BK\overline{K}$，可得

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{w}^{(K,0)})]-F^*\leq&\frac{F(\mathbf{w}^{(0,0)})-F^*}{MBK^2}+\widetilde{\mathcal{O}}\left(\frac{\kappa\gamma^2}{\mu NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right)+\widetilde{\mathcal{O}}\left(\frac{(B-1)}{B^2}\frac{\kappa^2\overline{\nu}^2}{\mu\overline{K}^2K^2}\right)\\
&+\widetilde{\mathcal{O}}\left(\frac{\kappa^2(B-1)^2\nu^2}{\mu B^2\overline{K}^2K^2}\right)+\widetilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu\overline{K}^2K^2}\right),
\end{aligned}
$$

˜其中 $\nu=\gamma+\alpha$，且 $\mathcal{O}(\cdot)$ 涵盖了所有对数项和常数。

推论 2。回顾一下，总通信轮次可表示为 $T=KK$，因此用 T表示时，该界变为

$$
\begin{aligned}
\mathbb{E}[F(\mathbf{w}^{(K,0)})]-F^*\leq&\frac{\overline{K}^2(F(\mathbf{w}^{(0,0)})-F^*)}{MBT^2}+\widetilde{\mathcal{O}}\left(\frac{\kappa^2(B-1)^2\nu^2}{\mu B^2T^2}\right)+\widetilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu T^2}\right)\\
&+\widetilde{\mathcal{O}}\left(\frac{\kappa^2\overline{\nu}^2}{\mu T^2}\frac{(B-1)}{B^2}\right)+\widetilde{\mathcal{O}}\left(\frac{\kappa\overline{K}\gamma^2}{\mu NT}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\right).
\end{aligned}
$$

由于 $K\geq 10\kappa\log\left(MBK^2\right)$，我们有 $T\geq 10\kappa\overline{K}\log\left(MB\overline{K}^2T^2\right)$，而且由于其下界取决于 K，因此无法增加 K而不相应地增加T。

证明。利用全局目标函数 F 的 L-光滑性，我们有

$$
\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})]-F(\mathbf{w}^{(k,0)})\leq\langle\nabla F(\mathbf{w}^{(k,0)}),\mathbb{E}_k[\mathbf{w}^{(k+1,0)}]-\mathbf{w}^{(k,0)}\rangle+\frac{L}{2}\mathbb{E}\left[\left\|\mathbf{w}^{(k+1,0)}-\mathbf{w}^{(k,0)}\right\|^2\right]
$$

$$\leq -\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\eta^2\left\|\mathbb{E}_k[\mathbf{r}^{(k,0)}]\right\|\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)}-\mathbf{w}^{(k,0)}\right\|^2\right]. \tag{76}$$

代入引理 E.1，我们对式 (76) 右端的第二项进行界定。

$$\begin{aligned}
&\eta^2\left\|\mathbb{E}_k[\mathbf{r}^{(k,0)}]\right\|\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\\
&\leq\left(\frac{2e^{1/10}}{3}+\frac{2\exp(e^{1/10}/10)e^{1/5}\eta BL(\overline{K}-1)}{3}\right)\sqrt{8\log\left(\frac{4MBK}{\delta}\right)}\eta^2\overline{K}L(B^{3/2}-1)\overline{\nu}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\\
&\quad+\left(\frac{e^{1/10}}{2}+\frac{\exp(e^{1/10}/10)e^{1/5}\eta BL(\overline{K}-1)}{2}\right)\eta^2\overline{K}LB(B-1)\nu\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\\
&\quad+\frac{\eta^2\exp(e^{1/10}/10)e^{1/10}B^2L\overline{K}(\overline{K}-1)\alpha}{2}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\\
&\quad+\left(\frac{e^{1/10}}{2}+\frac{\exp(e^{1/10}/10)e^{1/10}(\overline{K}-1)}{2}+\frac{\exp(e^{1/10}/10)e^{1/5}\eta(B-1)L(\overline{K}-1)}{2}\right)\eta^2\overline{K}LB^2\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&\leq-\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)}-\mathbf{w}^{(k,0)}\right\|^2\right]\\
&\quad+\frac{47}{20}\sqrt{\log\left(\frac{4MBK}{\delta}\right)}\eta^2\overline{K}L(B^{3/2}-1)\overline{\nu}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{63\eta^2\overline{K}LB(B-1)\nu}{100}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\\
&\quad+\frac{31\eta^2B^2L\overline{K}(\overline{K}-1)\alpha}{50}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\\
&\quad+\left(\frac{e^{1/10}}{20}+\frac{\exp(e^{1/10}/10)e^{1/10}}{20}+\frac{\exp(e^{1/10}/10)e^{1/5}}{200}\right)\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\\
&\leq\frac{13}{100}\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{47}{20}\sqrt{\log\left(\frac{4MBK}{\delta}\right)}\eta^2\overline{K}L(B^{3/2}-1)\overline{\nu}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\\
&\quad+\frac{63\eta^2\overline{K}LB(B-1)\nu}{100}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|+\frac{31\eta^2B^2L\overline{K}(\overline{K}-1)\alpha}{50}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|.
\end{aligned}\tag{77}$$

其中，在最后两个界中，我们使用了 $\eta=\log(MBK^2)/\mu KB\overline{K}$，$K\geq 10\kappa\log(MBK^2)$。接下来，我们对 (76) 中右端的第三项作如下界定：

$$\begin{aligned}
&\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)}-\mathbf{w}^{(k,0)}\right\|^2\right]=\frac{L}{2}\mathbb{E}_k\left[\left\|-\frac{\eta}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})+\eta^2\mathbf{r}^{(k,0)}\right\|^2\right]\\
&\leq L\mathbb{E}_k\left[\left\|-\frac{\eta}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right\|^2\right]+L\mathbb{E}_k\left[\left\|\eta^2\mathbf{r}^{(k,0)}\right\|^2\right]\\
&\leq L\eta^2B^2\overline{K}^2\mathbb{E}_k\left[\left\|\frac{1}{\overline{K}}\sum_{i=1}^{\overline{K}}\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right]+3L\eta^4\mathbb{E}_k\left[\left\|\mathbf{r}_1^{(k,0)}\right\|^2+\left\|\mathbf{r}_2^{(k,0)}\right\|^2+\eta^2\left\|\mathbf{r}_3^{(k,0)}\right\|^2\right]
\end{aligned}\tag{78}$$

$$\leq L\eta^2B^2\overline{K}^2\left[\frac{1}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\right]+3L\eta^4\mathbb{E}_k\left[\left\|\mathbf{r}_1^{(k,0)}\right\|^2+\left\|\mathbf{r}_2^{(k,0)}\right\|^2+\eta^2\left\|\mathbf{r}_3^{(k,0)}\right\|^2\right]. \tag{79}$$

其中，(78) 使用了 $\|a+b+c\|^2 \leq 3(\|a\|^2+\|b\|^2+\|c\|^2)$，而 (79) 使用了 (26)。将引理 E.2 中的界代入 (79)，得到

$$\frac{L}{2}\mathbb{E}_k\left[\left\|\mathbf{w}^{(k+1,0)}-\mathbf{w}^{(k,0)}\right\|^2\right] \leq \frac{21L\eta^2B^2\overline{K}^2}{10N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+\frac{31\eta^3\overline{K}L^2(B-1)^2}{10}\log\left(4MBK/\delta\right)\overline{\nu}^2$$
$$+\frac{\eta^3\overline{K}L^2B(B-1)^2\nu^2}{2}+\frac{\eta\overline{K}B}{20}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{31\overline{K}(\overline{K}-1)^2B^3L^2\eta^3\alpha^2}{50}. \tag{80}$$

最后，将（77）和（79）代入（76），得到

$$\begin{aligned}
&\mathbb{E}_k[F(\mathbf{w}^{(k+1,0)})]-F(\mathbf{w}^{(k,0)})\\
&\leq -\frac{82}{100}\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\frac{21L\eta^2B^2\overline{K}^2}{10N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2\\
&\quad+\left(\frac{\eta^{1/2}\overline{K}^{1/2}B^{1/2}}{10}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)\left(\frac{47\eta^{3/2}\overline{K}^{1/2}L(B^{3/2}-1)\sqrt{\log\left(\frac{4MBK}{\delta}\right)}\overline{\nu}}{2B^{1/2}}\right)\\
&\quad+\left(\frac{\eta^{1/2}\overline{K}^{1/2}B^{1/2}}{10}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)\left(\frac{63\eta^{3/2}\overline{K}^{1/2}LB^{1/2}(B-1)\nu}{10}\right)\\
&\quad+\left(\frac{\eta^{1/2}\overline{K}^{1/2}B^{1/2}}{25}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right)\left(\frac{31\eta^{3/2}\overline{K}^{1/2}(\overline{K}-1)LB^{3/2}\alpha}{2}\right)\\
&\quad+\frac{31\eta^3\overline{K}L^2(B-1)^2}{10}\log\left(4MBK/\delta\right)\overline{\nu}^2+\frac{\eta^3\overline{K}L^2B(B-1)^2\nu^2}{2}+\frac{31\overline{K}(\overline{K}-1)^2B^3L^2\eta^3\alpha^2}{50}\\
&\Rightarrow\mathbb{E}[F(\mathbf{w}^{(k+1,0)})]-F^* \leq (1-\mu\eta B\overline{K})(\mathbb{E}[F(\mathbf{w}^{(k,0)})]-F^*)+\frac{21L\eta^2B^2\overline{K}^2}{10N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2\\
&\quad+\left(\frac{47^2(B^{3/2}-1)^2}{8B}+\frac{31(B-1)^2}{10}\right)\eta^3\overline{K}L^2\log\left(\frac{4MBK}{\delta}\right)\overline{\nu}^2+\left(\frac{63^2}{200}+\frac{1}{2}\right)\eta^3\overline{K}L^2B(B-1)^2\nu^2\\
&\quad+121\overline{K}(\overline{K}-1)^2B^3L^2\eta^3\alpha^2. \quad (\because \textit{Lemma B.1})
\end{aligned} \tag{81}$$

展开（81）可得

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}^{(k+1,0)})]-F^* \leq (1-\mu\eta\overline{K}B)^K(F(\mathbf{w}^{(0,0)})-F^*)+\frac{21\kappa\eta B\overline{K}}{10N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2\\
&\quad+\left(\frac{47^2(B^{3/2}-1)^2}{8B^2}+\frac{31(B-1)^2}{10B}\right)\eta^2\kappa L\log\left(\frac{4MBK}{\delta}\right)\overline{\nu}^2+\left(\frac{63^2}{200}+\frac{1}{2}\right)\eta^2\kappa L(B-1)^2\nu^2\\
&\quad+121(\overline{K}-1)^2B^2L\kappa\eta^2\alpha^2
\end{aligned} \tag{82}$$

由 $\eta=\log(MBK^2)/\mu BK\overline{K}$，可得

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}^{(K,0)})]-F^*\\
&\leq \frac{F(\mathbf{w}^{(0,0)})-F^*}{MBK^2}+\frac{21\kappa\log(MBK^2)}{10\mu NK}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2\\
&\quad+\left(\frac{47^2(B^{3/2}-1)^2}{8B^4}+\frac{31(B-1)^2}{10B^3}\right)\frac{\kappa^2\log\left(4MBK/\delta\right)\log^2(MBK^2)\overline{\nu}^2}{\mu\overline{K}^2K^2}+\frac{21\kappa^2(B-1)^2\log^2(MBK^2)\nu^2}{10\mu B^2\overline{K}^2K^2}\\
&\quad+\frac{121(\overline{K}-1)^2\kappa^2\log^2(MBK^2)\alpha^2}{\mu\overline{K}^2K^2}
\end{aligned}$$

$$= \frac{F(\mathbf{w}^{(0,0)}) - F^*}{MBK^2} + \widetilde{\mathcal{O}}\left(\frac{\kappa\gamma^2}{\mu NK}\left(\frac{M/\overline{K} - N}{M/\overline{K} - 1}\right)\right) + \widetilde{\mathcal{O}}\left(\left(\frac{(B^{3/2}-1)^2}{B^4} + \frac{(B-1)^2}{B^3}\right)\frac{\kappa^2\overline{\nu}^2}{\mu\overline{K}^2K^2}\right)$$
$$+ \widetilde{\mathcal{O}}\left(\frac{\kappa^2(B-1)^2\nu^2}{\mu B^2\overline{K}^2K^2}\right) + \widetilde{\mathcal{O}}\left(\frac{\kappa^2(\overline{K}-1)^2\alpha^2}{\mu\overline{K}^2K^2}\right),$$

至此，证明完毕。 □

E.3 推论3的证明

回顾一下，CyCP+打乱式 SGD 和 LocalRR 的总成本分别如下:

$$C_{\mathrm{SSGD}}(\epsilon) = \widetilde{\mathcal{O}}\left(\frac{c_{\mathrm{SSGD}}\kappa}{\sqrt{\mu\epsilon}}\left(\frac{\overline{K}\sqrt{\mu}}{\kappa\sqrt{MB}} + \nu + \overline{K}\alpha + \frac{\overline{\nu}}{\sqrt{B}}\right)\right) \tag{83}$$

$$C_{\mathrm{LocalRR}}(\epsilon) = \widetilde{\mathcal{O}}\left(\frac{\overline{K}c_{\mathrm{SSGD}}}{\sqrt{\epsilon}}\left(\frac{1}{\sqrt{MB}} + \nu + \frac{\overline{\nu}}{\sqrt{B}}\right)\right) \tag{84}$$

要使 $C_{\mathrm{SSGD}}(\epsilon) < C_{\mathrm{LocalRR}}(\epsilon)$ 成立，我们需要满足因此，等价地，我们需要满足

$$(1-\overline{K})\nu + \overline{K}\alpha + (1-\overline{K})\frac{\overline{\nu}}{\sqrt{B}} < 0.$$

由于我们有 $K = M/\overline{N}$，因此

$$\left(\frac{N-M}{N}\right)\nu + \frac{M\alpha}{N} + \frac{(N-M)\overline{\nu}}{N\sqrt{B}} < 0 \tag{85}$$
$$(N-M)\nu + M\alpha + \frac{(N-M)\overline{\nu}}{\sqrt{B}} < 0$$
$$\left(\alpha - \nu - \frac{\overline{\nu}}{\sqrt{B}}\right)M < -N\nu - \frac{N\overline{\nu}}{\sqrt{B}}$$
$$\left(-\gamma - \frac{\overline{\nu}}{\sqrt{B}}\right)M < -N\left(\alpha + \gamma + \frac{\overline{\nu}}{\sqrt{B}}\right)$$
$$M > N\left(\alpha + \gamma + \frac{\overline{\nu}}{\sqrt{B}}\right) / \left(\gamma + \frac{\overline{\nu}}{\sqrt{B}}\right) = N\left(1 + \frac{\alpha}{\gamma + \frac{\overline{\nu}}{\sqrt{B}}}\right),$$

至此，证明完成。

E.4 中间引理的证明

引理的证明 题 *E.1.* 接下来，我们分别对$\mathbf{r}^{(k,0)}$ 的三个分量进行如下界定。

$$\left\|\mathbb{E}_k[\mathbf{r}_1^{(k,0)}]\right\| = \left\|\mathbb{E}_k\left[\frac{1}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-2}\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d - \eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right]\right\|$$
$$= \left\|\frac{\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\mathbb{E}_k\left[\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d - \eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right]\right\|$$
(∵ uniform i.i.d. 采样)

$$
\begin{aligned}
&\leq \frac{\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\left\|\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\| \\
&\leq \frac{\overline{K}}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\left\|\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\right\|\left\|\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\right\|\left\|\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\| \quad (\because \text{Submultiplicativity of Norms}) \\
&\leq \frac{(1+\eta L)^B\overline{K}L}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\left\|\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\| \quad (\because \left\|\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\right\|\leq L) \\
&\leq \frac{e^{1/10}\overline{K}L}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\left\|\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\| \quad (\text{Using } \eta\leq\tfrac{1}{10BL}) \\
&\leq \frac{e^{1/10}\overline{K}L}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\sum_{l=0}^{B-2}\left(\overline{\nu}\sqrt{8(l+1)\log\left(\frac{4MBK}{\delta}\right)}+(l+1)\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|\right) \quad (\text{Lemma 8 in [58]}) \\
&\leq \frac{2e^{1/10}\overline{K}L(B^{3/2}-1)\overline{\nu}}{3}\sqrt{8\log\left(\frac{4MBK}{\delta}\right)}+\frac{e^{1/10}\overline{K}LB(B-1)}{2M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma(i)}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\| \\
&\leq \frac{2e^{1/10}\overline{K}L(B^{3/2}-1)\overline{\nu}}{3}\sqrt{8\log\left(\frac{4MBK}{\delta}\right)}+\frac{e^{1/10}\overline{K}LB(B-1)\nu}{2}\left[\nu+\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right]. \qquad (86)
\end{aligned}
$$

其中，在最后一个界中我们使用了引理 4.1。类似地，我们可以将下一个噪声项界定为

$$
\begin{aligned}
\left\|\mathbb{E}_k[\mathbf{r}_2^{(k,0)}]\right\| &= \left\|\mathbb{E}_k\left[\frac{1}{N}\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right)\mathbf{T}^{(k,i+1)}\left(\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right)\right]\right\| \\
&\leq \frac{1}{N}\sum_{i=1}^{\overline{K}-1}\left\|\mathbb{E}_k\left[\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right)\mathbf{T}^{(k,i+1)}\left(\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right)\right]\right\| \\
&\qquad\qquad (\text{Using Lemma B.2}) \\
&\leq \frac{1}{N}\sum_{i=1}^{\overline{K}-1}\left\|\mathbb{E}_k\left[\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right)\right]\right\|\left\|\mathbb{E}_k\left[\mathbf{T}^{(k,i+1)}\right]\right\|\left\|\mathbb{E}_k\left[\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right]\right\| \qquad (87)
\end{aligned}
$$

由于

$$
\begin{aligned}
\left\|\mathbf{T}^{(k,i)}\right\| &= \left\|\frac{1}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\left(\prod_{j=B-1}^{l+1}\left(\mathbf{I}-\eta\widehat{\mathbf{H}}_{m,j}^{(k,i-1)}\right)\right)\widetilde{\mathbf{H}}_{m,l}^{(k,i-1)}\right\| \\
&\leq \frac{1}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\left\|\left(\prod_{j=B-1}^{l+1}\left(\mathbf{I}-\eta\widehat{\mathbf{H}}_{m,j}^{(k,i-1)}\right)\right)\widetilde{\mathbf{H}}_{m,l}^{(k,i-1)}\right\| \\
&\leq \frac{1}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}\left\|\prod_{j=B-1}^{l+1}\left(\mathbf{I}-\eta\widehat{\mathbf{H}}_{m,j}^{(k,i-1)}\right)\right\|\left\|\widetilde{\mathbf{H}}_{m,l}^{(k,i-1)}\right\| \\
&\leq \frac{1}{N}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-1}(1+\eta L)^B L \leq e^{1/10}BL, \qquad (88)
\end{aligned}
$$

由于 $\left\|\widetilde{\mathbf{H}}_m^{(k,i-1)}{}_{,l}\right\| \leq L$ 且 $\eta \leq 1/(10LBK)$，我们可以将 (87) 的上界写为

$$
\begin{aligned}
\left\|\mathbb{E}_k[\mathbf{r}_2^{(k,0)}]\right\| &\leq (1+e^{1/10}\eta BL)^{\overline{K}} e^{1/10} BL \sum_{i=1}^{\overline{K}-1} \left\| \mathbb{E}_k \left[\sum_{j=1}^{i} \frac{1}{N} \sum_{m\in\mathcal{S}^{(k,j)}} \sum_{l=0}^{B-1} \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) \right] \right\| \\
&= (1+e^{1/10}\eta BL)^{\overline{K}} e^{1/10} BL \sum_{i=1}^{\overline{K}-1} \left\| \sum_{j=1}^{i} \frac{\overline{K}}{M} \sum_{m\in\sigma^{(j)}} \sum_{l=0}^{B-1} \nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)}) \right\| \qquad (\because \text{unbiased client sampling}) \\
&= \frac{(1+e^{1/10}\eta BL)^{\overline{K}} e^{1/10} B^2 L\overline{K}}{M} \sum_{i=1}^{\overline{K}-1} \left\| \sum_{j=1}^{i} \sum_{m\in\sigma^{(j)}} \nabla F_m(\mathbf{w}^{(k,0)}) \right\| \\
&\leq \exp(e^{1/10}/10) e^{1/10} B^2 L \sum_{i=1}^{\overline{K}-1} \left(i \left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + i\alpha \right) \qquad (\text{using Lemma C.2, and } \eta \leq \tfrac{1}{10LB\overline{K}}) \\
&\leq \exp(e^{1/10}/10) e^{1/10} B^2 L \frac{\overline{K}(\overline{K}-1)}{2} \left[\left\| \nabla F(\mathbf{w}^{(k,0)}) \right\| + \alpha \right]. \qquad (89)
\end{aligned}
$$

对于最后一个噪声项，我们有

$$
\begin{aligned}
&\left\|\mathbb{E}_k[\mathbf{r}_3^{(k,0)}]\right\| \leq \\
&\left\| \mathbb{E}_k \left[\frac{1}{N} \sum_{i=1}^{\overline{K}-1} \left(\prod_{j=\overline{K}}^{i+2} (\mathbf{I}_d - \eta \mathbf{T}^{(k,j)}) \right) \mathbf{T}^{(k,i+1)} \left(\sum_{j=1}^{i} \sum_{m\in\mathcal{S}^{(k,j)}} \sum_{l=0}^{B-2} \left(\prod_{t=B-1}^{l+2} (\mathbf{I}_d - \eta \widehat{\mathbf{H}}_{m,t}^{(k,i-1)}) \right) \widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)} \sum_{j'=0}^{l} \nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)}) \right) \right] \right\| \\
&\leq \sum_{i=1}^{\overline{K}-1} \left\| \mathbb{E}_k \left[\prod_{j=\overline{K}}^{i+2} (\mathbf{I}_d - \eta \mathbf{T}^{(k,j)}) \right] \right\| \left\| \mathbb{E}_k \left[\mathbf{T}^{(k,i+1)} \right] \right\| \\
&\qquad \times \left\| \mathbb{E}_k \left[\sum_{j=1}^{i} \frac{1}{N} \sum_{m\in\mathcal{S}^{(k,j)}} \sum_{l=0}^{B-2} \left(\prod_{t=B-1}^{l+2} (\mathbf{I}_d - \eta \widehat{\mathbf{H}}_{m,t}^{(k,i-1)}) \right) \widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)} \sum_{j'=0}^{l} \nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)}) \right] \right\| \\
&\leq (1+\eta e^{1/10} BL)^{\overline{K}} e^{1/10} BL \sum_{i=1}^{\overline{K}-1} \left\| \sum_{j=1}^{i} \frac{\overline{K}}{M} \sum_{m\in\sigma^{(j)}} \sum_{l=0}^{B-2} \left(\prod_{t=B-1}^{l+2} (\mathbf{I}_d - \eta \widehat{\mathbf{H}}_{m,t}^{(k,i-1)}) \right) \widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)} \sum_{j'=0}^{l} \nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)}) \right\| \\
&\qquad \text{(Using (88))} \\
&\leq \frac{\exp(e^{1/10}/10) e^{1/10} BL\overline{K}}{M} \sum_{i=1}^{\overline{K}-1} \sum_{j=1}^{i} \sum_{m\in\sigma^{(j)}} \sum_{l=0}^{B-2} \mathbb{E}_k \left[\left\| \prod_{t=B-1}^{l+2} (\mathbf{I}_d - \eta \widehat{\mathbf{H}}_{m,t}^{(k,i-1)}) \right\| \left\| \widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)} \right\| \left\| \sum_{j'=0}^{l} \nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)}) \right\| \right] \\
&\qquad \text{(Using } \eta \leq \tfrac{1}{10BL\overline{K}}) \\
&\leq \frac{(1+\eta L)^B \exp(e^{1/10}/10) e^{1/10} BL^2 \overline{K}(\overline{K}-1)}{M} \sum_{j=1}^{\overline{K}} \sum_{m\in\sigma^{(j)}} \sum_{l=0}^{B-2} \mathbb{E}_k \left\| \sum_{j'=0}^{l} \nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)}) \right\| \\
&\qquad \text{(Using } \left\| \widehat{\mathbf{H}}_{m,l+1}^{(k,j)} \right\| \leq L, \text{ for all } j) \\
&\leq \frac{\exp(e^{1/10}/10) e^{1/5} BL^2 \overline{K}(\overline{K}-1)}{M} \sum_{j=1}^{\overline{K}} \sum_{m\in\sigma^{(j)}} \sum_{l=0}^{B-2} \left(\overline{\nu} \sqrt{8(l+1)\log\left(\frac{4MBK}{\delta}\right)} + (l+1) \left\| \nabla F_m(\mathbf{w}^{(k,0)}) \right\| \right) \\
&\qquad \text{(Lemma 8 in [58])}
\end{aligned}
$$

$$\leq \frac{2\exp(e^{1/10}/10)e^{1/5}B(B^{3/2}-1)L^2\overline{K}(\overline{K}-1)}{3}\overline{\nu}\sqrt{8\log\left(\frac{4MBK}{\delta}\right)}$$
$$+\frac{\exp(e^{1/10}/10)e^{1/5}B^2(B-1)L^2\overline{K}(\overline{K}-1)}{2M}\sum_{j=1}^{\overline{K}}\sum_{m\in\sigma^{(j)}}\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|$$
$$\leq \exp(e^{1/10}/10)e^{1/5}BL^2\overline{K}(\overline{K}-1)\left[\frac{2(B^{3/2}-1)}{3}\overline{\nu}\sqrt{8\log\left(\frac{4MBK}{\delta}\right)}+\frac{B(B-1)\nu}{2}+\frac{B(B-1)}{2}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|\right]. \tag{90}$$

最后，利用式 (86)、(89)、(90)，可得

$$\left\|\mathbb{E}_k[\mathbf{r}^{(k,0)}]\right\| \leq \left\|\mathbb{E}_k[\mathbf{r}_1^{(k,0)}]\right\| + \left\|\mathbb{E}_k[\mathbf{r}_2^{(k,0)}]\right\| + \eta\left\|\mathbb{E}_k[\mathbf{r}_3^{(k,0)}]\right\|$$

我们得到最终界。 □

引理 *E.2* 的证明。与界定噪声项的范数类似，我们可以如下界定噪声项的范数平方

$$3\mathbb{E}_k\left[\left\|\mathbf{r}_1^{(k,0)}\right\|^2\right]$$
$$=3\mathbb{E}_k\left[\left\|\frac{1}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-2}\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\|^2\right]$$
$$\leq 3\overline{K}(B-1)\mathbb{E}_k\left[\frac{1}{N}\sum_{i=1}^{\overline{K}}\sum_{m\in\mathcal{S}^{(k,i)}}\sum_{l=0}^{B-2}\left\|\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\|^2\right]nn$$
(Using Lemma B.2, Lemma B.3)
$$\leq \frac{3\overline{K}^2(B-1)}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma^{(i)}}\sum_{l=0}^{B-2}(1+\eta L)^{2B}L^2\mathbb{E}_k\left\|\sum_{j=0}^{l}\nabla F_{m,\pi_m^k(j)}(\mathbf{w}^{(k,0)})\right\|^2$$
$$\leq \frac{6e^{1/5}\overline{K}^2L^2(B-1)}{M}\sum_{i=1}^{\overline{K}}\sum_{m\in\sigma^{(i)}}\sum_{l=0}^{B-2}\left((l+1)(8\log(4MBK/\delta)\overline{\nu}^2+(l+1)^2\left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right)$$
(Using Lemma 8 in [58] and $\eta\leq 1/(10BL\overline{K})$)
$$\leq 4e^{1/5}\overline{K}^2L^2B(B-1)^2\left[6\log(4MBK/\delta)\overline{\nu}^2+B\nu^2+B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\right], \tag{91}$$

and

$$3\mathbb{E}_k\left[\left\|\mathbf{r}_2^{(k,0)}\right\|^2\right]$$
$$=3\mathbb{E}_k\left[\left\|\frac{1}{N}\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right)\mathbf{T}^{(k,i+1)}\left(\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right)\right\|^2\right]$$
$$\leq \frac{3(\overline{K}-1)}{N^2}\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right\|^2\left\|\mathbf{T}^{(k,i+1)}\right\|^2\left\|\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right\|^2\right]$$
(L范数的总和) d b lti li ti it

$$\leq \frac{3(\overline{K}-1)}{N^2}\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[(1+\eta BLe^{1/10})^{2\overline{K}}(B^2L^2e^{1/5})\left\|\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-1}\nabla F_{m,\pi_m^k(l)}(\mathbf{w}^{(k,0)})\right\|^2\right] \quad \text{(Using (88))}$$

$$\leq \frac{3(\overline{K}-1)B^4L^2\exp(e^{1/10}/5)e^{1/5}}{N^2}\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left\|\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right] \quad \text{(Using } \eta\leq 1/(10BL\overline{K}))$$

$$\leq \frac{3(\overline{K}-1)B^4L^2\exp(e^{1/10}/5)e^{1/5}}{N^2}\left(\frac{1}{3}(\overline{K}-1)\overline{K}^2N\left[\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+2N\left(\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\alpha^2\right)\right]\right) \quad \text{(Using (37))}$$

$$= \frac{(\overline{K}-1)^2\overline{K}^2B^4L^2\exp(e^{1/10}/5)e^{1/5}}{N}\left[\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2+2N\left(\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2+\alpha^2\right)\right]. \quad (92)$$

并使用相同的技巧对（91）和（92）进行界定，我们得到以下结果:

$$3\eta^2\mathbb{E}_k\left[\left\|\mathbf{r}_3^{(k,0)}\right\|^2\right] = 3\eta^2\mathbb{E}_k\left[\left\|\frac{1}{N}\sum_{i=1}^{\overline{K}-1}\left(\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right)\mathbf{T}^{(k,i+1)}\right.\right.$$
$$\left.\left.\times\left(\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-2}\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j'=0}^{l}\nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)})\right)\right\|^2\right]$$

$$\leq \frac{3\eta^2(\overline{K}-1)}{N^2}\sum_{i=1}^{\overline{K}-1}\mathbb{E}_k\left[\left(\left\|\prod_{j=\overline{K}}^{i+2}(\mathbf{I}_d-\eta\mathbf{T}^{(k,j)})\right\|\left\|\mathbf{T}^{(k,i+1)}\right\|\right.\right.$$
$$\left.\left.\times\left\|\sum_{j=1}^{i}\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-2}\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j'=0}^{l}\nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)})\right\|\right)^2\right]$$

$$\leq \frac{3\eta^2(\overline{K}-1)}{N^2}\sum_{i=1}^{\overline{K}-1}(1+\eta BLe^{1/10})^{2\overline{K}}(B^2L^2e^{1/5})$$
$$\times(iN(B-1))\sum_{j=1}^{i}\mathbb{E}_k\left[\sum_{m\in\mathcal{S}^{(k,j)}}\sum_{l=0}^{B-2}\left\|\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j'=0}^{l}\nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)})\right\|^2\right]$$

$$\leq \frac{3\eta^2(\overline{K}-1)\overline{K}B^2(B-1)L^2\exp(e^{1/10}/5)e^{1/5}}{M}$$
$$\times\sum_{i=1}^{\overline{K}-1}i\sum_{j=1}^{i}\sum_{m\in\sigma^{(j)}}\sum_{l=0}^{B-2}\mathbb{E}_k\left[\left\|\left(\prod_{t=B-1}^{l+2}(\mathbf{I}_d-\eta\widehat{\mathbf{H}}_{m,t}^{(k,i-1)})\right)\widehat{\mathbf{H}}_{m,l+1}^{(k,i-1)}\sum_{j'=0}^{l}\nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)})\right\|^2\right]$$

$$\leq \frac{3\eta^2(\overline{K}-1)\overline{K}B^2(B-1)L^2\exp(e^{1/10}/5)e^{1/5}}{M}$$
$$\times\sum_{i=1}^{\overline{K}-1}i\sum_{j=1}^{i}\sum_{m\in\sigma^{(j)}}\sum_{l=0}^{B-2}(1+\eta L)^{2B}L^2\mathbb{E}_k\left[\left\|\sum_{j'=0}^{l}\nabla F_{m,\pi_m^k(j')}(\mathbf{w}^{(k,0)})\right\|^2\right]$$

$$
\begin{aligned}
&\leq \frac{6\eta^2(\overline{K}-1)\overline{K}B^2(B-1)L^4\exp(e^{1/10}/5)e^{2/5}}{M} \\
&\quad \times \sum_{i=1}^{\overline{K}-1} i \sum_{j=1}^{i} \sum_{m\in\sigma^{(j)}} \sum_{l=0}^{B-2} \left(8(l+1)\log\left(4MBK/\delta\right)\overline{\nu}^2 + (l+1)^2 \left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2\right) \\
&\leq 12\eta^2(\overline{K}-1)^2\overline{K}^2B^3(B-1)^2L^4\exp(e^{1/10}/5)e^{2/5}\log\left(4MBK/\delta\right)\overline{\nu}^2 \\
&\quad + \frac{\eta^2(\overline{K}-1)^2\overline{K}^2B^4(B-1)^2L^4\exp(e^{1/10}/5)e^{2/5}}{M} \sum_{j=1}^{\overline{K}} \sum_{m\in\sigma^{(j)}} \left\|\nabla F_m(\mathbf{w}^{(k,0)})\right\|^2 \\
&\leq 2\eta^2(\overline{K}-1)^2\overline{K}^2B^3(B-1)^2L^4\exp(e^{1/10}/5)e^{2/5}\left[6\log\left(4MBK/\delta\right)\overline{\nu}^2 + B\left(\nu^2 + \left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2\right)\right].
\end{aligned} \tag{93}
$$

利用 (91)、(92) 和 (93)，我们得到

$$
\begin{aligned}
&L\eta^4\left(3\mathbb{E}_k\left[\left\|\mathbf{r}_1^{(k,0)}\right\|^2\right] + 3\mathbb{E}_k\left[\left\|\mathbf{r}_2^{(k,0)}\right\|^2\right] + 3\eta^2\mathbb{E}_k\left[\left\|\mathbf{r}_3^{(k,0)}\right\|^2\right]\right) \\
&\leq \left(24e^{1/5}\eta\overline{K}LB + \frac{2\eta\overline{K}BL\exp(e^{1/10}/5)e^{2/5}}{25}\right)\eta^3\overline{K}L^2(B-1)^2\log\left(4MBK/\delta\right)\overline{\nu}^2 \\
&\quad + \left(4e^{1/5}\eta\overline{K}LB + \frac{\eta(\overline{K}-1)BL\exp(e^{1/10}/5)e^{2/5}}{50}\right)\eta^3\overline{K}L^2B(B-1)^2\nu^2 \\
&\quad + \left(4e^{1/5}\eta^3\overline{K}L^3B^3 + 12\eta^3\overline{K}^3B^3L^3\exp(e^{1/10}/5)e^{1/5} + \frac{\eta^3\overline{K}B^3L^3\exp(e^{1/10}/5)e^{2/5}}{50}\right)\eta\overline{K}B\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 \\
&\quad + \frac{6(\overline{K}-1)\eta^4\overline{K}^3B^4L^3\exp(e^{1/10}/5)e^{1/5}}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 + 4\overline{K}^2(\overline{K}-1)^2B^4L^3\eta^4\alpha^2\exp(e^{1/10}/5)e^{1/5} \\
&\leq \frac{31\eta^3\overline{K}L^2(B-1)^2}{10}\log\left(4MBK/\delta\right)\overline{\nu}^2 + \frac{\eta^3\overline{K}L^2B(B-1)^2\nu^2}{2} + \frac{3\eta\overline{K}B}{100}\left\|\nabla F(\mathbf{w}^{(k,0)})\right\|^2 \\
&\quad + \frac{6(\overline{K}-1)\eta^4\overline{K}^3B^4L^3\exp(e^{1/10}/5)e^{1/5}}{N}\left(\frac{M/\overline{K}-N}{M/\overline{K}-1}\right)\gamma^2 + 4\overline{K}^2(\overline{K}-1)^2B^4L^3\eta^4\alpha^2\exp(e^{1/10}/5)e^{1/5},
\end{aligned} \tag{94}
$$

由此完成证明。 □