arXiv:2302.04249v2 [cs.LG] 9 Feb 2023

具有客户端异质性的联邦极小极大优化

Pranay Sharma、Rohan Panda 和 Gauri Joshi 电气与计算机工程系，卡内基梅隆大学，匹兹堡，PA 15213{pranaysh、rohanpan、gaurij}@andrew.cmu.edu

2023年2月10日

摘要

随着生成对抗网络等现代应用的出现，极小极大优化受到越来越多的关注，而且与简单的最小化相比，其本质上更具挑战性。当训练数据分布在多个边缘设备或客户端上时，这一难题会进一步加剧，尤其是当这些客户端拥有异质数据集和不同的本地计算能力时。我们提出了一个通用的联邦极小极大优化框架，涵盖了此类设置以及本地SGDA等已有方法。我们表明，对异质的局部进展进行朴素聚合会导致优化一个不匹配的目标函数——这一现象此前已在标准联邦最小化中被观察到。为解决这一问题，我们提出根据连续通信轮次之间所执行的本地步数，对客户端更新进行归一化。我们分析了所提算法在非凸-凹函数和非凸非凹函数类别下的收敛性，并刻画了客户端数据异质性、客户端部分参与以及异质本地计算所产生的影响。与现有文献中的研究相比，我们的分析在客户端内噪声和客户端间异质性方面采用了更一般的假设。对于所考虑的所有函数类别，我们都显著改进了已有的计算复杂度和通信复杂度结果。实验结果支持我们的理论结论。

1 引言

过去十年中，机器学习（ML）研究的爆发式增长催生了许多无法建模为简单最小化问题的新应用。其中许多问题，包括生成对抗网络（GANs）Goodfellow 等人 （2014年）；Arjovsky 等人 （2017）；Sanjabi 等人 （2018年）、对抗神经网络训练Madry 等人 （2018年）、鲁棒优化 Namkoong 与 Duchi （2016年）；Mohajerin Esfahani 与 Kuhn （2018年）、分布式非凸优化 Lu 等人 （2019年）以及公平机器学习Madras 等人 （2018年）；Mohri 等人 （2019年），都具有潜在的极小极大结构。然而，其底层问题通常是非凸的，而经典极小极大理论几乎完全处理凸-凹问题。

现代机器学习应用的另一个特点是训练数据天然具有分布式特性 Xing 等人（2016）。数据收集工作通常外包给边缘设备或客户端。然而，客户端可能因资源限制而无法，或因隐私顾虑而不愿将数据共享给中央服务器。为缓解这一问题，联邦学习（FL）Koneˇcnˋy 等人（2016）；Kairouz 等人（2019）应运而生。客户端在保留数据控制权的同时，也承担部分计算负载，仅使用自己的数据在本地运行部分训练过程。与服务器的通信并不频繁，从而进一步节省资源。自问世以来，联邦学习一直是活跃的研究领域，并取得了一些显著成果 Li 等人（2020）；Wang 等人（2021）。研究已证实一些常用实用技术的实际效益，并为其提供了理论依据，例如客户端执行多次本地更新 Stich（2018）；Khaled 等人（2020）；Koloskova 等人（2020）；Wang 与 Joshi（2021）、客户端部分参与 Yang 等（2021）以及通信压缩 Hamer 等人（2020）；Chen 等人（2021）。此外，研究人员还考察了客户端本地数据的异质性 Zhao 等人（2018）；Sattler 等人（2019），以及客户端系统能力的影响 Wang 等人（2020）；Mitra 等人（2021）。然而，所有这些研究几乎都只关注简单最小化问题。

表1：不同类别的非凸极小极大问题中，达到一个**每客户端**随机梯度复杂度以及所需通信轮次数的比较 ϵ-驻点解（见定义1）。这里 n 表示客户端总数。为与已有工作进行公平比较，本表中的结果专门针对以下情况：（i）所有客户端的权重相同（$p_i = 1/n$），（ii）执行相同次数的本地更新（$\tau_i = \tau$），以及（iii）使用相同的本地更新算法 SGDA。然而，我们的结果（第4节）适用于（i）至（iii）不成立时更一般的情形。

Work	设定与假设		全客户端参与（FCP）	
	系统异质性[a]	部分客户端参与	随机梯度复杂度	通信轮次
非凸-强凹（NC-SC）/非凸-Polyak-Łojasiewicz（NC-PL）				
（$n = 1$）Lin 等人（2020年a）	-	-	$\mathcal{O}(1/\epsilon^4)$	-
Sharma 等人(2022年)	✗	✗	$\mathcal{O}(1/(n\epsilon^4))$	$\mathcal{O}(1/\epsilon^3)$
Yang 等人 (2022a)[b]	✗	✓	$\mathcal{O}(1/(n\epsilon^4))$	$\mathcal{O}(1/\epsilon^2)$
我们的工作(定理 1，推论 1.2)	✓	✓	$\mathcal{O}\left(1/(n\epsilon^4)\right)$	$\mathcal{O}\left(1/\epsilon^2\right)$
非凸-凹（NC-C）				
($n = 1$) Lin 等人(2020a)	-	-	$\mathcal{O}(1/\epsilon^8)$	-
Sharma 等人(2022年)	✗	✗	$\mathcal{O}(1/(n\epsilon^8))$	$\mathcal{O}(1/\epsilon^7)$
我们的工作：(定理 2，推论 2.2)	✓	✓	$\mathcal{O}\left(1/(n\epsilon^8)\right)$	$\mathcal{O}\left(1/\epsilon^4\right)$
非凸-单点凹（NC-1PC）				
Deng& Mahdavi (2021年)	✗	✗	$\mathcal{O}(1/\epsilon^{12})$	$\mathcal{O}(n^{1/6}/\epsilon^8)$
Sharma 等人(2022年)	✗	✗	$\mathcal{O}(1/\epsilon^8)$	$\mathcal{O}(1/\epsilon^7)$
我们的工作：（定理 3）	✓	✓	$\mathcal{O}\left(1/(n\epsilon^8)\right)$	$\mathcal{O}\left(1/\epsilon^4\right)$

[a] 各客户端可以执行数量不等的本地迭代，并使用不同的本地优化器（见第 4 节）。[b] 我们在撰写本文期间读到了 Yang 等人 (2022a)。我们的算法 Fed-Norm-SGDA（算法 1）严格推广了他们的算法 FSGDA。

随着联邦学习系统在大规模应用中的使用日益广泛，它们必须适应各种各样的客户端。数据异质性受到了学界的广泛关注。然而，系统级异质性仍未得到充分研究。通过强制所有客户端执行相同数量的本地更新并使用相同的本地优化器，可以控制客户端差异或 异质性 （Yu 等人 (2019年);Haddadpour 等人 (2019年）。然而，如果客户端数据集大小相差悬殊，这种方法的效率就会很低。此外，这还会导致速度较快的客户端长时间闲置 （Reisizadeh 等人 (2022年）；Tziotis 等人 (2022年），等待速度较慢的客户端完成任务。另外，根据客户端的系统能力，使用相同的优化器可能效率低下或代价高昂。因此，适应系统级异质性是面向现实世界联邦学习方案的一项必要要求。

贡献。 我们考虑一个通用的联邦极小极大优化框架，该框架同时存在客户端间的数据异质性和系统异质性。我们考虑如下问题

$$\min_{\mathbf{x}\in\mathbb{R}^{d_x}} \max_{\mathbf{y}\in\mathbb{R}^{d_y}} \{F(\mathbf{x},\mathbf{y}) := \textstyle\sum_{i=1}^n p_i f_i(\mathbf{x},\mathbf{y})\}, \tag{1}$$

其中， f_i 是客户端 i的本地损失， p_i 是分配给客户端 i的权重，通常是客户端i的相对样本量， n 是客户端总数。我们研究几类非凸极小极大问题。此外，

• 在我们提出的广义联邦极小极大算法中，每一轮的参与客户端可以执行不同数量的本地步数，也可以使用不同的本地优化器。在这种情况下，朴素地聚合本地模型更新（如现有方法本地随机梯度下降上升所采用的做法）可能会导致算法针对一个不匹配的全局目标函数收敛。我们提出了一种简单的归一化策略来解决这一问题。

• 通过采用相互独立的服务器和客户端学习率，我们实现了阶数最优或最先进的计算复杂度，并显著降低了现有方法的通信复杂度。

• 在以下特殊情况下：所有客户端 (i) 均被赋予相同权重 $p_i = 1/n$，即 (1)；(ii) 执行相同次数的本地更新 ($\tau_i = \tau$ 对所有 i 均成立)；且 (iii) 使用相同的本地更新算法，我们的结果便可直接与现有研究进行比较（见表 1），并在以下方面优于它们。

1. 对于非凸-强凹（NC-SC）和非凸-PL（NC-PL）问题，我们的方法具有阶最优的梯度复杂度 $\mathcal{O}(1/(n\epsilon^4))$。此外，我们将通信成本从 Sharma 等人（2022）中的 $\mathcal{O}(1/\epsilon^3)$ 降至 $\mathcal{O}(1/\epsilon^2)$。[1]2. 对于非凸-凹（NC-C）和非凸-单点凹（NC-1PC）问题，我们实现了最先进的梯度复杂度，同时将通信成本从 Sharma 等人（2022）中的 $\mathcal{O}(1/\epsilon^7)$ 显著降低至 $\mathcal{O}(1/\epsilon^4)$。对于 NC-1PC 函数，我们证明了梯度复杂度随 n 线性加速，这正是 Sharma 等人（2022）提出的猜想，从而解决了一个开放问题。3. 作为证明中的一项中间结果，我们证明了本地随机梯度下降在单点凸函数最小化中的理论收敛性（见附录 C.4 中的引理C.5）。所达到的收敛速率与凸最小化问题的收敛速率相同。因此，我们将本地随机梯度下降的收敛性推广到了范围广得多的函数类。

2 相关工作

2.1 单客户端极小极大

非凸-强凹（NC-SC）。据我们所知，Lin 等人（2020a）是首篇分析随机（及确定性）非凸-强凹问题单循环算法的工作。尽管其展示的 $\mathcal{O}(\kappa^3/\epsilon^4)$ 复杂度在 ϵ 方面是最优的，但该算法要求批量大小为 $\mathcal{O}(\epsilon^{-2})$。Qiu等人（2020）利用动量，以 $\mathcal{O}(1)$ 的批量大小实现了 $\mathcal{O}(\epsilon^{-4})$ 收敛。近期的工作 Yang 等人（2022c）和 Sharma 等人（2022）在不使用动量的情况下达到了相同的速率。Yang 等人（2022c）还改进了对条件数 κ 的依赖。Luo和Chen（2021）近期研究了非凸-强凹问题的二阶驻点性。该类问题的下界见 Luo等人（2020）、Li等人（2021）和 Zhang等人（2021）。

非凸-凹（NC-C）。同样，Lin 等人（2020a）首次分析了用于随机NC-C问题的单循环算法，并证明了 $\mathcal{O}(\epsilon^{-8})$ 复杂度。在确定性问题中，借助嵌套算法 Nouiehed 等人（2019）；Thekumparampil等人（2019）以及单循环算法 Xu et al. (2020)；Zhang et al. (2020)，这一结果得到了改进。对于随机问题，Rafique et al. (2021) 以及近期的 Zhang et al. (2022) 将复杂度改进至$\mathcal{O}(\epsilon^{-6})$。然而，这两种算法都具有嵌套结构，即在每一步中迭代求解一个更简单的问题。迄今为止，利用单循环算法实现 $\mathcal{O}(\epsilon^{-6})$ 复杂度仍然难以实现。

2.2 分布式/联邦极小极大

近年来，分布式极小极大优化领域的相关研究也日益增多。其中一部分工作聚焦于去中心化设置，例如 Rogozin 等人 (2021年)；Beznosikov 等人 (2021b、c)；Metelev 等人 (2022年)。

与我们直接相关的是联邦设置，在这种设置中，客户端会在相邻通信轮次之间执行多次本地更新。针对凸-凹问题的相关研究包括 Reisizadeh 等人（2020）；Hou 等人（2021）；Liao 等人（2021）；Sun & Wei（2022）。近期研究考察了联邦设置下的特殊类别非凸极小极大问题，例如非凸-线性问题（Deng 等人，2020）、非凸-PL 问题（Deng & Mahdavi（2021）；Xie 等人（2021））以及非凸-单点凹问题（Deng & Mahdavi（2021））。Sharma 等人（2022）进一步改进了 Deng & Mahdavi（2021）中针对若干函数类别提出的复杂度保证。然而，所有这些研究都考虑了特定的联邦设置，假设全客户端参与，或客户端系统级同质，且每个客户端执行相同数量的本地更新。正如本文所示，客户端部分参与

[1]在撰写本文期间，我们注意到近期的一项工作 Yang 等人 (2022a) 提出了 FSGDA算法，并为非凸-Polyak-Łojasiewicz函数实现了 $\mathcal{O}(1/\epsilon^2)$ 通信成本。不过，我们的工作更具一般性，因为我们允许客户端进行异构数量的本地更新。

是简单联邦学习算法中最主要的误差来源。此外，系统级异质性会对算法性能产生重大影响。

与现有相关工作的区别。Wang 等人 (2020年) 首次研究了简单最小化问题中的系统异质性，并提出归一化平均方案，以避免优化不一致的目标函数。与 Wang 等人 (2020年) 相比，我们研究的问题更具挑战性，并实现了更高的通信节省（表 1）。Yang 等人 (2021年) 分析了联邦学习中的客户端部分参与，并从理论上证明了采用客户端/服务器独立学习率的益处。Deng & Mahdavi (2021年)；Sharma 等人 (2022年) 研究了联邦设置下的极小极大问题，但假设本地更新次数相同且所有客户端均参与。最新的工作 Yang 等人 (2022a) 研究了非凸-强凸问题，并实现了与我们相近的通信节省。然而，我们的工作研究了更一般的极小极大联邦学习框架，其中包含系统层面的客户端异质性和客户端部分参与。我们研究了若干函数类别，并降低了现有极小极大算法的通信与计算复杂度。

3 预备知识

记号。 我们用 $\|\cdot\|$ 表示欧几里得范数 $\|\cdot\|_2$。给定一个正整数 m，集合 $\{1,2,\ldots,m\}$ 记作 $[m]$。客户端 i 上的向量用下标 i 表示，例如 $\mathbf{x}_i$；迭代索引则用上标表示，例如 $\mathbf{y}^{(t)}$ 或 $\mathbf{y}^{(t,k)}$。给定一个函数 g，我们将其梯度向量定义为 $\left[\nabla_x g(\mathbf{x},\mathbf{y})^\top, \nabla_y g(\mathbf{x},\mathbf{y})^\top\right]^\top$，其随机梯度定义为 $\nabla g(\mathbf{x},\mathbf{y};\xi)$，其中 ξ 表示随机性。

收敛度量。 在非凸情形下，我们只能证明收敛到一个近似驻点，其定义如下。

定义 1 （ϵ-驻定性）。 点 $\mathbf{x}$ 是可微函数 g 的 ϵ-驻点，如果 $\|\nabla g(\mathbf{x})\| \le \epsilon$。

定义 2. 随机梯度（SG）复杂度是指单个客户端在算法执行过程中计算的梯度总数。

定义 3 （通信轮次）。 在单个通信轮次中，服务器将其全局模型发送给一组客户端，这些客户端从同一模型出发执行多次本地更新，并将其本地向量返回给服务器。随后，服务器聚合这些本地向量，得到新的全局模型。本文中，我们用 T 表示通信轮次数。

接下来，我们讨论本文中使用的一些假设。

假设 1 （光滑性）。 每个局部函数 f_i 均可微，且具有Lipschitz连续梯度。也就是说，存在一个常数 $L_f > 0$，使得对于每个客户端 $i \in [n]$，对所有$\mathbf{x}$，$\mathbf{x}' \in \mathbb{R}^{d_1}$ 和$\mathbf{y}$，$\mathbf{y}' \in \mathbb{R}^{d_2}$，

$$\|\nabla f_i(\mathbf{x},\mathbf{y}) - \nabla f_i(\mathbf{x}',\mathbf{y}')\| \le L_f \|(\mathbf{x},\mathbf{y}) - (\mathbf{x}',\mathbf{y}')\|.$$

假设 2 （局部方差）。 每个客户端上的随机梯度预言机都是无偏的。此外，存在常数σ_L，$\beta_L \ge 0$，使得对于每个客户端 $i \in [n]$，对所有$\mathbf{x}$，$\mathbf{y}$，

$$\mathbb{E}_{\xi_i}[\nabla f_i(\mathbf{x},\mathbf{y};\xi_i)] = \nabla f_i(\mathbf{x},\mathbf{y}),$$
$$\mathbb{E}_{\xi_i}\|\nabla f_i(\mathbf{x},\mathbf{y};\xi_i) - \nabla f_i(\mathbf{x},\mathbf{y})\|^2 \le \beta_L^2 \|\nabla f_i(\mathbf{x},\mathbf{y})\|^2 + \sigma_L^2.$$

假设 3 (全局 异质性)。 对于任意一组非负权重 $\{w_i\}_{i=1}^n$，使得 $\sum_{i=1}^n w_i = 1$，存在常数 $\beta_G \ge 1, \sigma_G \ge 0$，使得对于所有 $\mathbf{x}$,$\mathbf{y}$,

$$\sum_{i=1}^n w_i \|\nabla_x f_i(\mathbf{x},\mathbf{y})\|^2 \le \beta_G^2 \left\|\sum_{i=1}^n w_i \nabla_x f_i(\mathbf{x},\mathbf{y})\right\|^2 + \sigma_G^2,$$

$$\sum_{i=1}^{n} w_i \left\|\nabla_y f_i\left(\mathbf{x}, \mathbf{y}\right)\right\|^2 \leq \beta_G^2 \left\|\sum_{i=1}^{n} w_i \nabla_y f_i\left(\mathbf{x}, \mathbf{y}\right)\right\|^2 + \sigma_G^2.$$

如果所有 f_i 都相同，则有 $\beta_G = 1$，且 $\sigma_G = 0$。

大多数现有研究采用简化版的假设2、3，假设 $\beta_L = 0$和/或 $\beta_G = 0$。

4 异构联邦极小极大优化算法 n

在本节 中，我们提出了一种联邦极小极大算法，以处理客户端之间的系统异质性 .

4.1 本地SGDA的局限性

继 FedAvg McMahan 等人 （2020年）在联邦学习中的成功之后，Deng & Mahdavi （2021年）首次探索了在极小极大问题中对局部随机梯度下降-上升（SGDA）进行简单扩展的方法。在相邻通信轮次之间，客户端同时执行多次下降/上升步骤，分别更新极小变量$\mathbf{x}$ 和极大变量$\mathbf{y}$。后续研究中，Sharma 等人（2022年）改进了收敛结果，并表明 LocalSGDA 对于若干类非凸极小极大问题实现了最优梯度复杂度。然而，现有关于 LocalSGDA 的研究还假设每个 n 客户端都参与每一通信轮次。更为重要的是，正如在简单最小化问题中观察到的那样，Wang 等人（2020年）指出，如果客户端执行的本地更新次数不相等，或者它们的本地优化器并不完全相同，那么 LocalSGDA（与 FedAvg 类似）可能会收敛到另一个目标函数的驻点。第5.1节、5.2节对此进行了进一步讨论，并在图1中进行了说明，其中学习过程会不成比例地偏向执行更多本地更新的客户端。

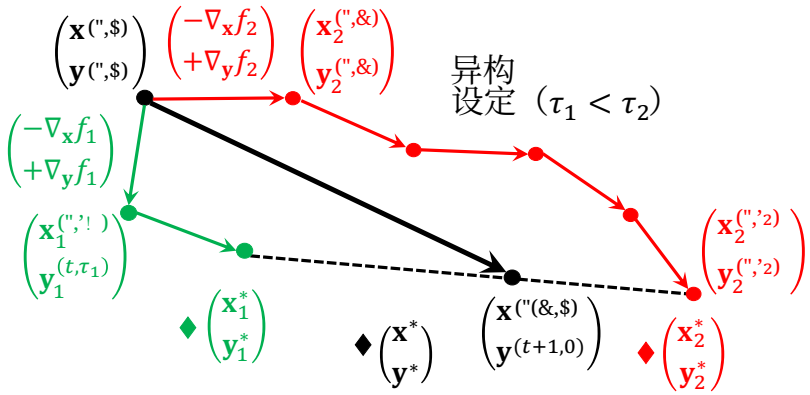

图 1：采用异构本地更新的 FedAvg。绿色（红色）三角形表示 f_1（f_2）的本地优化器，而（$\mathbf{x}^*$，$\mathbf{y}^*$）是全局优化器。客户端 i 的本地更新次数为 τ_i，其中 $\tau_1 = 2, \tau_2 = 5$）。

广义 Local SGDA 更新规则。为了理解本地SGDA中采用朴素聚合时出现的这种不匹配收敛现象，回顾一下，LocalSGDA 的更新形式为

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \gamma_x^s \sum_{i=1}^{n} p_i \Delta_{\mathbf{x},i}^{(t)}, \qquad \mathbf{y}^{(t+1)} = \mathbf{y}^{(t)} + \gamma_y^s \sum_{i=1}^{n} p_i \Delta_{\mathbf{y},i}^{(t)},$$

其中 γ_x^s，γ_y^s是服务器学习率，$\Delta^{(t)}{}_{\mathbf{x}\,i} = \frac{1}{\eta_x^c}\left(\mathbf{x}_{(t,\tau_i^{(t)})\,i} - \mathbf{x}^{(t)}\right)$，$\Delta^{(t)}{}_{\mathbf{y}\,i} = \frac{1}{\eta_y^c}\left(\mathbf{y}_{(t,\tau_i^{(t)})\,i} - \mathbf{y}^{(t)}\right)$ 是缩放后的本地更新。$\mathbf{x}_{(t,\tau_i^{(t)})\,i}$是客户端 i执行 $\tau_i^{(t)}$ 个本地步后的迭代点，而 η_x^c，η_y^c是客户端学习率。我们

考虑此更新规则的一个广义版本，其中 $\Delta^{(t)}{}_{\mathbf{x}\,i}$，$\Delta^{(t)}{}_{\mathbf{y}\,i}$ 是客户端 i 计算得到的局部随机梯度的线性组合，如下所示 $\Delta^{(t)}{}_{\mathbf{y}\,i} = \sum_{k=0}^{\tau_i^{(t)}-1} a_i^{(t,k)} \nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$，其中 $a_i^{(t,k)} \geq 0$。常用的客户端优化器，例如随机梯度下降、局部动量和可变的局部学习率，都可以纳入这一通用形式（示例见附录 A.1）。对于这种更一般的形式，我们可以将服务器端的 $\mathbf{x}$，$\mathbf{y}$ 更新改写如下

$$
\begin{aligned}
\mathbf{x}^{(t+1)} &= \mathbf{x}^{(t)} - \gamma_x^s \textstyle\sum_{i=1}^n p_i \mathbf{G}_{\mathbf{x},i}^{(t)} \frac{\bar{\boldsymbol{a}}_i^{(t)}}{\|\bar{\boldsymbol{a}}_i^{(t)}\|_1} \|\bar{\boldsymbol{a}}_i^{(t)}\|_1 \\
&= \mathbf{x}^{(t)} - \underbrace{\Big(\sum_{j=1}^n p_j \|\bar{\boldsymbol{a}}_j^{(t)}\|_1\Big)}_{\tau_{\text{eff}}^{(t)}} \gamma_x^s \sum_{i=1}^n \underbrace{\frac{p_i \|\bar{\boldsymbol{a}}_i^{(t)}\|_1}{\sum_{j=1}^n p_j \|\bar{\boldsymbol{a}}_j^{(t)}\|_1}}_{w_i} \underbrace{\frac{\mathbf{G}_i^{(t)} \bar{\boldsymbol{a}}_i^{(t)}}{\|\bar{\boldsymbol{a}}_i^{(t)}\|_1}}_{\mathbf{g}_{\mathbf{x},i}^{(t)}}, \\
\mathbf{y}^{(t+1)} &= \mathbf{y}^{(t)} + \tau_{\text{eff}}^{(t)} \gamma_y^s \textstyle\sum_{i=1}^n w_i \mathbf{g}_{\mathbf{y},i}^{(t)},
\end{aligned}
\tag{2}
$$

其中 $G^{(t)}{}_{\mathbf{x}\,i} = [\nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})]_{k=0}^{\tau_i^{(t)}} \in \mathbb{R}^{d_x \times \tau_i^{(t)}}$ 包含按列堆叠的 $\tau_i^{(t)}$ 随机梯度 $\bar{\boldsymbol{a}}_i^{(t)} = [a_i^{t,0}, a_i^{t,1}, \ldots, a_i^{t,\tau_i^{(t)}-1}]^\top$，$\mathbf{g}^{(t)}{}_{\mathbf{x}\,i}$、$\mathbf{g}^{(t)}{}_{\mathbf{y}\,i}$ 是随机梯度的归一化聚合，$\tau^{(t)}{}_{\text{eff}}$ 是有效本地步数。与 Wang 等人(2020年) 中对简单最小化问题的观察类似，

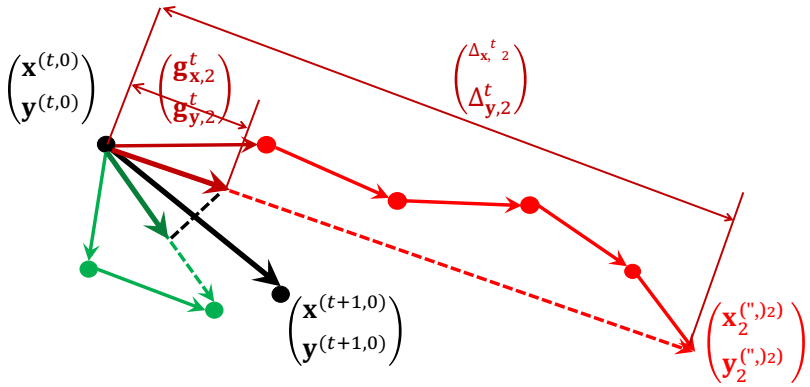

图 2：式（2）中的广义更新规则。注意，$(\mathbf{g}^{(t)}{}_{\mathbf{x}\,i}, \mathbf{g}^{(t)}{}_{\mathbf{y}\,i}) = \frac{1}{\tau_i}(\Delta^{(t)}{}_{\mathbf{x}\,i}, \Delta^{(t)}{}_{\mathbf{y}\,i})$。此外，在服务器端，加权和 $\sum_{i=1}^n w_i \mathbf{g}^{(t)}{}_{\mathbf{x}\,i}$ 会乘以 $\tau^{(t)}{}_{\text{eff}}$。

我们从定理1、2 中看到，这种通用算法最终得到的迭代序列会收敛到不同目标函数的驻点~。$F = \sum_{i=1}^n w_i f_i$此外，在推论1.1中，我们观察到，这种不匹配是由于使用权重 w_i 对（2）中的客户端贡献进行加权所致。

4.2 提出的归一化联邦极小极大算法

从广义更新规则可以看出，将权重 w_i 设为 p_i 可确保代理~目标函数 F 与原始全局目标函数 F一致。将 $w_i = p_i$ 设为可在服务器端聚合之前对各客户端的局部进展进行归一化。因此，即使存在异质性 F，我们仍能保持收敛到原始目标函数的驻点 $\{\tau_i^{(t)}\}$，正如定理1 和定理2所示。

该算法遵循算法 1 中给出的步骤。在每个通信轮次 t中，服务器选择一个客户端集合$\mathcal{C}^{(t)}$，并将其模型参数 $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ 发送给这些客户端。随后，选中的客户端执行多个本地随机梯度步骤。本地步数 $\{\tau^{(t)}{}_i\}$可以因客户端和轮次而异。在完成 $\tau_i^{(t)}$ 个本地步后，客户端 i 将其本地随机梯度聚合为 $\{\mathbf{g}^{(t)}{}_{\mathbf{x}\,i}, \mathbf{g}^{(t)}{}_{\mathbf{y}\,i}\}$，然后将其发送给服务器。注意，客户端 i, $\{\nabla f_i(\cdot, \cdot; \xi_i^{(t,k)})\}_{k=0}^{\tau_i^{(t)}}$处的梯度由 $\|\bar{\boldsymbol{a}}_i^{(t)}\|_1$归一化，其中 $\bar{\boldsymbol{a}}_i^{(t)} = [a_i^{t,0}, a_i^{t,1}, \ldots, a_i^{t,\tau_i^{(t)}-1}]^\top$ 是

算法 1 Fed-N ed-No

1: 输入： 初始化 $\mathbf{x}^{(0)}$，$\mathbf{y}^{(0)}$，通信轮次 T，学习率：客户端 $\{\eta_x^c, \eta_y^c\}$，服务器 $\{\gamma_x^s, \gamma_y^s\}$，本地更新数量 $\{\tau_i^{(t)}\}_{i,\ t}$，S，$s=-1$

2: **对于** $t=0$ 到 $T-1$ **执行**

3: 服务器选择客户端集合 $\mathcal{C}^{(t)}$；向其发送 $(\mathbf{x}^{(t)},\ \mathbf{y}^{(t)})$

4: **如** ，**则**

5: $s \leftarrow s+1$

6: 服务器向 $\widehat{\mathbf{x}}^{(s)} = \mathbf{x}^{(t)}$ 中的客户端发送 $\mathcal{C}^{(t)}$

7: **结束如果**

8: $\mathbf{x}_i^{(t,0)} = \mathbf{x}^{(t)}$，$\mathbf{y}_i^{(t,0)} = \mathbf{y}^{(t)}$ 对于 $i \in \mathcal{C}^{(t)}$

9: **对于** $k=0,\ldots,\tau_i^{(t)}-1$ **执行**

10: $\mathbf{x}_i^{(t,k+1)} = \mathbf{x}_i^{(t,k)} - \eta_x^c a_i^{(t,k)} \nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_{\xi_i^{(t,k)}}^{(t,k)\ \ i})$；

11: $\mathbf{y}_i^{(t,\ k+1)} = \mathbf{y}_i^{(t,k)} + \eta_y^c a_i^{(t,k)} \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$ # y-更新，用于 Fed-Norm-SGDA+

12: $\mathbf{y}_i^{(t,\ k+1)} = \mathbf{y}_i^{(t,k)} + \eta_y^c a_i^{(t,k)} \nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$ # y-更新（Fed-Norm-SGDA）

13: **结束循环**

14: 客户端 i 汇总其梯度以计算 $\mathbf{g}_{\mathbf{x}\ i}^{(t)}, \mathbf{g}_{\mathbf{y}\ i}^{(t)}$

15: $\mathbf{g}_{\mathbf{x}\ i}^{(t)} = \sum_{k=0}^{\tau_i^{(t)}-1} \frac{a_i^{(t,k)}}{\|\bar{\boldsymbol{a}}_i^{(t)}\|_1} \nabla_x f_i(\mathbf{x}_i^{(t,k)},\ \mathbf{y}_i^{(t,k)};\ \xi_i^{(t,k)})$

16: $\mathbf{g}_{\mathbf{y}\ i}^{(t)} = \sum_{k=0}^{\tau_i^{(t)}-1} \frac{a_i^{(t,\ k)}}{\|\bar{\boldsymbol{a}}_i^{(t)}\|_1} \nabla_y f_i\ (\widehat{\mathbf{x}}^{(s)},\ \mathbf{y}_i^{(t,k)};\ \xi_i^{(t,k)})$

17: $\mathbf{g}_{\mathbf{y}\ i}^{(t)} = \sum_{k=0}^{\tau_i^{(t)}-1} \frac{a_i^{(t,\ k)}}{\|\bar{\boldsymbol{a}}_i^{(t)}\|_1} \nabla_y f_i\ (\mathbf{x}_i^{(t,k)},\ \mathbf{y}_i^{(t,k)};\ \xi_i^{(t,k)})$

18: 客户端 $i \in \mathcal{C}^{(t)}$ 与 $\{\mathbf{g}^{(t)}{}_{\mathbf{x}\ i},\ \mathbf{g}^{(t)}{}_{\mathbf{y}\ i}\}$ 通信至服务器

19: 服务器计算聚合向量 $\{\mathbf{g}_{\mathbf{x}}^{(t)}, \mathbf{g}_{\mathbf{y}}^{(t)}\}$，使用 (3)

20: 服务器步：$\begin{cases} \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \tau^{(t)}{}_{\text{eff}}\, \gamma_x^s \mathbf{g}_{\mathbf{x}}^{(t)}, \\ \mathbf{y}^{(t+1)} = \mathbf{y}^{(t)} + \tau^{(t)}{}_{\text{效}}\ \gamma_y^s \mathbf{g}_{\mathbf{y}}^{(t)} \end{cases}$

21: **结束循环**

22: **返回：** $\bar{}\mathbf{x}^{(T)}$ 从 $\{\mathbf{x}^{(t)}\}_{t=1}^{T}$ 中均匀随机抽取

本地更新中分配给各个随机梯度的权重向量。[2] 服务器聚合这些本地向量，以计算全局方向估计 $\mathbf{g}_{\mathbf{x}}^{(t)}$，$\mathbf{g}_{\mathbf{y}}^{(t)}$，然后用它们更新服务器模型参数 $(\mathbf{x}^{(t)},\ \mathbf{y}^{(t)})$。

客户端选择。 在每一轮 t 中，服务器均匀随机抽取 $|\mathcal{C}^{(t)}|$ 个客户端不放回抽样（WOR）。在服务器聚合客户端更新时，客户端 i 更新的权重为 $\tilde{w}_i = w_i n/|\mathcal{C}^{(t)}|$，即，

$$\mathbf{g}_{\mathbf{x}}^{(t)} = \sum_{i\in\mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t)}, \qquad \mathbf{g}_{\mathbf{y}}^{(t)} = \sum_{i\in\mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{y},i}^{(t)}. \tag{3}$$

注意，$\mathbb{E}_{\mathcal{C}^{(t)}}[\mathbf{g}_{\mathbf{x}}^{(t)}] = \sum_{i=1}^n w_i \mathbf{g}^{(t)}{}_{\mathbf{x}\ i}$，$\mathbb{E}_{\mathcal{C}^{(t)}}[\mathbf{g}_{\mathbf{y}}^{(t)}] = \sum_{i=1}^n w_i \mathbf{g}^{(t)}{}_{\mathbf{y}\ i}$。

[2] 对于 LocalSGDA Deng & Mahdavi (2021年)；Sharma 等人 (2022年)，$a_i^{(t,k)} = 1$ 对所有 $i \in [n], t \in [T], k \in [\tau_i^{(t)}]$ 和 $\|\bar{\boldsymbol{a}}_i^{(t)}\|_1 = \tau_i^{(t)}$ 均成立。因此，$\mathbf{g}^{(t)}{}_{\mathbf{x}\ i}, \mathbf{g}^{(t)}{}_{\mathbf{y}\ i}$ 就是在第 t 轮计算得到的随机梯度的平均值。

5 收敛结果

接下来，我们给出不同类别非凸极小极大问题的收敛结果。为简单起见，在本节中，我们假设算法 1 中使用的参数在 t 中保持固定。因此，对于所有 t，$a_i^{(t,k)} \equiv a_i^{(k)}$、$\bar{\boldsymbol{a}}_i^{(t)} \equiv \boldsymbol{a}_i$、$\tau_i^{(t)} \equiv \tau_i$、$\tau^{(t)}{}_{\text{eff}} \equiv \tau_{\text{eff}}$ 和 $|\mathcal{C}^{(t)}| = P$。

5.1 非凸-强凹（NC-SC）情形

假设4(μ- $\mathbf{y}$ 上的强凹性（SC））。若函数 f 在$\mathbf{y}$ 上是 μ-强凹的（$\mu > 0$），则对所有$\mathbf{x}$ $\bar{}\mathbf{y}$,$\mathbf{y}_2$,

$$-f(\mathbf{x},\mathbf{y}_2) \geq -f(\mathbf{x},\bar{\mathbf{y}}) - \langle \nabla_y f(\mathbf{x},\bar{\mathbf{y}}), \mathbf{y}_2 - \bar{\mathbf{y}} \rangle + \frac{\mu}{2}\left\|\mathbf{y}_2 - \bar{\mathbf{y}}\right\|^2.$$

一般收敛结果。我们首先证明，使用算法 1 的本地更新后，迭代序列收敛到代理目标函数 $\widetilde{F}$, $\widetilde{F}(\mathbf{x},\mathbf{y}) \triangleq \sum_{i=1}^n w_i f_i(\mathbf{x},\mathbf{y})$ 的驻点。完整陈述见附录 B 。

定理 1。假设局部损失函数 $\{f_i\}_i$ 满足假设 *1*, *2*, *3*, *4*。假设服务器在每一轮中选择 $|\mathcal{C}^{(t)}| = P$个客户端 t。在适当选择客户端和服务器学习率的情况下，(η_x^c, η_y^c) 和 (γ_x^s, γ_y^s) 分别取相应值时，*Fed-Norm-SGDA* 生成的迭代序列满足

$$\min_{t\in[T]} \mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 \leq \mathcal{O}\Bigg(\underbrace{\kappa^2 \frac{(\bar{\tau}/\tau_{\mathit{eff}}) + A_w\sigma_L^2 + B_w\beta_L^2\sigma_G^2}{\sqrt{P\bar{\tau}T}}}_{\textit{Error with full synchronization}}\Bigg) + \underbrace{\mathcal{O}\Bigg(\kappa^2 \frac{C_w\sigma_L^2 + D\sigma_G^2}{\bar{\tau}^2 T}\Bigg)}_{\textit{Error due to local updates}} + \underbrace{\mathcal{O}\Bigg(\frac{n-P}{n-1}\cdot\frac{\kappa^2 E_w \tau_{\mathit{eff}}\sigma_G^2}{\sqrt{P\bar{\tau}T}}\Bigg)}_{\textit{Partial Participation Error}}. \quad (4)$$

其中，$\kappa = L_f/\mu$ 是条件数，$\widetilde{\Phi}(x) \triangleq max_{\mathbf{y}}\, \widetilde{F}(x,y)$ 是包络函数，$\bar{\tau} = \frac{1}{n}\sum_{i=1}^n \tau_i$, $A_w \triangleq n\tau_{\mathit{eff}} \sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$, $B_w \triangleq n\tau_{\mathit{eff}}\, max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$, $C_w \triangleq \sum_{i=1}^n w_i(\|\boldsymbol{a}_i\|_2^2 - [\alpha^{(t,\tau_i-1)}{}_i \quad]^2)$, $D \triangleq max_i(\beta_L^2\,\|\boldsymbol{a}_{i,-1}\|_2^2 + \|\boldsymbol{a}_{i,-1}\|_1^2)$, 其中 $\boldsymbol{a}_{i,-1} \triangleq [a_i^{(0)}, a_i^{(1)}, \ldots, a^{(\tau_i-2)}{}_i \quad]^\top$ 对所有 i 和 $E_w \triangleq n\, max_i\, w_i$ 均成立。

证明见附录 B 。界（4）中的第一项表示集中式算法的优化误差（参见 Lin 等人 (2020a) 中的附录 C.3）。第二项表示至少有一个客户端执行多次 ($\tau_i > 1$) 本地更新时产生的误差。最后一项源于客户端子采样。这也解释了它为何取决于数据异质性 σ_G。

定理1 陈述 了代理目标函数的收敛性 $\widetilde{F}$。接下来，我们来看真实目标函数的收敛性 F。

推论 1.1(按 F) **衡量的收敛性**。给定 $\Phi(\mathbf{x}) \triangleq \max_{\mathbf{y}} F(\mathbf{x},\mathbf{y})$，在定理 *1* 的条件下 ，

$$\min_{t\in[T]} \left\|\nabla\Phi(\mathbf{x}^{(t)})\right\|^2 \leq 2\left(2\chi^2_{\mathbf{p}\|\mathbf{w}}\beta_H^2 + 1\right)\epsilon_{\mathit{opt}} + 4\chi^2_{\mathbf{p}\|\mathbf{w}}\sigma_G^2 + \frac{4L_f^2}{T}\sum_{t=0}^{T-1}\left\|\mathbf{y}^*(\mathbf{x}^{(t)}) - \widetilde{\mathbf{y}}^*(\mathbf{x}^{(t)})\right\|^2. \quad (5)$$

其中，$\chi^2_{\mathbf{p}\|\mathbf{w}} \triangleq \sum_{i=1}^n \frac{(p_i - w_i)}{w_i}^2$，$\epsilon_{\mathit{opt}} \triangleq \frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\widetilde{\Phi}(x^{(t)})\right\|^2$ 表示式（*4*）中的优化误差。如果对所有 $i \in [n]$都有 $p_i = \widetilde{w_i}$，那么 $\chi^2_{\mathbf{p}\|\mathbf{w}} = 0$。此外，此时 $F(x,y) \equiv \widetilde{F}(x,y)$。因此，对于所有 x，$y^*(x)$ =最大值点集合$_{\mathbf{y}}$ $F(x,y)$ 与 $\widetilde{y}^*(x)$ =最大值点集合$_{\mathbf{y}}$ $\widetilde{F}(x,y)$ 相同。因此，由（*5*）可得 $min_{t\in[T]}\left\|\nabla\Phi(x^{(t)})\right\|^2 \leq 2\epsilon_{\mathit{opt}}$。

由推论 1.1可知，若在算法 $\{w_i\}$ 的服务器更新中将 $\{p_i\}$ 替换为 $F1$，则可获得关于真实目标函数 F的收敛性。

备注 1. 若客户端采用相同权重（$w_i = 1/n$ 对所有 i均成立），且每个客户端执行 τ步本地SGDA，则有$\bar{\tau} = \tau$, $A_w = B_w = 1$，$C_w = \tau - 1$，以及 $D = (\tau-1)(\tau - 1 + \beta_L^2)$。因此，式（4）中的界可大幅简化为

$$\mathcal{O}\left(\frac{\sigma_L^2 + \beta_L^2\sigma_G^2}{\sqrt{P\bar{\tau}T}} + \frac{\sigma_L^2 + \tau\sigma_G^2}{\tau T} + \left(\frac{n-P}{n-1}\right)\sigma_G^2\sqrt{\frac{\tau}{PT}}\right). \quad (6)$$

从（6）可以得出几个关键见解。

• 客户端部分参与（PCP）误差 $\mathcal{O}\left(\frac{n-P}{n-1}\cdot\sigma_G^2\sqrt{\frac{\tau}{PT}}\right)$ 是收敛误差中最显著的组成部分。此外，与另外两种误差不同，PCP 导致的误差实际上会随着本地更新次数的增加而增大 τ。因此，除非 $\sigma_G=0$（见表1）这一特殊情况，否则在客户端执行多次本地更新并不能带来通信节省。Yang 等人在最小化问题中也观察到了类似现象 (2021年)；Jhunjhunwala 等人(2022年)，而最近，Yang 等人在极小极大问题中也观察到了类似现象 (2022a)。

• 在不存在多次本地更新的情况下（即 $\tau_i=1$ 对所有 i），且采用全参与（$P=n$）时，所得误差 $\mathcal{O}\left(\frac{\sigma_L^2+\beta_L^2\sigma_G^2}{\sqrt{nTT}}\right)$ 尽管完全同步，仍取决于全局异质性 σ_G。这是由于采用了更一般的局部方差界（假设 2）。对于 $\beta_L>0$，这种对 σ_G 的依赖不可避免。本文中的所有结果均有这一观察。其依据见备注 6 （附录 A.2）。

推论 1.2（改进的通信节省）。 假设所有客户端的权重相同（$p_i=1/n$ 对所有 i），且每个客户端执行 τ 步本地*SGDA*。要达到一个 ϵ-驻点，即$\mathbf{x}$ 满足 $\mathbb{E}\|\nabla\Phi(\mathbf{x})\|\le\epsilon$,
• 在全参与情况下，*Fed-Norm-SGDA* 的每客户端梯度复杂度为 $T\tau=\mathcal{O}\left(\frac{\kappa^4}{n\epsilon^4}\right)$。所需的通信轮次为 $T=\mathcal{O}\left(\frac{\kappa^2}{\epsilon^2}\right)$。• 在部分参与情况下，当客户端间数据异质性 $\sigma_G=0$ 这一特殊情形下，*Fed-Norm-SGDA* 的每客户端梯度复杂度为 $\mathcal{O}(\kappa^4/(P\epsilon^4))$，而通信成本为 $\mathcal{O}(\kappa^2/\epsilon^2)$。

备注 2。 推论 1.2 中的梯度复杂度在 ϵ 下是最优的，并且随着参与客户端数量的增加实现线性加速。通信复杂度在 ϵ 下也达到最优，并改进了Deng & Mahdavi(2021年)；Sharma 等人 (2022年) 的相应结果。我们的通信成本与近期工作 Yang 等人(2022a) 相当。然而，我们的工作研究了一个包含异构客户端的、更现实的联邦学习设置。

将结果推广至非凸 PL 函数

假设 5. 若函数 f 在$\mathbf{y}$ ($\mu>0$)中满足 μ-PL 条件，则对于任意固定的$\mathbf{x}$:

1. $\max_{\mathbf{y}'} f(\mathbf{x},\mathbf{y}')$ 的解集非空;
2. $\|\nabla_y f(\mathbf{x},\mathbf{y})\|^2\ge 2\mu(\max_{\mathbf{y}'} f(\mathbf{x},\mathbf{y}')-f(\mathbf{x},\mathbf{y}))$, for all $\mathbf{y}$.

备注 3. 若局部函数 $\{f_i\}$ 满足假设 1、2、3，且全局函数 F 满足假设 5，那么采用适当选择的学习率（见附录 B.5），定理 1 中的界 μ对非凸-PL 函数同样成立。

5.2 非凸-凹（NC-C）情形

在本小节中，我们考虑满足以下假设的光滑非凸函数。

假设 6 （凹性）。 函数 f 关于 $\mathbf{y}$ 是凹的，当固定 $\mathbf{x}\in\mathbb{R}^{d_1}$ 时，对所有 $\mathbf{y}$，$\mathbf{y}'\in\mathbb{R}^{d_2}$,

$$f(\mathbf{x},\mathbf{y})\le f(\mathbf{x},\mathbf{y}')+\langle\nabla_y f(\mathbf{x},\mathbf{y}'),\mathbf{y}-\mathbf{y}'\rangle\,.$$

假设 7 （关于 $\mathbf{x}$ 的 Lipschitz 连续性）。 给定函数 f，存在一个常数 $G_{\mathbf{x}}$，使得对于每个 $\mathbf{y}\in\mathbb{R}^{d_2}$，以及所有 $\mathbf{x}$，$\mathbf{x}'\in\mathbb{R}^{d_1}$,

$$\|f(\mathbf{x},\mathbf{y})-f(\mathbf{x}',\mathbf{y})\|\le G_{\mathbf{x}}\|\mathbf{x}-\mathbf{x}'\|\,.$$

目前使用的包络函数 $\Phi(\mathbf{x})=\max_{\mathbf{y}} f(\mathbf{x},\mathbf{y})$ 在不存在唯一最大化点时可能不再光滑。因此，我们采用 Davis& Drusvyatskiy(2019年) 提出的另一种驻定性定义，该定义利用了 Φ的 Moreau 包络，定义如下。

定义 4（Moreau 包络）。 若对于所有 $\mathbf{x}\in\mathbb{R}^{d_x}$，函数 ϕ_λ 是 λ-Moreau 包络 ϕ，其中 $\lambda>0$，则

$$\phi_\lambda(\mathbf{x})=\min_{\mathbf{x}'}\phi(\mathbf{x}')+\frac{1}{2\lambda}\|\mathbf{x}'-\mathbf{x}\|^2\,.$$

Drusvyatskiy & Paquette （2019年）表明，较小的 $\|\nabla\phi_\lambda(\mathbf{x})\|$意味着在 $\tilde{}\mathbf{x}$ 的邻域内存在某个点$\mathbf{x}$，该点对于 ϕ而言是近似驻点 。因此，在我们的情形中，我们着重于最小化 $\|\nabla\Phi_\lambda(\mathbf{x})\|$。

所提出的算法。 对于非凸-凹函数，我们使用 Fed-Norm-SGDA+。 $\mathbf{x}$更新与 Fed-Norm-SGDA 完全相同。然而，对于 $\mathbf{y}$ 更新，客户端计算随机梯度 $\nabla_y f_i(\hat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$ 并将 x分量固定在 $\hat{\mathbf{x}}^{(s)}$，持续 S 个通信轮次。这种 技巧最初由 Deng & Mahdavi（2021年）提出，它在分析上具有双循环算法的优势（先多次更新 $\mathbf{y}$，再更新一次 $\mathbf{x}$），同时也能同步更新$\mathbf{x}$。

定理 2。 假设局部损失函数 $\{f_i\}$ 满足假设 *1*、*2*、*3*、*6*、*7*， $\left\|\mathbf{y}^{(t)}\right\|^2 \leq R$ 对所有 t均成立，并且服务器对所有 $|\mathcal{C}^{(t)}| = P$选择 t个客户端。采用适当的客户端和服务器学习率，分别为 $(\eta_x^c,\ \eta_y^c)$ 和 $(\gamma_x^s,\ \gamma_y^s)$时，*Fed-Norm-SGDA* 的迭代序列+ 满足

$$\min_{t\in[T]} \mathbb{E}\left\|\nabla \widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \leq \underbrace{\mathcal{O}\left(\frac{(\bar{\tau}/\tau_{\mathit{eff}})^{1/4}}{(\bar{\tau}PT)^{1/4}} + \frac{(\tau_{\mathit{eff}}P)^{1/4}}{T^{3/4}}\right)}_{\textit{Error with full synchronization}} + \underbrace{\mathcal{O}\left(\frac{C_w\sigma_L^2 + D(G_{\mathbf{x}}^2+\sigma_G^2)}{\bar{\tau}^2T^{3/4}}\right)}_{\textit{Local updates error}} + \underbrace{\mathcal{O}\left(\left(\frac{n-P}{n-1}\cdot\frac{E_w}{PT}\right)^{1/4}\right)}_{\textit{Partial participation error}}, \tag{7}$$

w此处 $\Phi_{1/2L_f}$是 Φ的 *Moreau* 包络。常数 C_w、 D、 $\bar{\tau}$在定理 *1*中定义。

~证明见附录 C 。定理 2 给出了代理目标函数 F的收敛性。接下来，我们来看真实目标函数 F的收敛性。

推论 2.1 （关于 F的收敛性）。 给定包络函数 $\Phi(\mathbf{x}) \triangleq$最大值$_{\mathbf{y}}\, F(\mathbf{x},\ \mathbf{y})$， $\widetilde{\Phi}(\mathbf{x}) \triangleq$最大值$_{\mathbf{y}}\, \widetilde{F}(\mathbf{x},\ \mathbf{y})$，在定理 *2* 的条件下，

$$\min_{t\in[T]} \left\|\nabla \Phi_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \leq \epsilon'_{opt} + \frac{8L_f^2}{T}\sum_{t=0}^{T-1}\left\|\tilde{\mathbf{x}}^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2,$$

其中， $\Phi_{1/2L_f}$是 Φ的 *Moreau* 包络， $\tilde{}x^{(t)} \triangleq arg\ min_{\mathbf{x}'}\{\widetilde{\Phi}(x') + L_f\|x' - x^{(t)}\|^2\}$， $\bar{}x^{(t)} \triangleq arg\ min_{\mathbf{x}'}\{\Phi(x') + L_f\|x' - x^{(t)}\|^2\}$，对所有 t均成立， ϵ'_{opt}是式（7）中的误差界。

与推论 1.1 类似，如果在算法 1 的服务器更新中，对所有 $i \in [n]$，将 $\{w_i\}$ 替换为 $\{p_i\}$，那么$\widetilde{F} \equiv F$，并且对所有 t， $\tilde{}\mathbf{x}^{(t)}$和 $\bar{}\mathbf{x}^{(t)}$相同。因此，定理 2 给出 $\min_{t\in[T]} \left\|\nabla \Phi_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \leq \epsilon'_{\text{opt}}$。注释4。对于非凸-凹函数，一些现有工作不要求假设 7，同时还提高了收敛速率。然而，这些方法要么采用双循环结构 Rafique et al. (2021); Zhang et al. (2022)，要么处理确定性问题 Xu et al. (2020); Zhang et al. (2020)。提出一种具备相同优势、用于随机NC-C问题的单循环方法仍是一个开放问题。

推论 2.2（改进的通信节省）。 假设所有客户端的权重相等（$p_i = 1/n$ 对所有 i均成立），且每个客户端执行 τ 步本地 *SGDA*。要达到一个 ϵ-驻点，即满足 $\mathbb{E}\|\nabla \Phi_{1/2L_f}(\mathbf{x})\| \leq \epsilon$ 的 $\mathbf{x}$，
• 在全参与的情况下，*Fed-Norm-SGDA+* 的每客户端梯度复杂度为 $T\tau = \mathcal{O}\left(\frac{1}{n\epsilon^8}\right)$。所需的通信轮次数为 $T = \mathcal{O}\left(\frac{1}{\epsilon^4}\right)$。• 在部分参与的情况下，当客户端间数据异质性 $\sigma_G = 0$ 这一特殊情形下，*Fed-Norm-SGDA* 的每客户端梯度复杂度为 $\mathcal{O}\left(1/\left(P\epsilon^8\right)\right)$，而通信成本为 $\mathcal{O}\left(1/\epsilon^4\right)$。

在通信需求方面，我们实现了巨大的节省（相比于 $\mathcal{O}(1/\epsilon^7)$，见 Sharma 等人(2022年)）。我们的梯度复杂度结果实现了参与客户端数量上的线性加速。

5.3 非凸-单点凹（NC-1PC）情形

在神经网络训练期间的随机梯度下降动态中，已观察到单点凸性。

假设 8 （关于$\mathbf{y}$的单点凹性）。 若固定$\mathbf{y}$ f，则称函数 $\in \mathbb{R}^{d_1}$关于$\mathbf{x}$ $\in \mathbb{R}^{d_2}$是单点凹的，对于所有$\mathbf{y}$ $\in \mathbb{R}^{d_2}$，

$$\langle \nabla_y f(\mathbf{x}, \mathbf{y}'), \mathbf{y} - \mathbf{y}^*(\mathbf{x})\rangle \leq f(\mathbf{x},\mathbf{y}) - f(\mathbf{x},\mathbf{y}^*(\mathbf{x})),$$

其中 $\mathbf{y}^*(\mathbf{x})$ $\in$最大值点集合$_{\mathbf{y}}\, f(\mathbf{x},\ \mathbf{y})$。

由于篇幅限制，我们仅给出在所有客户端权重相等（$p_i = 1/n$ 对所有 i）且每个客户端执行 τ 步本地 SGDA 时的每客户端梯度复杂度和通信复杂度结果。更多详情请参见附录C.4。

定理 *3.* 假设局部损失函数 $\{f_i\}$ 满足假设 *1*、*2*、*3*、*7*。假设对于所有 x，所有 f_i 都在共同的全局最小值点 $y^*(x)$ 处满足假设 *8*，并且对于所有 t，$\left\|y^{(t)}\right\|^2 \leq R$。那么，要达到一个 ϵ-精确点，*Fed-Norm-SGDA+*（算法 *1*）的随机梯度复杂度为 $\mathcal{O}(1/(n\epsilon^8))$，所需的通信轮次为 $T/\tau = \mathcal{O}(1/\epsilon^4)$。

备注 5. 定理 3 证明了Sharma 等人(2022年)提出的猜想，即单点凸函数应当能够实现线性加速。此外，我们将其通信复杂度从 $\mathcal{O}(1/\epsilon^7)$ 降低到 $\mathcal{O}(1/\epsilon^4)$。作为证明中的一个中间结果，我们证明了本地随机梯度下降在单点凸函数上的收敛性，将凸优化界推广到了范围更广的一类函数。

6 实验

本节评估所提出算法的实证性能。我们考虑鲁棒神经网络训练问题 Sinha 等人(2017)；Madry 等人(2018年)，以及公平分类问题 Mohri 等人(2019年)；Deng 等人(2020年)。受篇幅限制，实验的更多细节及一些补充结果收录于附录 D。我们的实验在由 $n = 15$ 个客户端组成的网络上运行，每个客户端均配备一块 NVIDIA TitanX GPU。我们使用Wang 等人(2019年)提出的参数为 α的狄利克雷分布$_n(\alpha)$来建模客户端间的数据异质性。较小的 $\alpha \Rightarrow$ 表示客户端间的异质性更高。

鲁棒神经网络训练。我们考虑以下鲁棒神经网络（NN）训练问题。

$$\min_{\mathbf{x}} \max_{\|\mathbf{y}\|^2 \leq 1} \sum_{j=1}^{N} \ell\left(h_{\mathbf{x}}(\mathbf{a}_i + \mathbf{y}), b_i\right), \tag{8}$$

其中，$\mathbf{x}$ 表示神经网络参数，(a_i, b_i)表示第 i个样本的特征和标签，$\mathbf{y}$ 表示对抗性添加的特征扰动，而 $h_{\mathbf{x}}$ 表示神经网络输出。

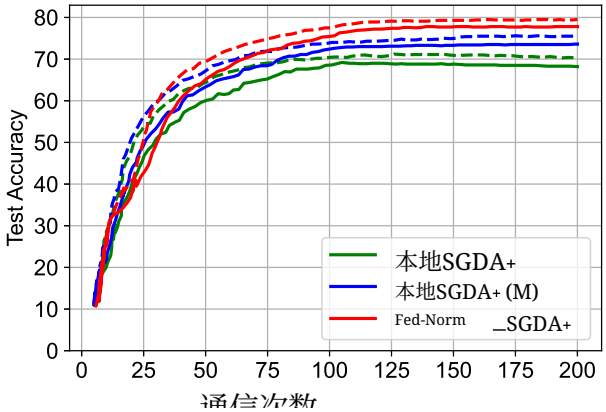

图3：异构数量的本地更新对 Fed-Norm-SGDA 的性能影响比较 $\{\tau_i\}$，+（算法1）、本地SGDA+以及采用动量的本地SGDA+，在使用 VGG11 模型求解 CIFAR10 数据集上的（8）时。实线（虚线）曲线对应于 $E = 5$（$E = 7$），以及 $\alpha = 0.1$。

系统异质性的影响。在图3中，我们比较了客户端之间本地更新数量异构对所提出的 Fed-Norm-SGDA 性能的影响+。我们将其与本地 SGDA+ Deng & Mahdavi(2021年) 以及带动量的本地 SGDA+ Sharma 等人(2022年) 进行比较。客户端通过 $\tau_i \sim Unif[2, E]$抽取各自在本地运行的训练轮次数。我们观察到，Fed-Norm-SGDA+能够很好地适应系统异质性，并优于这两种现有方法。

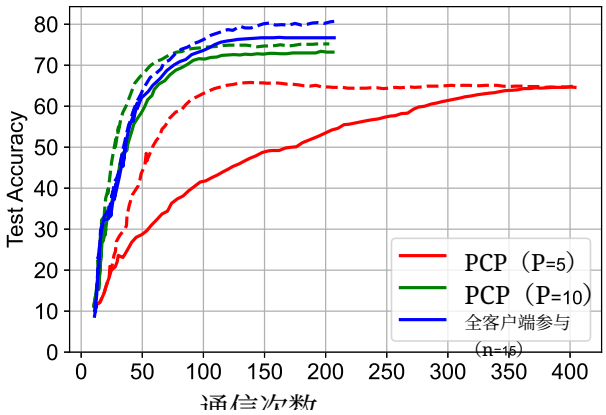

图4：在CIFAR10 数据集上使用VGG11 模型解决鲁棒神经网络训练问题时，比较客户端部分参与（PCP）对Fed-Norm-SGDA+性能的影响。图中展示了鲁棒测试准确率。实线（虚线）曲线对应 $\alpha = 0.1$ （$\alpha = 1.0$）。

部分参与与异质性的影响。 接下来，我们比较不同程度的客户端部分参与对性能的影响。我们将完全参与设置（$n = 15$）与 $P = 5$、10进行比较。客户端通过 $\tau_i \sim Unif[2, 5]$对本地训练轮次数量进行采样。我们绘制了数据异质性参数 $\alpha = 0.1$、1。0两种不同取值下的结果。正如我们所有理论结果所示，部分参与是收敛误差中最重要的组成部分， P取值越小，性能损失越大。此外，客户端间异质性越高（由较小的 α值建模），性能越差。我们还在附录D中进一步探讨了 α 对性能的影响。

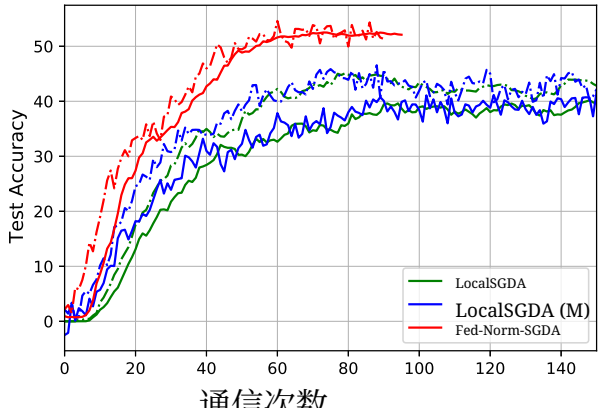

图5：在CIFAR10 数据集上使用VGG11 模型进行公平分类任务时，对比本地SGDA、带动量的本地SGDA和Fed-Norm-SGDA。实线（虚线）曲线对应 $E = 5$（$E = 7$）、 $\alpha = 0.1$。

公平分类。 我们考虑公平分类问题的极小极大问题形式Mohri 等人(2019年）； Nouiehed 等人 (2019年）。

$$\min_{\mathbf{x}} \max_{\mathbf{y} \in \mathcal{Y}} \sum_{c=1}^{C} y_c F_c(\mathbf{x}) - \frac{\lambda}{2} \|\mathbf{y}\|^2, \tag{9}$$

其中，$\mathbf{x}$ 表示神经网络的参数， $\{F_c\}_{c=1}^{C}$ 表示类别 c对应的损失函数， Δ_C 是 C维概率单纯形。在图5中，我们绘制了 Fed-Norm-SGDA、Local SGDA、Deng & Mahdavi (2021年) 以及带动量的 Local SGDA Sharma 等人(2022年) 所达到的最差分布测试准确率。与图3中一样，客户端采样 $\tau_i \sim Unif[2, E]$。我们绘制了每种情况下最差分布上的测试准确率。再次表明，Fed-Norm-SGDA 优于现有方法。

7 结论

在这项工作中，我们研究了联邦设置下的非凸极小极大问题。除了客户端间数据异质性和客户端部分参与外，还存在系统异质性。我们发现，现有方法（例如 Local SGDA）可能会收敛到与原定目标函数截然不同的目标函数的驻点。我们证明，归一化单个客户端的贡献可以解决这一问题。我们分析了几类非凸极小极大函数，并显著改进了现有的计算复杂度和通信复杂度结果。未来可能的研究方向包括分析客户端在线状态不可预测的联邦系统Yang 等人(2022b)。

致谢

本研究部分由美国国家科学基金会资助，项目编号为 CCF 2045694、CNS-2112471，以及美国海军研究办公室项目 N00014-23-1-2149。李佳睿协助制作了论文中的部分图表。

参考文献

Arjovsky, M.、Chintala, S. 和 Bottou, L. Wasserstein生成对抗网络。载于国际机器学习会议，第 214–223 页。PMLR，2017年。

Beznosikov, A.、Richt´arik, P.、Diskin, M.、Ryabinin, M. 和 Gasnikov, A. 采用压缩通信的变分不等式求解分布式方法及其理论保证。*arXiv*预印本 *arXiv:2110.03313*，2021a。

Beznosikov, A.、Rogozin, A.、Kovalev, D. 和 Gasnikov, A. 时变网络上求解鞍点问题的近最优去中心化算法。载于国际优化与应用会议，第 246–257 页。Springer，2021b。

Beznosikov, A., Scutari, G., Rogozin, A., and Gasnikov, A. 相似性条件下的分布式鞍点问题。载于神经信息处理系统进展，第34卷，2021年c。

Beznosikov, A.、Sushko, V.、Sadiev, A. 和 Gasnikov, A. 个性化去中心化联邦极小极大问题。*arXiv* 预印本 *arXiv:2106.07289*，2021年。Chen, M.、Shlezinger, N.、Poor, H. V.、Eldar, Y. C. 和 Cui, S. 通信高效的联邦学习。美国国家科学院院刊，118(17):e2024789118，2021年。Chen, Z.、Zhou, Y.、Xu, T. 和 Liang, Y. 近端梯度下降-上升法：基于 kł 几何的变量收敛。载于国际学习表征会议，2020年。

Cho, H. 和 Yun, C. 带随机打乱的 SGDA：非凸-p {\L} 极小极大优化的更快收敛。*arXiv* 预印本 *arXiv:2210.05995*，2022年。Davis, D. 和 Drusvyatskiy, D. 弱凸函数的随机模型化最小化。*SIAM* 优化杂志，29(1):207–239，2019年。Deng, Y. 和 Mahdavi, M. 本地随机梯度下降上升：收敛性分析与通信效率。载于国际人工智能与统计会议，第 1387–1395 页。PMLR，2021年。Deng, Y.、Kamani, M. M. 和 Mahdavi, M. 分布鲁棒联邦平均。载于神经信息处理系统进展，第 33 卷，第 15111–15122 页，2020年。Drusvyatskiy, D. 和 Paquette, C. 最小化凸函数与光滑映射复合函数的效率。数学规划，178(1):503–558，2019年。Goodfellow, I.、Pouget-Abadie, J.、Mirza, M.、Xu, B.、Warde-Farley, D.、Ozair, S.、Courville, A. 和 Bengio, Y. 生成对抗网络。载于神经信息处理系统进展，第 27 卷，2014年。Haddadpour, F.、Kamani, M. M.、Mahdavi, M. 和 Cadambe, V. 带周期性平均的本地随机梯度下降：更紧致的分析与自适应同步。神经信息处理系统进展，32:11082–11094，2019年。Hamer, J.、Mohri, M. 和 Suresh, A. T. Fedboost：一种通信高效的联邦学习算法。载于国际机器学习会议，第 3973–3983 页。PMLR，2020年。Hou, C.、Thekumparampil, K. K.、Fanti, G. 和 Oh, S. 联邦鞍点优化的高效算法。*arXiv* 预印本 *arXiv:2102.06333*，2021年。Jhunjhunwala, D.、Sharma, P.、Nagarkatti, A. 和 Joshi, G. FedVARP：解决联邦学习中客户端部分参与导致的方差问题。载于第*38*届人工智能不确定性会议，2022年。Jin, C.、Netrapalli, P. 和 Jordan, M. 非凸非凹极小极大优化中的局部最优性是什么？载于国际机器学习会议，第 4880–4889 页。PMLR，2020年。Kairouz, P.、McMahan, H. B.、Avent, B.、Bellet, A.、Bennis, M.、Bhagoji, A. N.、Bonawitz, K.、Charles, Z.、Cormode, G.、Cummings, R. 等。联邦学习的进展与开放问题。*arXiv* 预印本 *arXiv:1912.04977*，2019年。Karimi, H.、Nutini, J. 和 Schmidt, M. 满足 Polyak-Łojasiewicz 条件时梯度方法和近端梯度方法的线性收敛。载于欧洲机器学习与数据库知识发现联合会议，第 795–811 页。Springer，2016年。Khaled, A.、Mishchenko, K. 和 Richt´arik, P. 关于同质与异质数据上本地随机梯度下降的更紧理论。载于国际人工智能与统计会议，第 4519–4529 页。PMLR，2020年。Koloskova, A.、Loizou, N.、Boreiri, S.、Jaggi, M. 和 Stich, S. 具有变化拓扑和本地更新的去中心化随机梯度下降统一理论。载于国际机器学习会议，第 5381–5393 页。PMLR，2020年。Koneˇcn`y, J.、McMahan, H. B.、Ramage, D. 和 Richt´arik, P. 联邦优化：面向端侧智能的分布式机器学习。*arXiv* 预印本 *arXiv:1610.02527*，2016年。Lee, S. 和 Kim, D. 光滑结构化非凸非凹极小极大问题的快速外梯度方法。载于神经信息处理系统进展，第 34 卷，2021年。Lei, Y.、Yang, Z.、Yang, T. 和 Ying, Y. 极小极大问题中随机梯度方法的稳定性与泛化。载于国际机器学习会议，第 6175–6186 页。PMLR，2021年。Li, H.、Tian, Y.、Zhang, J. 和 Jadbabaie, A. 非凸强凹极小极大优化的复杂度下界。载于神经信息处理系统进展，第 34 卷，2021年。

Li, T., Sahu, A. K., Talwalkar, A., 和 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。Liao, L., Shen, L., Duan, J., Kolar, M., 和 Tao, D. 用于随机凸-凹极小极大问题的局部 Adagrad 型算法。*arXiv* 预印本 *arXiv:2106.10022*，2021年。Lin, T., Jin, C., 和 Jordan, M. 关于非凸-凹极小极大问题的梯度下降上升。在国际机器学习会议，第 6083–6093 页。PMLR，2020a。Lin, T., Jin, C., 和 Jordan, M. I. 极小极大优化的近乎最优算法。在学习理论会议，第 2738–2779 页。PMLR，2020b。Liu, W., Mokhtari, A., Ozdaglar, A., Pattathil, S., Shen, Z., 和 Zheng, N. 一种用于鞍点问题的去中心化近端点型方法。*arXiv* 预印本 *arXiv:1910.14380*，2019年。Lu, S., Tsaknakis, I., 和 Hong, M. 非凸极小极大问题的分块交替优化：信号处理与通信中的算法和应用。在*ICASSP 2019-2019 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 4754–4758 页。IEEE，2019年。Lu, S., Tsaknakis, I., Hong, M., 和 Chen, Y. 用于单侧非凸极小极大问题的混合分块逐次逼近：算法与应用。*IEEE* 信号处理汇刊，68:3676–3691，2020年。Luo, L. 和 Chen, C. 寻找非凸-强凹极小极大问题的二阶驻点。*arXiv* 预印本 *arXiv:2110.04814*，2021年。Luo, L., Ye, H., Huang, Z., 和 Zhang, T. 用于随机非凸-强凹极小极大问题的随机递归梯度下降上升。在神经信息处理系统进展，第 33 卷，第 20566–20577 页，2020年。Luo, L., Xie, G., Zhang, T., 和 Zhang, Z. 用于有限和非平衡凸-凹极小极大优化的近乎最优随机算法。*arXiv* 预印本 *arXiv:2106.01761*，2021年。Madras, D., Creager, E., Pitassi, T., 和 Zemel, R. 学习对抗公平且可迁移的表示。在国际机器学习会议，第 3384–3393 页。PMLR，2018年。Madry, A., Makelov, A., Schmidt, L., Tsipras, D., 和 Vladu, A. 迈向能够抵御对抗攻击的深度学习模型。在国际学习表征会议，2018年。McMahan, B., Moore, E., Ramage, D., Hampson, S., 和 y Arcas, B. A. 基于去中心化数据的深度网络通信高效学习。在人工智能与统计，第 1273–1282 页。PMLR，2017年。Metelev, D., Rogozin, A., Gasnikov, A., 和 Kovalev, D. 强凸性与强凹性常数不同的去中心化鞍点问题。*arXiv* 预印本 *arXiv:2206.00090*，2022年。Mitra, A., Jaafar, R., Pappas, G. J., 和 Hassani, H. 联邦学习中的线性收敛：应对客户端异质性与稀疏梯度。神经信息处理系统进展，34:14606–14619，2021年。Mohajerin Esfahani, P. 和 Kuhn, D. 使用 Wasserstein 度量的数据驱动分布鲁棒优化：性能保证与可处理的重构。数学规划，171(1):115–166，2018年。Mohri, M., Sivek, G., 和 Suresh, A. T. 无先验联邦学习。在国际机器学习会议，第 4615–4625 页。PMLR，2019年。Namkoong, H. 和 Duchi, J. C. 使用 f-散度进行分布鲁棒优化的随机梯度方法。在神经信息处理系统进展，第 29 卷，2016年。Nesterov, Y. 凸优化讲义，第 137 卷。Springer，2018年。

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., 和 Razaviyayn, M. 使用迭代一阶方法求解一类非凸极小极大博弈。在神经信息处理系统进展，第 32 卷，第 14934–14942 页，2019年。Ouyang, Y. 和 Xu, Y. 凸凹双线性鞍点问题的一阶方法复杂度下界。数学规划，185(1):1–35，2021年。Qiu, S., Yang, Z., Wei, X., Ye, J., 和 Wang, Z. 用于光滑非线性时序差分学习的单时间尺度随机非凸凹优化。*arXiv* 预印本 *arXiv:2008.10103*，2020年。Rafique, H., Liu, M., Lin, Q., 和 Yang, T. 弱凸凹极小极大优化：可证明的算法及其在机器学习中的应用。*Optimization Methods and Software*，第 1–35 页，2021年。Reisizadeh, A., Farnia, F., Pedarsani, R., 和 Jadbabaie, A. 鲁棒联邦学习：仿射分布偏移情形。在神经信息处理系统进展，第 33 卷，第 21554–21565 页，2020年。Reisizadeh, A., Tziotis, I., Hassani, H., Mokhtari, A., 和 Pedarsani, R. 抗掉队者联邦学习：利用统计精度与系统异质性之间的相互作用。*IEEE*信息论选题领域期刊，2022年。Rogozin, A., Beznosikov, A., Dvinskikh, D., Kovalev, D., Dvurechensky, P., 和 Gasnikov, A. 用于鞍点问题的去中心化分布式优化。*arXiv* 预印本 *arXiv:2102.07758*，2021年。Sanjabi, M., Ba, J., Razaviyayn, M., 和 Lee, J. D. 关于使用正则化最优传输训练生成对抗网络的收敛性与鲁棒性。神经信息处理系统进展，31，2018年。Sattler, F., Wiedemann, S., M¨uller, K.-R., 和 Samek, W. 基于非独立同分布数据的鲁棒且通信高效的联邦学习。*IEEE*神经网络与学习系统汇刊，31(9):3400–3413，2019年。Sharma, P., Panda, R., Joshi, G., 和 Varshney, P. 联邦极小极大优化：改进的收敛性分析与算法。在国际机器学习会议，第 19683–19730 页。PMLR，2022年。Sinha, A., Namkoong, H., 和 Duchi, J. 基于原理的对抗训练实现可认证的分布鲁棒性。在国际学习表征会议，2017年。Stich, S. U. 本地随机梯度下降收敛快、通信少。在国际学习表征会议，2018年。Sun, Z. 和 Wei, E. 一种用于联邦极小极大学习且具有线性收敛性的通信高效算法。*arXiv* 预印本 *arXiv:2206.01132*，2022年。Thekumparampil, K. K., Jain, P., Netrapalli, P., 和 Oh, S. 用于光滑极小极大优化的高效算法。在神经信息处理系统进展，第 32 卷，2019年。Tran-Dinh, Q., Liu, D., 和 Nguyen, L. M. 用于非凸线性函数极小极大问题的混合方差缩减随机梯度下降算法。在神经信息处理系统进展，第 33 卷，第 11096–11107 页，2020年。Tziotis, I., Shen, Z., Pedarsani, R., Hassani, H., 和 Mokhtari, A. 抗掉队者个性化联邦学习。*arXiv* 预印本 *arXiv:2206.02078*，2022年。Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., 和 Khazaeni, Y. 采用匹配平均的联邦学习。在国际学习表征会议，2019年。Wang, J. 和 Joshi, G. 协作随机梯度下降：本地更新随机梯度下降算法设计与分析的统一框架。机器学习研究期刊，22(213):1–50，2021年。

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. 解决异构联邦优化中的目标不一致问题。载于神经信息处理系统进展，第 33 卷，第 7611–7623 页，2020年。Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., 等。联邦优化实用指南。*arXiv* 预印本 *arXiv:2107.06917*，2021年。Wang, Y. and Li, J. 凸-凹极小极大优化的改进算法。载于神经信息处理系统进展，第 33 卷，第 4800–4810 页，2020年。Woodworth, B. E., Patel, K. K., and Srebro, N. 异构分布式学习中的小批量与本地随机梯度下降。神经信息处理系统进展，33:6281–6292，2020年。Xie, G., Luo, L., Lian, Y., and Zhang, Z. 有限和凸-凹极小极大优化问题的复杂度下界。载于国际机器学习会议，第 10504–10513 页。PMLR，2020年。Xie, J., Zhang, C., Zhang, Y., Shen, Z., and Qian, H. 非凸PL极小极大问题的联邦学习框架。*arXiv* 预印本 *arXiv:2105.14216*，2021年。Xing, E. P., Ho, Q., Xie, P., and Wei, D. 大数据分布式机器学习的策略与原则。《工程》，2(2):179–195，2016年。Xu, Z., Zhang, H., Xu, Y., and Lan, G. 用于非凸-凹和凸-非凹极小极大问题的统一单循环交替梯度投影算法。*arXiv* 预印本 *arXiv:2006.02032*，2020年。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。Yang, H., Liu, Z., Zhang, X., and Liu, J. SAGDA：在联邦极小极大学习中实现 $\mathcal{O}(\epsilon^{-2})$ 通信复杂度。*arXiv* 预印本 *arXiv:2210.00611*，2022a。Yang, H., Zhang, X., Khanduri, P., and Liu, J. 无序联邦学习。载于 国际机器学习会议，第 25331–25363 页。PMLR，2022b。Yang, J., Zhang, S., Kiyavash, N., and He, N. 极小极大优化的催化剂框架。载于 神经信息处理系统进展，第 33 卷，第 5667–5678 页，2020年。Yang, J., Orvieto, A., Lucchi, A., and He, N. 无强凹性条件下更快的极小极大优化单循环算法。载于国际人工智能与统计会议，第 5485–5517 页。PMLR，2022c。Yoon, T. and Ryu, E. K. 光滑凸-凹极小极大问题的加速算法：梯度范数平方以 o (1/kˆ2) 的速率收敛。载于国际机器学习会议，第 12098–12109 页。PMLR，2021年。Yu, H., Jin, R., and Yang, S. 关于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，第 7184–7193 页。PMLR，2019年。Yun, C., Rajput, S., and Sra, S. 随机打乱下小批量与本地随机梯度下降：紧致收敛界及拓展。载于国际学习表征会议，2022年。Zhang, J., Xiao, P., Sun, R., and Luo, Z. 用于非凸-凹极小极大问题的单循环平滑梯度下降-上升算法。载于 神经信息处理系统进展，第 33 卷，第 7377–7389 页，2020年。Zhang, S., Yang, J., Guzm´an, C., Kiyavash, N., and He, N. 非凸-强凹极小极大优化的复杂度。载于人工智能不确定性会议，第 482–492 页。PMLR，2021年。Zhang, X., Aybat, N. S., and Gurbuzbalaban, M. Sapd+：用于非凸-凹极小极大问题的加速随机方法。*arXiv* 预印本 *arXiv:2205.15084*，2022年。Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. 使用非独立同分布数据的联邦学习。*arXiv* 预印本 *arXiv:1806.00582*，2018年。

目录

1 引言 **1**

2 相关工作 3
2.1 单客户端极小极大.2.2 分布式/联邦极小极大. 3
3

3 预备知识 **4**

4 异构联邦极小极大优化算法 **5**
4.1 Local SGDA 的局限性.4.2 提出的归一化联邦极小极大算法. 5
6

5 收敛结果 **8**
5.1 非凸-强凹（NC-SC）情形.85.2 非凸-凹（NC-C）情形.95.3 非凸单点凹情形（NC-1PC）.10

6 实验 **11**

7 结论 13

A 背景 **19**
A.1 客户端采用不同求解器时的梯度聚合.19A.2 辅助结果.20A.3 有用的不等式.20A.4 收敛速率比较.21

B 非凸强凹函数下 Fed-Norm-SGDA 的收敛性（定理 1） **21**
B.1 中间引理.22B.2 定理 1 的证明.24B.3 中间引理的证明.26B.4 辅助引理.37B.5 Polyak–Łojasiewicz（PL）条件下的收敛性.38

C Fed-Norm-SGDA 的收敛性+ 适用于非凸凹函数（定理 2） **39**
C.1 中间引理.39C.2 定理的证明2 .41C.3 中间引理的证明.46C.4 将结果推广至非凸单点凹函数（定理 3）
. 50

D 补充实验 51

附录

附录的安排如下。在第A节中，我们介绍了全文中用到的一些基本数学结果和不等式。在附录B中，我们证明了 Fed-Norm-SGDA（算法1）针对光滑非凸-强凹（以及非凸-PŁ）函数的非渐近收敛性，并推导出该算法达到 ϵ-驻点所需的梯度复杂度和通信成本。在附录C中，我们证明了 Fed-Norm-SGDA+ （算法1）针对光滑非凸-凹和非凸-单点凹函数的非渐近收敛性。最后，在附录D 中，我们给出了所进行的补充实验的详细信息。

A 背景

A.1 客户端使用不同求解器时的梯度聚合

本地SGDA。 假设 $\tau_i^{(t)} = \tau^{(t)}$ eff$= \tau$对所有 $i \in [n]$， $t \in [T]$。此外， $a_i^{(t,k)} = 1$对所有 $k \in [\tau]$， t。那么，算法 1-Fed-Norm-SGDA 中的本地迭代点更新可简化为（Fed-Norm-SGDA 的更新+ 类似）

$$\begin{aligned}\mathbf{x}_i^{(t,k+1)} &= \mathbf{x}_i^{(t,k)} - \eta_x^c \nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}),\\ \mathbf{y}_i^{(t,k+1)} &= \mathbf{y}_i^{(t,k)} + \eta_y^c \nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}),\end{aligned}$$

对于 $k \in \{0, \ldots, \tau - 1\}$ 且梯度聚合向量（$\mathbf{g}^{(t)}{}_{\mathbf{x}\,i}$， $\mathbf{g}^{(t)}{}_{\mathbf{y}\,i}$）只是各个梯度的平均值

$$\mathbf{g}_{\mathbf{x},i}^{(t)} = \frac{1}{\tau}\sum_{k=0}^{\tau-1} \nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}), \quad \mathbf{g}_{\mathbf{y},i}^{(t)} = \frac{1}{\tau}\sum_{k=0}^{\tau-1} \nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$$

请注意，这些恰好是 Deng & Mahdavi (2021年)； Sharma 等人(2022年)提出的 LocalSGDA 的迭代点；唯一的区别是，在 LocalSGDA 中，客户端将迭代点 $\{\mathbf{x}_i^{(t,\tau)}, \mathbf{y}_i^{(t,\tau)}\}$ 发送给服务器，由服务器对其取平均以计算 $\{\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}\}$。而在这里，客户端发送的是 $\{\mathbf{g}^{(t)}{}_{\mathbf{x}\,i}, \mathbf{g}^{(t)}{}_{\mathbf{y}\,i}\}$。此外，在 Fed-Norm-SGDA 中，客户端和服务器使用不同的学习率，因此局部更新误差的界更紧。

在本地更新中使用动量。 假设每个本地客户端都使用动量尺度为 ρ 的动量缓冲区。那么，对于 $k \in \{0, \ldots, \tau_i^{(t)} - 1\}$

$$\begin{aligned}\mathbf{d}_{\mathbf{x},i}^{t,k+1} &= \rho \mathbf{d}_{\mathbf{x},i}^{t,k} + \nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}), \quad \mathbf{x}_i^{(t,k+1)} = \mathbf{x}_i^{(t,k)} - \eta_x^c \mathbf{d}_{\mathbf{x},i}^{t,k+1}\\ \mathbf{d}_{\mathbf{y},i}^{t,k+1} &= \rho \mathbf{d}_{\mathbf{y},i}^{t,k} + \nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}), \quad \mathbf{y}_i^{(t,k+1)} = \mathbf{y}_i^{(t,k)} + \eta_y^c \mathbf{d}_{\mathbf{y},i}^{t,k+1},\end{aligned}$$

简单计算可知，梯度聚合向量 $(\mathbf{g}^{(t)}{}_{\mathbf{x}\,i}, \mathbf{g}^{(t)}{}_{\mathbf{y}\,i})$ 中 $\nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$ 和 $\nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$ 的系数为

$$\sum_{j\geq k}^{\tau_i^{(t)}-1} = 1 + \rho + \cdots + \rho^{\tau_i^{(t)}-1-k} = \frac{1-\rho^{\tau_i^{(t)}-k}}{1-\rho}.$$

因此，聚合向量为 $\bar{\boldsymbol{a}}_i^{(t)} = \frac{1}{1-\rho}[1-\rho^{\tau_i^{(t)}}, 1-\rho^{\tau_i^{(t)}-1}, \ldots, 1-\rho]$，并且

$$\|\bar{\boldsymbol{a}}_i^{(t)}\|_1 = \sum_{k=0}^{\tau_i^{(t)}-1} \frac{1-\rho^{\tau_i^{(t)}-k}}{1-\rho} = \frac{1}{1-\rho}\left[\tau_i^{(t)} - \frac{\rho(1-\rho^{\tau_i^{(t)}})}{1-\rho}\right].$$

A.2 辅助结果

备注 6　（异质性的影响 σ_G 即使有 $\tau = 1$）。考虑两个简单最小化问题：

$$\textbf{(P1):}\quad \min_{\mathbf{x}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \qquad \text{and} \qquad \textbf{(P2):}\quad \min_{\mathbf{x}} f(x).$$

(P1) 是一个简单的分布式最小化问题，包含 n 个客户端，我们使用同步分布式随机梯度下降来求解。在迭代 t 时，每个客户端 i 计算随机梯度 $\nabla f_i(\mathbf{x}^{(t)}; \xi_i^{(t)})$，并将其发送给服务器。服务器对这些梯度取平均，然后沿着方向 $\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}^{(t)}; \xi_i^{(t)})$迈出一步。另一方面，**(P1)** 是一个集中式最小化问题，在每次迭代 t 中，智能体计算一个批量大小为 n 的随机梯度估计量 $\frac{1}{n}\sum_{i=1}^{n} \nabla f(\mathbf{x}^{(t)}; \xi_i^{(t)})$。我们比较这两个全局梯度估计量的方差如下。

(P1)		**(P2)**
$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}^{(t)}; \xi_i^{(t)}) - \nabla f(\mathbf{x}^{(t)})\right\|^2$		$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} \nabla f(\mathbf{x}^{(t)}; \xi_i^{(t)}) - \nabla f(\mathbf{x}^{(t)})\right\|^2$
$\leq \frac{1}{n^2}\sum_{i=1}^{n}\left[\sigma_L^2 + \beta_L^2 \mathbb{E}\left\|\nabla f_i(\mathbf{x}^{(t)})\right\|^2\right]$	vs	$= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}\left\|\nabla f(\mathbf{x}^{(t)}; \xi_i^{(t)}) - \nabla f(\mathbf{x}^{(t)})\right\|^2$
$\leq \frac{\sigma_L^2}{n} + \frac{\beta_L^2}{n}\left[\beta_H^2 \mathbb{E}\left\|\nabla f(\mathbf{x}^{(t)})\right\|^2 + \sigma_G^2\right].$		$\leq \frac{\sigma_L^2}{n} + \frac{\beta_L^2}{n}\mathbb{E}\left\|\nabla f(\mathbf{x}^{(t)})\right\|^2.$

由于几乎所有现有研究都考虑局部方差界（假设 2），且 $\beta_L = 0$，同步分布式随机梯度下降 **(P1)** 和单智能体小批量随机梯度下降**(P2)** 中的全局梯度估计量具有相同的 $\frac{\sigma_L^2}{n}$ 方差界。因此，在大多数关于最小化问题 Wang 等人 (2020年)；Yang 等人 (2021年) 以及极小极大问题 Sharma 等人 (2022年) 的联邦学习研究中，完全同步误差仅取决于局部方差 σ_L^2。然而，如上所示，对于 $\beta_L > 0$，这种 表面上的等价性 不再成立。 Koloskova 等人 (2020年) 针对最小化问题考虑了与我们类似的局部方差假设，也显示出对异质性的类似依赖关系 σ_G。

A.3 有用的不等式

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

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

对于任意常数 $\gamma > 0$。

引理 A.2 (强凹性)。 若存在一个常数 $\mu > 0$，使得对于所有$\mathbf{x} \in \mathcal{X}$以及所有$\mathbf{y}$、$\mathbf{y}' \in \mathcal{Y}$，以下不等式成立，则函数 $g\colon \mathcal{X} \times \mathcal{Y}$ 关于 $\mathbf{y}$ 强凹。

$$g(\mathbf{x}, \mathbf{y}) \leq g(\mathbf{x}, \mathbf{y}') + \langle \nabla_y g(\mathbf{x}, \mathbf{y}'), \mathbf{y}' - \mathbf{y} \rangle - \frac{\mu}{2}\|\mathbf{y} - \mathbf{y}'\|^2.$$

引理 A.3　(詹森不等式)。 给定一个凸函数 f 和一个随机变量 X，以下结论成立 *s.*

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

引理 A.4 (平方和)。 对于正整数 K 和向量集合 $x_1, \ldots, x_K$，以下结论成立：

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

引理 A.5 (二次增长条件 Karimi 等人 (2016年))。 若函数 g 满足假设 *1*、*4*，则对所有 x，以下条件成立

$$g(x) - \min_z g(z) \geq \frac{\mu}{2}\|x_p - x\|^2,$$
$$\|\nabla g(x)\|^2 \geq 2\mu\left(g(x) - \min_z g(z)\right).$$

引理 A.6。 对于 L-光滑凸函数 g，以下不等式成立

$$\mathbb{E}\left\|\nabla g(\mathbf{y})-\nabla g(\mathbf{x})\right\|^2 \leq 2L\left[g(\mathbf{y})-g(\mathbf{x})-\nabla g(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})\right]. \tag{10}$$

引理 A.7 （Cho 和 Yun（2022年）中的命题 6）。 对于 L-光滑函数 g，其下界为 g^*，以下不等式对所有 $\mathbf{x}$

$$\mathbb{E}\left\|\nabla g(\mathbf{x})\right\|^2 \leq 2L\left[g(\mathbf{x})-g^*\right]. \tag{11}$$

A.4 收敛速率比较

表 2：不同类别非凸极小极大问题的收敛速率比较。n 表示客户端总数，而 P 表示客户端部分参与时每轮抽样的客户端数量。 T 表示通信轮次， τ 表示每个客户端的本地更新次数，σ_L^2， σ_G^2 分别表示随机梯度方差和全局异质性（假设 2、3）。 $\kappa = L_f/\mu$ 表示条件数。为与现有研究进行公平比较，本表中的结果特化到以下情形：所有客户端 (i) 权重相同（$p_i = 1/n$），(ii) 执行相同次数的本地更新（$\tau_i = \tau$），且 (iii) 使用相同的本地更新算法 SGDA。不过，当 (i)–(iii) 不成立时，我们的结果（第 4节）仍适用于更一般的设置。

Work	部分参与	系统异质性	收敛速率
非凸-强凹（NC-SC）/非凸-Polyak-Łojasiewicz（NC-PL）			
Sharma 等人(2022年)	✗	✗	$\mathcal{O}\left(\frac{1}{\sqrt{n\tau T}}+\frac{n\tau(\sigma_L^2+\sigma_G^2)}{T}\right)$
Yang 等人 (2022a)	✓	✗	$\left(\frac{(n-P)(\sigma_G^2)}{n}\sqrt{\frac{\tau}{PT}}+\frac{1}{\sqrt{P\tau T}}\right)+\frac{(\sigma_L^2+\tau\sigma_G^2)}{\tau T}$
我们的工作： 定理1	✓	✓	$\left(\frac{(n-P)\sigma_G^2(n-1)}{n}\sqrt{\frac{\tau}{PT}}+\frac{1}{\sqrt{P\tau T}}+\frac{\sigma_L^2+\tau\sigma_G^2}{\tau T}\right)$
非凸-凹（NC-C）			$\mathcal{O}\left(\frac{1}{\tau nT()^{1/4}}\right)+\mathcal{O}\left(\frac{(n\tau)^{3/2}}{\sqrt{T}}\right)$
Sharma 等人(2022年)	✗	✗	$\mathcal{O}\left(\sqrt[4]{\frac{n-P}{n-1}\frac{1}{()PT}}+\frac{1}{\tau PT0^{1/4}}+\frac{(P\tau)^{1/4}}{T^{3/4}}\right)$
我们的工作： 定理2	✓	✓	
非凸-单点凹（NC-1PC）			$\mathcal{O}\left(\frac{1}{\tau T()^{1/4}}\right)+\mathcal{O}\left(\frac{\tau^{3/2}}{\sqrt{T}}\right)$
Sharma 等人（2022年）	✗	✗	$\mathcal{O}\left(\sqrt[4]{\frac{n-P}{n-1}\frac{1}{()PT}}+\frac{1}{\tau PT0^{1/4}}+\frac{(P\tau)^{1/4}}{T^{3/4}}\right)$
我们的工作： 定理3	✓	✓	

B Fed-Norm-SGDA 在非凸-强凹函数上的收敛性（定理1）

本节安排如下。首先，在附录 B.1 中，我们给出一些中间结果，并利用它们证明主要定理。接下来，在附录 B.2 中，我们给出定理 1 的证明，随后在附录 B.3 中证明这些中间结果。附录 B.4 包含一些辅助结果。最后，在附录 B.5 中，我们讨论非凸 PL 函数的收敛结果。

我们要解决的问题是

$$\min_{\mathbf{x}}\max_{\mathbf{y}}\left\{\widetilde{F}(\mathbf{x},\mathbf{y}) \triangleq \sum_{i=1}^{n} w_i f_i(\mathbf{x},\mathbf{y})\right\}.$$

我们定义 $\widetilde{\Phi}(\mathbf{x}) \triangleq \max_{\mathbf{y}} \widetilde{F}(\mathbf{x}, \mathbf{y})$ 和 $\tilde{}\mathbf{y}^*(\mathbf{x}) \in \arg\max_{\mathbf{y}} \widetilde{F}(\mathbf{x}, \mathbf{y})$。由于 $\widetilde{F}(\mathbf{x}\ \cdot)$ 是 μ-强凹的，$\tilde{}\mathbf{y}^*(\mathbf{x})$ 是唯一的。在 Fed-Norm-SGDA（算法 1）中，客户端更新如下：

$$\begin{aligned}\mathbf{x}_i^{(t,k)} &= \mathbf{x}^{(t)} - \eta_x^c \sum_{j=0}^{k-1} a_i^{(j)}(k) \nabla_x f_i(\mathbf{x}_i^{(t,j)}, \mathbf{y}_i^{(t,j)}; \xi_i^{(t,j)}), \\ \mathbf{y}_i^{(t,k)} &= \mathbf{y}^{(t)} + \eta_y^c \sum_{j=0}^{k-1} a_i^{(j)}(k) \nabla_y f_i(\mathbf{x}_i^{(t,j)}, \mathbf{y}_i^{(t,j)}; \xi_i^{(t,j)}),\end{aligned} \tag{12}$$

其中 $1 \le k \le \tau_i$。随后对这些客户端更新进行聚合，以计算 $\{\mathbf{g}^{(t)}{}_{\mathbf{x}\,i},\ \mathbf{g}^{(t)}{}_{\mathbf{y}\,i}\}$

$$\begin{aligned}\mathbf{g}_{\mathbf{x},i}^{(t)} &= \frac{1}{\|\boldsymbol{a}_i\|} \sum_{k=0}^{\tau_i - 1} a_i^{(k)}(\tau_i) \nabla_x f_i \left(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}\right); \quad \mathbf{h}_{\mathbf{x},i}^{(t)} = \frac{1}{\|\boldsymbol{a}_i\|} \sum_{k=0}^{\tau_i - 1} a_i^{(k)}(\tau_i) \nabla_x f_i \left(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}\right) \\ \mathbf{g}_{\mathbf{y},i}^{(t)} &= \frac{1}{\|\boldsymbol{a}_i\|} \sum_{k=0}^{\tau_i - 1} a_i^{(k)}(\tau_i) \nabla_y f_i \left(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}\right); \quad \mathbf{h}_{\mathbf{y},i}^{(t)} = \frac{1}{\|\boldsymbol{a}_i\|} \sum_{k=0}^{\tau_i - 1} a_i^{(k)}(\tau_i) \nabla_y f_i \left(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}\right).\end{aligned}$$

备注7。请注意，在上文的 $a_i^{(j)}(k)$ 中，我们明确写出了其对 k 的依赖关系。为避免繁琐的记号，主论文中省略了这一点。然而，对于某些本地优化器，例如客户端处的局部动量（附录 A.1），系数 $a_i^{(j)}(k)$ 会随 k 变化。在后续分析中，我们假设对于所有 $j \in \{0, 1, \ldots, k-1\}$ 以及所有 $k \in \{1, 2, \ldots, \tau_i\}$，$a_i^{(j)}(k) \le \alpha$。此外，我们假设对于可行的 k，$\|\boldsymbol{a}_i(k)\|_1 \le \|\boldsymbol{a}_i(k)\|_1$，且 $\|\boldsymbol{a}_i(k+1)\|_2 \le \|\boldsymbol{a}_i(k+1)\|_2$。我们还使用记号 $\|\boldsymbol{a}_i\| \triangleq \|\boldsymbol{a}_i(\tau_i)\|$。

在第 t次迭代中，服务器均匀随机地进行 $|\mathcal{C}^{(t)}|$ 客户端无放回（**WOR**）抽样。在服务器进行聚合时，客户端 i 的更新按 $\tilde{w}_i = w_i n / |\mathcal{C}^{(t)}|$加权。服务器计算得到的聚合结果（$\mathbf{g}_{\mathbf{x}}^{(t)}$，$\mathbf{g}_{\mathbf{y}}^{(t)}$）形式如下：

$$\begin{aligned}\mathbf{g}_{\mathbf{x}}^{(t)} &= \sum_{i \in \mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t)}, \quad \text{such that} \quad \mathbb{E}_{\mathcal{C}^{(t)}}[\mathbf{g}_{\mathbf{x}}^{(t)}] = \mathbb{E}_{\mathcal{C}^{(t)}}\left[\sum_{i=1}^{n} \mathbb{I}(i \in \mathcal{C}^{(t)}) \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t)}\right] = \sum_{i=1}^{n} w_i \mathbf{g}_{\mathbf{x},i}^{(t)} \\ \mathbf{g}_{\mathbf{y}}^{(t)} &= \sum_{i \in \mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{y},i}^{(t)}, \quad \text{such that} \quad \mathbb{E}_{\mathcal{C}^{(t)}}[\mathbf{g}_{\mathbf{y}}^{(t)}] = \mathbb{E}_{\mathcal{C}^{(t)}}\left[\sum_{i=1}^{n} \mathbb{I}(i \in \mathcal{C}^{(t)}) \tilde{w}_i \mathbf{g}_{\mathbf{y},i}^{(t)}\right] = \sum_{i=1}^{n} w_i \mathbf{g}_{\mathbf{y},i}^{(t)}\end{aligned} \tag{13}$$

为便于分析，除非另有说明，我们假设 $|\mathcal{C}^{(t)}| = P$对所有 t均成立。最后，服务器按如下方式更新$\mathbf{x}$、$\mathbf{y}$变量：

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \tau_{\text{eff}} \gamma_x^s \mathbf{g}_{\mathbf{x}}^{(t)}, \qquad \mathbf{y}^{(t+1)} = \mathbf{y}^{(t)} + \tau_{\text{eff}} \gamma_y^s \mathbf{g}_{\mathbf{y}}^{(t)}.$$

B.1 中间引理

~我们先给出 Nouiehed 等人 （2019年）关于 $\Phi(\cdot)$ 的光滑性的以下结果。

引理B.1。 如果函数 $f(\cdot, \cdot)$ 满足假设*1*、*4* （L_f-光滑性且 μ-强凹性（关于$\mathbf{y}$），则$\phi(\cdot) \triangleq \max_{\mathbf{y}} f(\cdot, \mathbf{y})$ 是 L_Φ-光滑的，其中 $L_\Phi = \kappa L_f / 2 + L_f$，$\kappa = L_f / \mu$为条件数。

引理 B.2. 假设客户端本地损失函数 $\{f_i\}$ 满足假设 *1*、*3*，且局部函数的随机预言机满足假设 *2*。假设服务器每轮选择 P个客户端。那么，由以下方法生成的迭代序列

Fed-Norm-SGDA（算法 *1*）满足

$$\begin{aligned}\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{(t)}\right\|^2 &= \mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{g}_{\mathbf{x},i}^{(t)}\right\|^2 \\ &\le \frac{n}{P}\left(\frac{P-1}{n-1}\right)\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 + \frac{n}{P}\sum_{i=1}^n\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\left[\sigma_L^2+\beta_L^2\mathbb{E}\left\|\nabla_x f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)})\right\|^2\right] \\ &\quad + \frac{n(n-P)}{n-1}\left[\frac{2L_f^2}{P}\sum_{i=1}^n\frac{w_i^2}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)+(\max_i w_i)\frac{2}{P}\left(\beta_G^2\left\|\nabla_x\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2+\sigma_G^2\right)\right],\end{aligned} \tag{14}$$

其中，$\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \triangleq \mathbb{E}\left[\|\mathbf{x}_i^{(t,k)}-\mathbf{x}^{(t)}\|^2+\|\mathbf{y}_i^{(t,k)}-\mathbf{y}^{(t)}\|^2\right]$ 是客户端 i 在第t个通信轮次的本地迭代 k 处的迭代漂移。

引理 *B.3*。假设客户端本地损失函数 $\{f_i\}$ 满足假设 *1*、*3*、*4*，且本地函数的随机预言机满足假设 *2*。此外，服务器学习率 γ_x^s 满足 $64\tau_{\mathit{eff}}\gamma_x^s L_\Phi\beta_L^2\beta_G^2\frac{n}{P}(\max_i w_i\|\boldsymbol{a}_i\|_2^2/\|\boldsymbol{a}_i\|_1^2) \le 1$, $8\tau_{\mathit{eff}}\gamma_x^s L_\Phi(\max_i w_i)\frac{n}{P}\binom{n-P}{n-1}\max\{8\beta_G^2,1\} \le 1$，且 $8\tau_{\mathit{eff}}\gamma_x^s L_\Phi\beta_L^2\frac{n}{P}(\max_{i,k} w_i a_i^{(k)}(\tau_i)/\|\boldsymbol{a}_i\|_1) \le 1$。那么，算法 *1* 生成的迭代序列满足

$$\begin{aligned}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)})-\widetilde{\Phi}(\mathbf{x}^{(t)})\right] &\le -\frac{7\tau_{\mathit{eff}}\gamma_x^s}{16}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 - \frac{\tau_{\mathit{eff}}\gamma_x^s}{2}\left(1-\frac{n(P-1)}{P(n-1)}\tau_{\mathit{eff}}\gamma_x^s L_\Phi\right)\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \\ &\quad + \frac{5}{4}\tau_{\mathit{eff}}\gamma_x^s L_f^2\sum_{i=1}^n\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) + \frac{9\tau_{\mathit{eff}}\gamma_x^s L_f^2}{4\mu}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right] \\ &\quad + \frac{\tau_{\mathit{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+\sigma_G^2\left(2(\max_i w_i)\frac{n-P}{n-1}+2\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right].\end{aligned} \tag{15}$$

备注。公式（15）中的界与简单光滑最小化问题对应的单步衰减界非常相似。主要区别在于存在 $\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]$，它量化了 $\mathbf{y}^{(t)}$ 在求解最大化问题 $\max_{\mathbf{y}}\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y})$ 时的不准确程度。项 $\sum_{i=1}^n\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)$ 是客户端漂移，并在引理 B.4 中给出了界。

引理 B.4。假设局部损失函数 $\{f_i\}$ 满足假设*1*、*3*、*4*，且局部函数的随机预言机满足假设 *2*。此外，在算法 *1* 中，我们选择学习率 η_x^c，η_y^c 使得 $\max\{\eta_x^c,\eta_y^c\} \le \frac{1}{2L_f\max_i\|\boldsymbol{a}_i\|_1\sqrt{2(1+\beta_L^2)}}$。然后，由 *Fed-Norm-SGDA*（算法 *1*）生成的迭代序列 $\{\mathbf{x}_i^{(t)}、\mathbf{y}_i^{(t)}\}$ 满足

$$\begin{aligned}L_f^2\sum_{i=1}^n\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) &\le 2\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\sigma_L^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2 + 4L_f^2M_{\mathbf{a}_{-1}}\left([\eta_x^c]^2+[\eta_y^c]^2\right)\sigma_G^2 \\ &\quad + 8L_f^2M_{\mathbf{a}_{-1}}\beta_G^2[\eta_x^c]^2\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 + 8L_f^3M_{\mathbf{a}_{-1}}\beta_G^2\left(2\kappa[\eta_x^c]^2+[\eta_y^c]^2\right)\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right],\end{aligned}$$

其中 $M_{\mathbf{a}_{-1}} \triangleq$ 最大值 $_i\left(\|\boldsymbol{a}_{i,-1}\|_1^2+\beta_L^2\|\boldsymbol{a}_{i,-1}\|_2^2\right)$。

引理 B.5。假设局部损失函数 $\{f_i\}$ 满足假设*1*、*3*、*4*，且局部函数的随机预言机满足假设 *2*。服务器学习率 γ_x^s，γ_y^s 满足以下条件：

$$2\tau_{\mathit{eff}}\gamma_y^s L_f \le 1, \tau_{\mathit{eff}}\gamma_y^s\kappa L_f\beta_G^2\frac{n}{P}\max\left\{\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}, \frac{n-P}{n-1}\max_i w_i\right\} \le \frac{1}{64}, \gamma_x^s \le \frac{\gamma_y^s}{156\kappa^2},$$

$8\tau_{\mathit{eff}}\gamma_x^s L_\Phi \beta_L^2 \frac{n}{P}$最大 $\left\{\text{最大}_{i,\ k} \frac{w_i a_i^{(k)}}{(\tau_i)\|\boldsymbol{a}_i\|_1},\ \frac{n-P}{n-1}\text{最大}_i\ w_i\right\} \le 1$，$\tau_{\mathit{eff}}\gamma_x^s L_f \beta_G \sqrt{\frac{n}{P}}$最大 $\left\{\frac{n-P}{n-1}\text{最大}_i \frac{w_i}{\beta_L\sqrt{\text{最大}_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}}}\right\} \le \frac{1}{40\kappa}$客户端学习率 η_x^c，η_y^c满足 $\eta_y^c L_f \beta_G \le \frac{1}{16\sqrt{\kappa M_{\mathbf{a}_{-1}}}}$且 η_x^c：$\eta_x^c L_f \beta_G \le \frac{1}{64\kappa\sqrt{M_{\mathbf{a}_{-1}}}}$，分别。然后，由 *Fed-Norm-SGDA*（算法 *1*）生成的迭代序列满足

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right] \\
&\le \frac{4\left[\widetilde{\Phi}(\mathbf{x}^{(0)}) - \widetilde{F}(\mathbf{x}^{(0)},\mathbf{y}^{(0)})\right]}{\tau_{\mathit{eff}}\gamma_y^s\mu T} + \frac{1}{12\mu\kappa^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 + \frac{8\tau_{\mathit{eff}}[\gamma_x^s]^2 L_\Phi}{\gamma_y^s\mu}\frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \\
&\quad + 6\tau_{\mathit{eff}}\gamma_y^s\kappa\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + 2\sigma_G^2\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right] \\
&\quad + 18\kappa L_f\left([\eta_x^c]^2 + [\eta_y^c]^2\right)\left[\sigma_L^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2 + 2\sigma_G^2 M_{\mathbf{a}_{-1}}\right] \\
&\quad + \frac{8\tau_{\mathit{eff}}[\gamma_x^s]^2\kappa}{\gamma_y^s}\frac{n}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\left([\eta_x^c]^2 + [\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2 + 2\sigma_G^2 M_{\mathbf{a}_{-1}}\right].
\end{aligned}
\tag{16}
$$

B.2 定理的证明1

为完整起见，我们先在此给出定理1 的完整表述。

定理。假设局部损失函数 $\{f_i\}_i$ 满足假设*1*、*2*、*3*、*4*。假设服务器采用无放回抽样方案***(WOR)***选择客户端。此外，服务器学习率 γ_x^s、γ_y^s 以及客户端学习率η_x^c、η_y^c 满足引理*B.5*中规定的条件。那么，*Fed-Norm-SGDA*（算法*1*）生成的迭代序列满足

$$
\begin{aligned}
\min_{t\in[0:T-1]}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 \le \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 &\le \underbrace{\mathcal{O}\left(\kappa^2\left[\frac{\Delta_{\widetilde{\Phi}}}{\tau_{\mathit{eff}}\gamma_y^s T} + \frac{\gamma_y^s L_f}{P}\left(A_w\sigma_L^2 + B_w\beta_L^2\sigma_G^2\right)\right]\right)}_{\textit{Error with full synchronization}} \\
&\quad + \underbrace{\mathcal{O}\left(\kappa^2\left([\eta_x^c]^2 + [\eta_y^c]^2\right)L_f^2\left[C_w\sigma_L^2 + D\sigma_G^2\right]\right)}_{\textit{Error due to local updates}} + \underbrace{\mathcal{O}\left(\kappa^2\frac{n-P}{n-1}\frac{\gamma_y^s L_f E_w\tau_{\mathit{eff}}\sigma_G^2}{P}\right)}_{\textit{Partial Participation Error}},
\end{aligned}
$$

其中 $\kappa = L_f/\mu$ 为条件数，$\widetilde{\Phi}(x) \triangleq \max_{\mathbf{y}} \widetilde{F}(x,y)$ 为包络函数，$\Delta_{\widetilde{\Phi}} \triangleq \widetilde{\Phi}(x^{(0)}) - \min_{\mathbf{x}}\widetilde{\Phi}(x)$，$A_w \triangleq n\tau_{\mathit{eff}}\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$，$B_w \triangleq n\tau_{\mathit{eff}}\max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$，$C_w \triangleq \sum_{i=1}^n w_i(\|\boldsymbol{a}_i\|_2^2 - [\alpha^{(t,\tau_i-1)}{}_i\ \ ]^2)$，$D \triangleq \max_i(\beta_L^2\|\boldsymbol{a}_{i,-1}\|_2^2 + \|\boldsymbol{a}_{i,-1}\|_1^2)$，其中 $\boldsymbol{a}_{i,-1} \triangleq [a_i^{(0)}, a_i^{(1)}, \ldots, a^{(\tau_i-2)}{}_i\ \ ]^\top$ 对所有 i 以及 $E_w \triangleq n\max_i w_i$均成立。

利用上述界中的 $\gamma_y^s = \Theta\left(\frac{1}{L_f}\sqrt{\frac{P}{\bar{\tau}T}}\right)$ 和 $\eta_x^c \le \eta_y^c = \Theta\left(\frac{1}{L_f\bar{\tau}\sqrt{T}}\right)$（其中 $\bar{\tau} = \frac{1}{n}\sum_{i=1}^n\tau_i$），可得

$$
\min_{t\in[T]}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 \le \underbrace{\mathcal{O}\left(\kappa^2\frac{(\bar{\tau}/\tau_{\mathit{eff}}) + A_w\sigma_L^2 + B_w\beta_L^2\sigma_G^2}{\sqrt{P\bar{\tau}T}}\right)}_{\textit{Error with full synchronization}} + \underbrace{\mathcal{O}\left(\kappa^2\frac{C_w\sigma_L^2 + D\sigma_G^2}{\bar{\tau}^2 T}\right)}_{\textit{Local updates error}} + \underbrace{\mathcal{O}\left(\frac{n-P}{n-1}\cdot\frac{\kappa^2 E_w\tau_{\mathit{eff}}\sigma_G^2}{\sqrt{P\bar{\tau}T}}\right)}_{\textit{Partial participation error}}.
$$

证明。利用引理 B.3，并代入引理 B.4给出的迭代量的漂移界，我们可以得到以下界

$$
\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)}) - \widetilde{\Phi}(\mathbf{x}^{(t)})\right] \le -\frac{7\tau_{\text{eff}}\gamma_x^s}{16}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 - \frac{\tau_{\text{eff}}\gamma_x^s}{2}\left(1 - \frac{n(P-1)}{P(n-1)}\tau_{\text{eff}}\gamma_x^s L_\Phi\right)\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2
$$

$$
\begin{aligned}
&+\frac{9\tau_{\text{eff}}\gamma_x^s L_f^2}{4\mu}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\\
&+\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+\sigma_G^2\left(2(\max_i w_i)\frac{n-P}{n-1}+2\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]\\
&+\frac{5}{2}\tau_{\text{eff}}\gamma_x^s\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\left\|\boldsymbol{a}_{i,-1}\right\|_2^2+2\sigma_G^2 M_{\mathbf{a}_{-1}}\right]\\
&+10\tau_{\text{eff}}\gamma_x^s L_f^2 M_{\mathbf{a}_{-1}}\beta_G^2\left[[\eta_x^c]^2\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2+L_f\left(2\kappa[\eta_x^c]^2+[\eta_y^c]^2\right)\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\right]. \qquad (17)
\end{aligned}
$$

对 17）从 $t=0,\ldots,T-1$ 求和，代入引理 $\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\ \mathbf{y}^{(t)})\right]$ 给出的界，并重新整理各项，可得

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\\
&=\mathcal{O}\left(\frac{\kappa^2\Delta_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_y^s T}+\tau_{\text{eff}}\gamma_y^s L_f\kappa^2\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+\sigma_G^2\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]\right)\\
&\quad+\mathcal{O}\left(\kappa^2\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\left(\|\boldsymbol{a}_i\|_2^2-[a_i^{(t,\tau_i-1)}]^2\right)+\sigma_G^2\max_i\left(\|\boldsymbol{a}_{i,-1}\|_1^2+\beta_L^2\|\boldsymbol{a}_{i,-1}\|_2^2\right)\right]\right) \qquad (18)
\end{aligned}
$$

因此，利用常数 A_w、B_w、C_w、D、E_w，式（18）可简化为

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\le\mathcal{O}\left(\kappa^2\left[\frac{\Delta_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_y^s T}+\frac{\gamma_y^s L_f}{P}\left(A_w\sigma_L^2+(B_w\beta_L^2+E_w\tau_{\text{eff}})\sigma_G^2\right)\right]+\kappa^2\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[C_w\sigma_L^2+D\sigma_G^2\right]\right).
$$

至此，证明完成。 □

关于 F的收敛

~推论1.1的证明。根据 $F(\mathbf{x})$ 和 $F(\mathbf{x})$的定义，可得

$$
\begin{aligned}
\nabla\Phi(\mathbf{x})-\nabla\widetilde{\Phi}(\mathbf{x})&=\sum_{i=1}^n\left[p_i\nabla_x f_i(\mathbf{x},\mathbf{y}^*(\mathbf{x}))-w_i\nabla_x f_i(\mathbf{x},\widetilde{\mathbf{y}}^*(\mathbf{x}))\right] \qquad (\mathbf{y}^*(\mathbf{x})\in\arg\max_{\mathbf{y}}F(\mathbf{x},\mathbf{y}))\\
&=\sum_{i=1}^n p_i\left[\nabla_x f_i(\mathbf{x},\mathbf{y}^*(\mathbf{x}))-\nabla_x f_i(\mathbf{x},\widetilde{\mathbf{y}}^*(\mathbf{x}))\right]+\sum_{i=1}^n(p_i-w_i)\nabla_x f_i(\mathbf{x},\widetilde{\mathbf{y}}^*(\mathbf{x}))\\
&=\left[\nabla_x F(\mathbf{x},\mathbf{y}^*(\mathbf{x}))-\nabla_x F(\mathbf{x},\widetilde{\mathbf{y}}^*(\mathbf{x}))\right]+\sum_{i=1}^n\frac{p_i-w_i}{\sqrt{w_i}}\cdot\sqrt{w_i}\nabla_x f_i(\mathbf{x},\widetilde{\mathbf{y}}^*(\mathbf{x})).
\end{aligned}
$$

取范数，利用 L_f-光滑性并应用柯西–施瓦茨不等式，可得

$$
\begin{aligned}
\left\|\nabla\Phi(\mathbf{x})-\nabla\widetilde{\Phi}(\mathbf{x})\right\|^2&\le 2L_f^2\left\|\mathbf{y}^*(\mathbf{x})-\widetilde{\mathbf{y}}^*(\mathbf{x})\right\|^2+2\left[\sum_{i=1}^n\frac{(p_i-w_i)^2}{w_i}\right]\left[\sum_{i=1}^n w_i\left\|\nabla_x f_i(\mathbf{x},\widetilde{\mathbf{y}}^*(\mathbf{x}))\right\|^2\right]\\
&\le 2L_f^2\left\|\mathbf{y}^*(\mathbf{x})-\widetilde{\mathbf{y}}^*(\mathbf{x})\right\|^2+2\chi^2_{\mathbf{p}\|\mathbf{w}}\left[\beta_G^2\left\|\nabla\widetilde{\Phi}(\mathbf{x})\right\|^2+\sigma_G^2\right],
\end{aligned}
$$

其中，最后一个不等式用到了假设3。接下来，注意到

$$
\left\|\nabla\Phi(\mathbf{x})\right\|^2\le 2\left\|\nabla\Phi(\mathbf{x})-\nabla\widetilde{\Phi}(\mathbf{x})\right\|^2+2\left\|\nabla\widetilde{\Phi}(\mathbf{x})\right\|^2.
$$

因此，我们得到

$$\begin{aligned}
&\min_{t\in[T]}\left\|\nabla\Phi(\mathbf{x}^{(t)})\right\|^2 \le \frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\Phi(\mathbf{x}^{(t)})\right\|^2 \\
&\le 2\left[2\chi^2_{\mathbf{p}\|\mathbf{w}}\beta_G^2+1\right]\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 + 4\left[\chi^2_{\mathbf{p}\|\mathbf{w}}\sigma_G^2 + L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{y}^*(\mathbf{x}^{(t)})-\widetilde{\mathbf{y}}^*(\mathbf{x}^{(t)})\right\|^2\right] \\
&= 2\left[2\chi^2_{\mathbf{p}\|\mathbf{w}}\beta_G^2+1\right]\epsilon_{\text{opt}} + 4\left[\chi^2_{\mathbf{p}\|\mathbf{w}}\sigma_G^2 + L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{y}^*(\mathbf{x}^{(t)})-\widetilde{\mathbf{y}}^*(\mathbf{x}^{(t)})\right\|^2\right].
\end{aligned}$$

其中，ϵ_{opt} 表示定理1中（4）右侧的优化误差。 □

推论*1.2*的证明。如果客户端的权重相同（$w_i = p_i = 1/n$ 对所有 i），且每个客户端执行 τ 步本地SGDA，如式（6）所示，我们得到

$$\min_{t\in[T]}\left\|\nabla\Phi(\mathbf{x}^{(t)})\right\|^2 \le \mathcal{O}\Big(\frac{\sigma_L^2+\beta_L^2\sigma_G^2}{\sqrt{P\tau T}} + \frac{\sigma_L^2+\tau\sigma_G^2}{\tau T} + \Big(\frac{n-P}{n-1}\Big)\sigma_G^2\sqrt{\frac{\tau}{PT}}\Big).$$

- 对于全客户端参与，这可化简为

$$\min_{t\in[T]}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 \le \mathcal{O}\left(\frac{1}{\sqrt{n\tau T}}+\frac{1}{T}\right).$$

  要达到一个 ϵ-驻点，假设 $n\tau \le T$，每客户端梯度复杂度为 $T\tau = \mathcal{O}\left(\frac{1}{n\epsilon^4}\right)$。由于$\tau \le T/n$，所需的最少通信轮次为 $T = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$。

- 对于部分参与，$\mathcal{O}\Big(\Big(\frac{n-P}{n-1}\Big)\sigma_G^2\sqrt{\frac{\tau}{PT}}\Big)$ 是主导项，多个本地更新不会带来任何收敛收益。因此，对于 $\tau = \mathcal{O}(1)$，每梯度的客户端复杂度和通信轮次数均为 $T\tau = \mathcal{O}\left(\frac{1}{P\epsilon^4}\right)$。然而，如果不同客户端的数据来自相同分布（$\sigma_G = 0$），则每客户端梯度复杂度可恢复为 $\mathcal{O}\left(\frac{1}{P\epsilon^4}\right)$，通信轮次数为 $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$。

□

B.3 中间引理的证明

Proof of Lemma B.2.

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{g}^{(t)}_{\mathbf{x},i}\right\|^2 &= \mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\left(\mathbf{g}^{(t)}_{\mathbf{x},i}-\mathbf{h}^{(t)}_{\mathbf{x},i}+\mathbf{h}^{(t)}_{\mathbf{x},i}\right)\right\|^2 \\
&= \mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\left(\mathbf{g}^{(t)}_{\mathbf{x},i}-\mathbf{h}^{(t)}_{\mathbf{x},i}\right)\right\|^2 + \mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{h}^{(t)}_{\mathbf{x},i}\right\|^2 \\
&= \mathbb{E}\left[\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i^2\left\|\mathbf{g}^{(t)}_{\mathbf{x},i}-\mathbf{h}^{(t)}_{\mathbf{x},i}\right\|^2\right] + \mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{h}^{(t)}_{\mathbf{x},i}\right\|^2 && \text{（采样方案）} \\
&= \frac{n}{P}\sum_{i=1}^n w_i^2\mathbb{E}\left\|\mathbf{g}^{(t)}_{\mathbf{x},i}-\mathbf{h}^{(t)}_{\mathbf{x},i}\right\|^2 + \mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{h}^{(t)}_{\mathbf{x},i}\right\|^2 && (\because \tilde{w}_i = w_i n/P)
\end{aligned}$$

$$\leq \frac{n}{P}\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\left[\sigma_L^2+\beta_L^2\mathbb{E}\left\|\nabla_x f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)})\right\|^2\right]+\mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2. \tag{19}$$

其中，最后一个不等式由假设1 和2推出。此外，第二项可以如下界定。

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{h}_{\mathbf{x},i}^{(t)}-\sum_{i=1}^{n}w_i\mathbf{h}_{\mathbf{x},i}^{(t)}+\sum_{i=1}^{n}w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\\
&=\mathbb{E}\left\|\sum_{i=1}^{n}w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2+\mathbb{E}\left\|\sum_{i=1}^{n}\mathbb{I}(i\in\mathcal{C}^{(t)})\tilde{w}_i\mathbf{h}_{\mathbf{x},i}^{(t)}-\sum_{i=1}^{n}w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 && \text{((WOR) sampling)}\\
&=\mathbb{E}\left\|\sum_{i=1}^{n}w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2+\sum_{i=1}^{n}\mathbb{E}\left[\left((\mathbb{I}(i\in\mathcal{C}^{(t)}))^2\tilde{w}_i^2+w_i^2-2\mathbb{I}(i\in\mathcal{C}^{(t)})\tilde{w}_iw_i\right)\left\|\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\right]\\
&\qquad+\sum_{i\neq j}\mathbb{E}\left\langle(\mathbb{I}(i\in\mathcal{C}^{(t)})\tilde{w}_i-w_i)\mathbf{h}_{\mathbf{x},i}^{(t)},(\mathbb{I}(j\in\mathcal{C}^{(t)})\tilde{w}_j-w_j)\mathbf{h}_{\mathbf{x},j}^{(t)}\right\rangle\\
&=\mathbb{E}\left\|\sum_{i=1}^{n}w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2+\sum_{i=1}^{n}\mathbb{E}\left[w_i^2\left(\frac{n}{P}-1\right)\left\|\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\right]\\
&\qquad+\sum_{i\neq j}\mathbb{E}\left[\left(\mathbb{I}(i\in\mathcal{C}^{(t)})\cdot\mathbb{I}(j\in\mathcal{C}^{(t)})\tilde{w}_i\tilde{w}_j-\mathbb{I}(j\in\mathcal{C}^{(t)})\tilde{w}_jw_i-\mathbb{I}(i\in\mathcal{C}^{(t)})\tilde{w}_iw_j+w_iw_j\right)\left\langle\mathbf{h}_{\mathbf{x},i}^{(t)},\mathbf{h}_{\mathbf{x},j}^{(t)}\right\rangle\right]\\
&=\mathbb{E}\left\|\sum_{i=1}^{n}w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2+\left(\frac{n}{P}-1\right)\sum_{i=1}^{n}\mathbb{E}\left[w_i^2\left\|\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\right]+\sum_{i\neq j}\mathbb{E}\left[w_iw_j\left(\frac{n}{P}\left(\frac{P-1}{n-1}\right)-1\right)\left\langle\mathbf{h}_{\mathbf{x},i}^{(t)},\mathbf{h}_{\mathbf{x},j}^{(t)}\right\rangle\right]\\
&=\frac{n}{P}\left(\frac{P-1}{n-1}\right)\mathbb{E}\left\|\sum_{i=1}^{n}w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2+\frac{n}{P}\frac{n-P}{n-1}\sum_{i=1}^{n}w_i^2\mathbb{E}\left\|\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2,
\end{aligned} \tag{20}$$

接下来，我们对（20）中的第二项进行界定。

$$\begin{aligned}
&\sum_{i=1}^{n}w_i^2\mathbb{E}\left\|\mathbf{h}_{\mathbf{x},i}^{(t)}-\nabla_x f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})+\nabla_x f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2\\
&\leq 2L_f^2\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)+2(\max_i w_i)\left(\beta_G^2\left\|\nabla_x\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2+\sigma_G^2\right),
\end{aligned} \tag{21}$$

根据假设 3。 $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\triangleq\mathbb{E}\left[\|\mathbf{x}_i^{(t,k)}-\mathbf{x}^{(t)}\|^2+\|\mathbf{y}_i^{(t,k)}-\mathbf{y}^{(t)}\|^2\right]$。将（21）代入（20），并将所得界代入（19），可得（14）中的界。 □

引理 *B.3* 的证明。由于局部函数 $\{f_i\}$ 满足假设 4， $F(\mathbf{x},\cdot)$ 是 μ-SC，对任意 $\mathbf{x}$均成立。在证明中，我们使用 μ-SC 函数 $F(\mathbf{x},\cdot)$的二次增长性质，即对于任意给定的 $\mathbf{x}$

$$\frac{\mu}{2}\|\mathbf{y}-\mathbf{y}^*(\mathbf{x})\|^2\leq F(\mathbf{x},\mathbf{y}^*(\mathbf{x}))-F(\mathbf{x},\mathbf{y}),\quad\text{for all }\mathbf{y}, \tag{22}$$

~其中，$\mathbf{y}^*(\mathbf{x})$ =最大值点集合$_{\mathbf{y}'}$ $F(\mathbf{x},\mathbf{y}')$。利用 L_Φ-光滑性 $\Phi(\cdot)$,

$$\mathbb{E}\widetilde{\Phi}(\mathbf{x}^{(t+1)})\leq\mathbb{E}\widetilde{\Phi}(\mathbf{x}^{(t)})-\mathbb{E}\left\langle\nabla\widetilde{\Phi}(\mathbf{x}^{(t)}),\tau_{\text{eff}}\gamma_x^s\frac{1}{|\mathcal{C}^{(t)}|}\sum_{i\in\mathcal{C}^{(t)}}\mathbf{g}_{\mathbf{x},i}^{(t)}\right\rangle+\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2L_\Phi}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{C}^{(t)}|}\sum_{i\in\mathcal{C}^{(t)}}\mathbf{g}_{\mathbf{x},i}^{(t)}\right\|^2$$

$$= \mathbb{E}\widetilde{\Phi}(\mathbf{x}^{(t)}) - \tau_{\text{eff}}\gamma_x^s \mathbb{E}\left\langle \nabla\widetilde{\Phi}(\mathbf{x}^{(t)}), \sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)} \right\rangle + \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{(t)}\right\|^2 \qquad \text{(using Assumption 2 and (3))}$$

$$= \mathbb{E}\widetilde{\Phi}(\mathbf{x}^{(t)}) - \frac{\tau_{\text{eff}}\gamma_x^s}{2}\mathbb{E}\left[\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 + \left\|\sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\right] + \frac{\tau_{\text{eff}}\gamma_x^s}{2}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)}) - \sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2$$
$$+ \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{(t)}\right\|^2. \tag{23}$$

接下来,

$$\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)}) - \sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2$$
$$= \mathbb{E}\left\|\sum_{i=1}^n w_i \left(\nabla_x f_i(\mathbf{x}^{(t)}, \mathbf{y}^*(\mathbf{x}^{(t)})) - \nabla_x f_i(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) + \nabla_x f_i(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \mathbf{h}_{\mathbf{x},i}^{(t)}\right)\right\|^2$$
(since $\mathbf{y}^*(\mathbf{x}) = \arg\max_{\mathbf{y}'} F(\mathbf{x}, \mathbf{y}')$)
$$\le 2L_f^2 \mathbb{E}\left\|\mathbf{y}^*(\mathbf{x}^{(t)}) - \mathbf{y}^{(t)}\right\|^2 + 2\mathbb{E}\left\|\sum_{i=1}^n w_i \left(\nabla_x f_i(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\nabla_x f_i\left(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}\right)\right)\right\|^2$$
(Jensen's inequality; L_f-smoothness; Young's inequality)
$$\le \frac{4L_f^2}{\mu}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right] + 2\sum_{i=1}^n w_i \frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\mathbb{E}\left\|\nabla_x f_i(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)})\right\|^2$$
(Quadratic growth of μ-SC functions (22); Jensen's inequality)
$$\le \frac{4L_f^2}{\mu}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right] + 2L_f^2\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\mathbb{E}\left[\left\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\right\|^2 + \left\|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\right\|^2\right]$$
(L_f-smoothness)
$$= \frac{4L_f^2}{\mu}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right] + 2L_f^2\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i). \tag{24}$$

其中 $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \triangleq \mathbb{E}\left[\left\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\right\|^2 + \left\|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\right\|^2\right]$。此外，包含 $\left\|\nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)})\right\|^2$ 在（14）中的项由引理 B.7给出界。

将（24）、（14）中的界以及引理B.7 代入（23），得到

$$\mathbb{E}\widetilde{\Phi}(\mathbf{x}^{(t+1)}) \le \mathbb{E}\widetilde{\Phi}(\mathbf{x}^{(t)}) - \frac{\tau_{\text{eff}}\gamma_x^s}{2}\mathbb{E}\left[\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 + \left\|\sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\right]$$
$$+ \frac{\tau_{\text{eff}}\gamma_x^s}{2}\left[\frac{4L_f^2}{\mu}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right] + 2L_f^2\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\right]$$
$$+ \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\left[\frac{n}{P}\left(\frac{P-1}{n-1}\right)\mathbb{E}\left\|\sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 + \frac{n}{P}\left(\frac{n-P}{n-1}\right)2L_f^2\sum_{i=1}^n \frac{w_i^2}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\right]$$
$$+ \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\frac{n\sigma_L^2}{P}\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\frac{n}{P}\left(\frac{n-P}{n-1}\right)2(\max_i w_i)\left(\beta_G^2\left\|\nabla_x\widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right\|^2 + \sigma_G^2\right)$$
$$+ \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\frac{n}{P}\beta_L^2\left[2L_f^2\sum_{i=1}^n \frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) + 2\sigma_G^2\left(\max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]$$

$$
\begin{aligned}
&+\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2L_\Phi}{2}\frac{n}{P}\beta_L^2 4\beta_G^2\left(\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left[\frac{2L_f^2}{\mu}\mathbb{E}\left(\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right)+\left\|\nabla_x\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\right]\\
&\le\mathbb{E}\widetilde{\Phi}(\mathbf{x}^{(t)})-\frac{7\tau_{\text{eff}}\gamma_x^s}{16}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2-\frac{\tau_{\text{eff}}\gamma_x^s}{2}\left(1-\frac{n}{P}\left(\frac{P-1}{n-1}\right)\tau_{\text{eff}}\gamma_x^sL_\Phi\right)\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\\
&+\frac{9\tau_{\text{eff}}\gamma_x^sL_f^2}{4\mu}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]+\frac{5}{4}\tau_{\text{eff}}\gamma_x^sL_f^2\sum_{i=1}^n\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\\
&+\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2L_\Phi}{2}\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+\sigma_G^2\left(2(\max_i w_i)\frac{n-P}{n-1}+2\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right],
\end{aligned}
$$

其中，利用关于学习率的假设简化了系数 γ_x^s

$$
\begin{aligned}
\tau_{\text{eff}}\gamma_x^sL_\Phi\left[(\max_i w_i)\left(\frac{n-P}{n-1}\right)+\beta_L^2\max_{i,k}\frac{w_ia_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right]&\le\frac{P}{4n}\\
\tau_{\text{eff}}\gamma_x^sL_\Phi\left[(\max_i w_i)\left(\frac{n-P}{n-1}\right)+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right]&\le\frac{P}{32\beta_G^2n}.
\end{aligned}
$$

至此，证明完成。 □

引理证明 *B.4*。 我们使用个体迭代的客户端更新方程（12）。为了界定 $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)$，我们先界定单个分量项 $\mathbb{E}\left\|\mathbf{x}_i^{(t,k)}-\mathbf{x}^{(t)}\right\|^2$。对于 $1\le k\le\tau_i$，根据修正方差假设

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{x}_i^{(t,k)}-\mathbf{x}^{(t)}\right\|^2=[\eta_x^c]^2\mathbb{E}\left\|\sum_{j=0}^{k-1}a_i^{(j)}(k)\left(\nabla_xf_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)};\xi_i^{(t,j)}\right)-\nabla_xf_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)+\nabla_xf_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right)\right\|^2\\
&=[\eta_x^c]^2\left[\mathbb{E}\left\|\sum_{j=0}^{k-1}a_i^{(j)}(k)\left(\nabla_xf_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)};\xi_i^{(t,j)}\right)-\nabla_xf_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right)\right\|^2+\mathbb{E}\left\|\sum_{j=0}^{k-1}a_i^{(j)}(k)\nabla_xf_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right\|^2\right]\\
&\qquad\text{(using unbiasedness in Assumption 2)}\\
&=[\eta_x^c]^2\left[\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2\mathbb{E}\left\|\nabla_xf_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)};\xi_i^{(t,j)}\right)-\nabla_xf_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right\|^2+\mathbb{E}\left\|\sum_{j=0}^{k-1}a_i^{(j)}(k)\nabla_xf_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right\|^2\right]\\
&\le[\eta_x^c]^2\left[\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2\left(\sigma_L^2+\beta_L^2\left\|\nabla_xf_i(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)})\right\|^2\right)+\left(\sum_{j=0}^{k-1}a_i^{(j)}(k)\right)\sum_{j=0}^{k-1}a_i^{(j)}(k)\mathbb{E}\left\|\nabla_xf_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right\|^2\right].
\end{aligned}
\tag{25}
$$

其中最后一个不等式由詹森不等式（引理 A.3）得出。接下来，注意到

$$
\begin{aligned}
&\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2\le\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\sum_{k=0}^{\tau_i-2}[a_i^{(j)}(k)]^2=\sum_{k=0}^{\tau_i-2}[a_i^{(j)}(k)]^2\le\|\boldsymbol{a}_i\|_2^2-[a_i^{(t,\tau_i-1)}(\tau_i)]^2,\\
&\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\sum_{j=0}^{k-1}a_i^{(j)}(k)\le\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\sum_{k=0}^{\tau_i-2}a_i^{(j)}(k)=\sum_{k=0}^{\tau_i-2}a_i^{(j)}(k)\le\|\boldsymbol{a}_i\|_1-[a_i^{(t,\tau_i-1)}(\tau_i)],\\
&\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2\le\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\sum_{j=0}^{\tau_i-2}[a_i^{(j)}(k)]^2\le\alpha\cdot\sum_{k=0}^{\tau_i-2}a_i^{(k)}(\tau_i),
\end{aligned}
\tag{26}
$$

为简洁起见，我们定义 $\|\boldsymbol{a}_{i,\ -1}\|_2^2 \triangleq \|\boldsymbol{a}_i\|_2^2 - [a_i^{(t,\tau_i-1)}(\tau_i)]^2$，$\|\boldsymbol{a}_{i,\ -1}\|_1 \triangleq \|\boldsymbol{a}_i\|_1 - [a_i^{(t,\tau_i-1)}(\tau_i)]$。利用（26），我们对（25）中的各项进行界定。

$$\begin{aligned}
&\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2\beta_L^2\left\|\nabla_x f_i(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)})\right\|^2 \\
&\le 2\beta_L^2\alpha L_f^2\sum_{j=0}^{\tau_i-2}[a_i^{(j)}(k)]\Delta_{\mathbf{x},\mathbf{y}}^{(t,j)} + 2\beta_L^2\left\|\boldsymbol{a}_{i,-1}\right\|_2^2\left\|\nabla_x f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2. \qquad (27)
\end{aligned}$$

同理，

$$\begin{aligned}
&\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\left(\sum_{j=0}^{k-1} a_i^{(j)}(k)\right)\sum_{j=0}^{k-1} a_i^{(j)}(k)\mathbb{E}\left\|\nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right\|^2 \\
&\le \frac{2}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\left(\sum_{j=0}^{\tau_i-2} a_i^{(j)}(k)\right)\sum_{j=0}^{\tau_i-2} a_i^{(j)}(k)\left[L_f^2\Delta_{\mathbf{x},\mathbf{y}}^{(t,j)} + \left\|\nabla_x f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2\right] \\
&\le 2\left\|\boldsymbol{a}_{i,-1}\right\|_1 L_f^2\sum_{j=0}^{\tau_i-2} a_i^{(j)}(k)\Delta_{\mathbf{x},\mathbf{y}}^{(t,j)} + 2\left\|\boldsymbol{a}_{i,-1}\right\|_1^2\left\|\nabla_x f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2. \qquad (28)
\end{aligned}$$

将 (27)、(28) 代入 (25)，可得

$$\begin{aligned}
\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\mathbb{E}\left\|\mathbf{x}_i^{(t,k)}-\mathbf{x}^{(t)}\right\|^2 &\le [\eta_x^c]^2\sigma_L^2\left\|\boldsymbol{a}_{i,-1}\right\|_2^2 + 2[\eta_x^c]^2L_f^2\left(\|\boldsymbol{a}_{i,-1}\|_1+\beta_L^2\alpha\right)\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \\
&\quad + 2[\eta_x^c]^2\left(\|\boldsymbol{a}_{i,-1}\|_1^2+\beta_L^2\left\|\boldsymbol{a}_{i,-1}\right\|_2^2\right)\mathbb{E}\left\|\nabla_x f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2. \qquad (29)
\end{aligned}$$

同理，我们可以给$\mathbf{y}$ 误差设界

$$\begin{aligned}
\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\mathbb{E}\left\|\mathbf{y}_i^{(t,k)}-\mathbf{y}^{(t)}\right\|^2 &\le [\eta_y^c]^2\sigma_L^2\left\|\boldsymbol{a}_{i,-1}\right\|_2^2 + 2[\eta_y^c]^2L_f^2\left(\|\boldsymbol{a}_{i,-1}\|_1+\beta_L^2\alpha\right)\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \\
&\quad + 2[\eta_y^c]^2\left(\|\boldsymbol{a}_{i,-1}\|_1^2+\beta_L^2\left\|\boldsymbol{a}_{i,-1}\right\|_2^2\right)\mathbb{E}\left\|\nabla_y f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2. \qquad (30)
\end{aligned}$$

结合 (29) 和 (30) 中的两个界，可得

$$\begin{aligned}
&\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\mathbb{E}\left[\left\|\mathbf{x}_i^{(t,k)}-\mathbf{x}^{(t)}\right\|^2+\left\|\mathbf{y}_i^{(t,k)}-\mathbf{y}^{(t)}\right\|^2\right] \le \left([\eta_x^c]^2+[\eta_y^c]^2\right)\sigma_L^2\left\|\boldsymbol{a}_{i,-1}\right\|_2^2 \\
&+ 2\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\|\boldsymbol{a}_i\|_1\left(\|\boldsymbol{a}_{i,-1}\|_1+\beta_L^2\alpha\right)\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \\
&+ 2\left(\|\boldsymbol{a}_{i,-1}\|_1^2+\beta_L^2\left\|\boldsymbol{a}_{i,-1}\right\|_2^2\right)\left[[\eta_x^c]^2\mathbb{E}\left\|\nabla_x f_i\left(\mathbf{x}^{(t)},\mathbf{y}^{(t)}\right)\right\|^2+[\eta_y^c]^2\mathbb{E}\left\|\nabla_y f_i\left(\mathbf{x}^{(t)},\mathbf{y}^{(t)}\right)\right\|^2\right]. \qquad (31)
\end{aligned}$$

定义 $A_m \triangleq 2L_f^2\left([\eta_x^c]^2+[\eta_y^c]^2\right)$最大值$_i\ \|\boldsymbol{a}_i\|_1\left(\|\boldsymbol{a}_{i,\ -1}\|_1+\beta_L^2\alpha\right)$。重新排列（31）中的各项，并对智能体取加权和，可得

$$L_f^2\sum_{i=1}^{n}\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)$$

$$
\begin{aligned}
&\leq \frac{\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\sigma_L^2}{1-A_m}\sum_{i=1}^n w_i\left\|\boldsymbol{a}_{i,-1}\right\|_2^2 \\
&\quad+\frac{2L_f^2}{1-A_m}\sum_{i=1}^n w_i\left(\left\|\boldsymbol{a}_{i,-1}\right\|_1^2+\beta_L^2\left\|\boldsymbol{a}_{i,-1}\right\|_2^2\right)\left[[\eta_x^c]^2\mathbb{E}\left\|\nabla_x f_i\left(\mathbf{x}^{(t)},\mathbf{y}^{(t)}\right)\right\|^2+[\eta_y^c]^2\mathbb{E}\left\|\nabla_y f_i\left(\mathbf{x}^{(t)},\mathbf{y}^{(t)}\right)\right\|^2\right] \\
&\leq \frac{\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2}{1-A_m}\sigma_L^2\sum_{i=1}^n w_i\left\|\boldsymbol{a}_{i,-1}\right\|_2^2+\frac{2L_f^2M_{\mathbf{a}_{-1}}}{1-A_m}[\eta_x^c]^2\left(\beta_G^2\mathbb{E}\left\|\sum_{i=1}^n w_i\nabla_x f_i\left(\mathbf{x}^{(t)},\mathbf{y}^{(t)}\right)\right\|^2+\sigma_G^2\right) \\
&\quad+\frac{2L_f^2M_{\mathbf{a}_{-1}}}{1-A_m}[\eta_y^c]^2\left(\beta_G^2\mathbb{E}\left\|\sum_{i=1}^n w_i\nabla_y f_i\left(\mathbf{x}^{(t)},\mathbf{y}^{(t)}\right)\right\|^2+\sigma_G^2\right). \qquad (32)
\end{aligned}
$$

其中（32）由假设 3推出，我们定义 $M_{\mathbf{a}_{-1}} \triangleq$ 最大值$_i\left(\left\|\boldsymbol{a}_{i,-1}\right\|_1^2+\beta_L^2\left\|\boldsymbol{a}_{i,-1}\right\|_2^2\right)$。我们在引理 B.6中对 $\mathbb{E}\left\|\nabla_x F\left(\mathbf{x}^{(t)}、\mathbf{y}^{(t)}\right)\right\|^2$ 给出了界。类似地，我们可以如下界定 $\mathbb{E}\left\|\nabla_y F\left(\mathbf{x}^{(t)}、\mathbf{y}^{(t)}\right)\right\|^2$。

$$
\begin{aligned}
\mathbb{E}\left\|\nabla_y F\left(\mathbf{x}^{(t)},\mathbf{y}^{(t)}\right)\right\|^2 &= \mathbb{E}\left\|\nabla_y F\left(\mathbf{x}^{(t)},\mathbf{y}^{(t)}\right)-\nabla_y F\left(\mathbf{x}^{(t)},\mathbf{y}^*(\mathbf{x}^{(t)})\right)\right\|^2 \qquad (\because \mathbf{y}^*(\mathbf{x})=\arg\max_{\mathbf{y}'} F(\mathbf{x},\mathbf{y}')) \\
&\leq 2L_f\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]. \qquad (33)
\end{aligned}
$$

利用 L_f-光滑性以及 $F(\mathbf{x}\ \cdot)$(引理 A.6) 的凹性。此外，对于 η_x^c,η_y^c 的取值，我们得到 $A_m \leq 1/2$。因此，将两个界代入式（32），即可完成证明。 □

引理 *B.5* 的证明。 利用 L_f-光滑性（假设 1）以及 $F(\mathbf{x}\ \cdot)$,

$$
\begin{aligned}
&\mathbb{E}\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)}) \leq \mathbb{E}\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t+1)})-\mathbb{E}\left\langle\nabla_y\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)}),\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\right\rangle+\frac{L_f}{2}\left\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\right\|^2 \\
&=\mathbb{E}\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t+1)})-\tau_{\text{eff}}\gamma_y^s\mathbb{E}\left\langle\nabla_y\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)}),\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\rangle+\frac{\tau_{\text{eff}}^2[\gamma_y^s]^2L_f}{2}\mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^{(t)}\right\|^2, \\
&\leq \mathbb{E}\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t+1)})-\frac{\tau_{\text{eff}}\gamma_y^s}{2}\mathbb{E}\left[\left\|\nabla_y\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\right\|^2+\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2-\left\|\nabla_y\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})-\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2\right] \\
&\quad+\frac{\tau_{\text{eff}}^2[\gamma_y^s]^2L_f}{2}\mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^{(t)}\right\|^2. \qquad (34)
\end{aligned}
$$

接下来，我们对式（34）中的各项分别取界。利用 μ-SC 函数的二次增长性质（引理 A.5）

$$
\mathbb{E}\left\|\nabla_y\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\right\|^2 \geq 2\mu\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)})-\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\right].
$$

接下来，我们将 $\left\|\nabla_y\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})-\sum_{i=1}^n w_i\mathbf{h}^{(t)}{}_{\mathbf{y}\ i}\right\|^2$，并采用与式（24）中类似的推理。

$$
\begin{aligned}
&\mathbb{E}\left\|\nabla_y\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})-\nabla_y\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})+\nabla_y\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})-\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2 \\
&\leq 2L_f^2\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|^2+2\left\|\sum_{i=1}^n w_i\left(\nabla_y f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})-\mathbf{h}_{\mathbf{y},i}^{(t)}\right)\right\|^2 \\
&\leq 2L_f^2\tau_{\text{eff}}^2[\gamma_x^s]^2\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{(t)}\right\|^2+2L_f^2\sum_{i=1}^n\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i), \qquad (35)
\end{aligned}
$$

利用引理**B.2** 中的式（14），我们可以将 $\mathbb{E}\|\mathbf{g}_{\mathbf{x}}^{(t)}\|^2$ 界定为

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{(t)}\right\|^2 \leq{}& \frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^{n} w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 + \frac{\sigma_L^2 n}{P}\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \frac{2\sigma_G^2}{P}\left(\frac{n(n-P)}{n-1}\max_i w_i\right) \\
&+ \frac{\beta_L^2 n}{P}\left[2L_f^2\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) + 4\beta_G^2\left(\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\mathbb{E}\left\|\nabla_x\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\right] \\
&+ \frac{\beta_L^2 n}{P}2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\left[\sigma_G^2 + 4\beta_G^2 L_f\kappa\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\right] \\
&+ \frac{n(n-P)}{n-1}\frac{2L_f^2}{P}\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \\
&+ \left(\frac{n(n-P)}{n-1}\max_i w_i\right)\frac{2\beta_G^2}{P}\left[4L_f\kappa\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right] + 2\left\|\nabla_x\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\right] \\
\leq{}& \frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^{n} w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 + \frac{\sigma_L^2 n}{P}\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \frac{2n\sigma_G^2}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right) \\
&+ \left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\frac{2nL_f^2}{P}\sum_{i=1}^{n}\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \\
&+ \frac{4n\beta_G^2}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left\|\nabla_x\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 \\
&+ \frac{8n\beta_G^2 L_f\kappa}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]. \qquad (36)
\end{aligned}
$$

同理，我们可以对 $\mathbb{E}\|\mathbf{g}_{\mathbf{y}}^{(t)}\|^2$ 进行界定，得到

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^{(t)}\right\|^2 \leq{}& \frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^{n} w_i \mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2 + \frac{\sigma_L^2 n}{P}\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \frac{2\sigma_G^2 n}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right) \\
&+ \left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\frac{2nL_f^2}{P}\sum_{i=1}^{n}\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \\
&+ \frac{4\beta_G^2 L_f n}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]. \qquad (37)
\end{aligned}
$$

将式（35）、（36）、（37）和引理 B.4 代入式（34），并重新整理各项，可得

$$
\begin{aligned}
&\mathbb{E}\widetilde{F}(\mathbf{x}^{(t+1)},\ \mathbf{y}^{(t)}) \\
&\leq \mathbb{E}\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t+1)}) - \frac{\tau_{\text{eff}}\gamma_y^s}{2}\left(1 - \tau_{\text{eff}}\gamma_y^s L_f\frac{n(P-1)}{P(n-1)}\right)\mathbb{E}\left\|\sum_{i=1}^{n} w_i \mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2 + \tau_{\text{eff}}^3[\gamma_x^s]^2\gamma_y^s L_f^2\frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^{n} w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \\
&\quad - \tau_{\text{eff}}\gamma_y^s\mu\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)}) - \widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\right] \\
&\quad + \tau_{\text{eff}}^3[\gamma_x^s]^2\gamma_y^s L_f^2\frac{4n\beta_G^2}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left\|\nabla_x\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 \\
&\quad + \frac{\tau_{\text{eff}}^2[\gamma_y^s]^2 L_f}{2}\left(1 + \frac{2\tau_{\text{eff}}[\gamma_x^s]^2 L_f}{\gamma_y^s}\right)\left[\frac{\sigma_L^2 n}{P}\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \frac{2\sigma_G^2 n}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\left[\tau_{\text{eff}}\gamma_y^s+\frac{\tau_{\text{eff}}^2[\gamma_y^s]^2L_fn}{P}\left(\frac{2\tau_{\text{eff}}[\gamma_x^s]^2L_f}{\gamma_y^s}+1\right)\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_{i,k}\frac{w_ia_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\right]L_f^2\sum_{i=1}^n\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\\
&+2\tau_{\text{eff}}^2[\gamma_y^s]^2L_f\left(\frac{4\tau_{\text{eff}}[\gamma_x^s]^2L_f\kappa}{\gamma_y^s}+1\right)\frac{\beta_G^2nL_f}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\\
\le\ &\mathbb{E}\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t+1)})-\frac{\tau_{\text{eff}}\gamma_y^s}{2}\left(1-\tau_{\text{eff}}\gamma_y^sL_f\frac{n(P-1)}{P(n-1)}\right)\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2+\tau_{\text{eff}}^3[\gamma_x^s]^2\gamma_y^sL_f^2\frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\\
&-\tau_{\text{eff}}\gamma_y^s\mu\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)})-\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\right]\\
&+\tau_{\text{eff}}^3\gamma_y^s[\gamma_x^s]^2L_f^2\frac{4\beta_G^2n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\mathbb{E}\left\|\nabla_x\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\\
&+\tau_{\text{eff}}^2[\gamma_y^s]^2L_f\left[\frac{\sigma_L^2n}{P}\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+\frac{2\sigma_G^2n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]\\
&\qquad\qquad(\because 4\tau_{\text{eff}}\gamma_x^sL_f\le 1,\gamma_x^s\kappa\le\gamma_y^s)\\
&+2\tau_{\text{eff}}\gamma_y^sL_f^2\sum_{i=1}^n\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\qquad(\because 8\tau_{\text{eff}}\gamma_x^sL_\Phi\le P)\\
&+4\tau_{\text{eff}}^2[\gamma_y^s]^2L_f^2\frac{\beta_G^2n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]
\end{aligned}\tag{38}
$$

其中，我们利用 $4\tau_{\text{eff}}\gamma_x^sL_f\le 1$、$\gamma_x^s\le\kappa\gamma_y^s$ 对部分系数进行了简化。重新整理各项，并使用引理 B.4中的界，可得

$$
\begin{aligned}
&\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)})-\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t+1)})\right]\\
\le\ &(1-\tau_{\text{eff}}\gamma_y^s\mu)\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)})-\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\right]\\
&-\frac{\tau_{\text{eff}}\gamma_y^s}{2}\left(1-\tau_{\text{eff}}\gamma_y^sL_f\frac{n(P-1)}{P(n-1)}\right)\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2+\tau_{\text{eff}}^3[\gamma_x^s]^2\gamma_y^sL_f^2\frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\\
&+4\tau_{\text{eff}}\gamma_y^sL_f^2\beta_G^2\left[\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)+4M_{\mathbf{a}_{-1}}[\eta_x^c]^2\right]\mathbb{E}\left\|\nabla_x\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\\
&+\tau_{\text{eff}}^2[\gamma_y^s]^2L_f\left[\frac{n\sigma_L^2}{P}\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+\frac{2n\sigma_G^2}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]\\
&+4\tau_{\text{eff}}\gamma_y^s\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2+2M_{\mathbf{a}_{-1}}\sigma_G^2\right]\\
&+4\tau_{\text{eff}}\gamma_y^sL_f^2\beta_G^2\left[\frac{\tau_{\text{eff}}\gamma_y^sn}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)+4L_fM_{\mathbf{a}_{-1}}\left(2\kappa[\eta_x^c]^2+[\eta_y^c]^2\right)\right]\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]
\end{aligned}\tag{39}
$$

接下来，注意到

$$
\begin{aligned}
&\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)})-\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\right]\\
&=\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)})-\widetilde{\Phi}(\mathbf{x}^{(t)})+\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})+\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})-\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\right].
\end{aligned}\tag{40}
$$

将引理 B.4 中的界代入引理 B.3，$\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)})-\widetilde{\Phi}(\mathbf{x}^{(t)})\right]$ 可得以下界。

$$
\mathbb{E}\widetilde{\Phi}(\mathbf{x}^{(t+1)})-\mathbb{E}\widetilde{\Phi}(\mathbf{x}^{(t)})
$$

$$
\begin{aligned}
&\leq -\frac{3\tau_{\text{eff}}\gamma_x^s}{8}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 - \frac{\tau_{\text{eff}}\gamma_x^s}{2}\left(1-\tau_{\text{eff}}\gamma_x^s L_\Phi\frac{n(P-1)}{P(n-1)}\right)\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 + \frac{5\tau_{\text{eff}}\gamma_x^s L_f^2}{2\mu}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\\
&\quad+\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2L_\Phi}{2}\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+\sigma_G^2\left(2(\max_i w_i)\frac{n-P}{n-1}+2\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]\\
&\quad+\frac{5}{2}\tau_{\text{eff}}\gamma_x^s\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\left\|\boldsymbol{a}_{i,-1}\right\|_2^2+2M_{\mathbf{a}_{-1}}\sigma_G^2\right]. \qquad (41)
\end{aligned}
$$

其中，我们使用了 $20L_f^2M_{\mathbf{a}_{-1}}\beta_G^2[\eta_x^c]^2\leq\frac{1}{8}, 40L_fM_{\mathbf{a}_{-1}}\beta_G^2\left(2\kappa[\eta_x^c]^2+[\eta_y^c]^2\right)\leq\frac{1}{\mu}$。接下来，我们对 $\mathbb{E}\left[\widetilde{F}(\mathbf{x}^{(t)},\ \mathbf{y}^{(t)})-\widetilde{F}(\mathbf{x}^{(t+1)},\ \mathbf{y}^{(t)})\right]$ 进行界定。再次利用 $F(\cdot,\ \mathbf{y})$ 的 L_f-光滑性，

$$
\begin{aligned}
&\mathbb{E}\left[\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})-\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\right]\\
&\leq\mathbb{E}\left[\left\langle-\nabla_x\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)}),\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\rangle+\frac{L_f}{2}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|^2\right]\\
&\leq\frac{\tau_{\text{eff}}\gamma_x^s}{2}\mathbb{E}\left[\left\|\nabla_x\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})-\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})+\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2+\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\right]+\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2L_f}{2}\mathbb{E}\left\|\frac{1}{|\mathcal{C}^{(t)}|}\sum_{i\in\mathcal{C}^{(t)}}\mathbf{g}_{\mathbf{x},i}^{(t)}\right\|^2\\
&\leq\tau_{\text{eff}}\gamma_x^s\mathbb{E}\left[\frac{2L_f^2}{\mu}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]+\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2+\frac{1}{2}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\right] \qquad (L_f\text{-smoothness, Lemma A.5})\\
&\quad+\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2L_f}{2}\mathbb{E}\left\|\mathbf{g}_{\mathbf{x}}^{(t)}\right\|^2\\
&\leq\frac{3\tau_{\text{eff}}\gamma_x^sL_f^2}{\mu}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]+2\tau_{\text{eff}}\gamma_x^s\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2+\frac{\tau_{\text{eff}}\gamma_x^s}{2}\left(1+\tau_{\text{eff}}\gamma_x^sL_f\frac{n(P-1)}{P(n-1)}\right)\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\\
&\qquad\qquad\qquad

$$
\begin{aligned}
&\le (1-\tau_{\text{eff}}\gamma_y^s\mu)\left[1+\frac{3\tau_{\text{eff}}\gamma_x^s L_f^2}{\mu}+\frac{5\tau_{\text{eff}}\gamma_x^s L_f^2}{2\mu}\right]\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\\
&+4\tau_{\text{eff}}\gamma_y^s L_f^2\beta_G^2\left[\frac{\tau_{\text{eff}}\gamma_y^s n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)+4L_f M_{\mathbf{a}_{-1}}\left(2\kappa[\eta_x^c]^2+[\eta_y^c]^2\right)\right]\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\\
&+4\tau_{\text{eff}}\gamma_y^s L_f^2\beta_G^2\left[\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)+4M_{\mathbf{a}_{-1}}[\eta_x^c]^2\right]\mathbb{E}\left\|\nabla_x\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\\
&-\frac{\tau_{\text{eff}}\gamma_y^s}{2}\left(1-\tau_{\text{eff}}\gamma_y^s L_f\frac{n(P-1)}{P(n-1)}\right)\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2+\tau_{\text{eff}}^3[\gamma_x^s]^2\gamma_y^s L_f^2\frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\\
&+(1-\tau_{\text{eff}}\gamma_y^s\mu)\left[\frac{13\tau_{\text{eff}}\gamma_x^s}{8}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2+\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2}{2}(L_\Phi+L_f)\frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\right]\\
&+\tau_{\text{eff}}^2[\gamma_y^s]^2 L_f\left[\frac{\sigma_L^2 n}{P}\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+\frac{2\sigma_G^2 n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]\\
&+4\tau_{\text{eff}}\gamma_y^s\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2+2M_{\mathbf{a}_{-1}}\sigma_G^2\right]\\
&+(1-\tau_{\text{eff}}\gamma_y^s\mu)\frac{\sigma_L^2\tau_{\text{eff}}^2[\gamma_x^s]^2 L_f}{2}\frac{n}{P}\left[\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+4\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2\right]\\
&+(1-\tau_{\text{eff}}\gamma_y^s\mu)\sigma_G^2\tau_{\text{eff}}^2[\gamma_x^s]^2 L_f\frac{n}{P}\left[\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)+4\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)L_f^2 M_{\mathbf{a}_{-1}}\left([\eta_x^c]^2+[\eta_y^c\right.\right.\\
&+(1-\tau_{\text{eff}}\gamma_y^s\mu)\frac{5}{2}\tau_{\text{eff}}\gamma_x^s\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2+2M_{\mathbf{a}_{-1}}\sigma_G^2\right]\\
&+(1-\tau_{\text{eff}}\gamma_y^s\mu)\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+\sigma_G^2\left(2(\max_i w_i)\frac{n-P}{n-1}+2\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right] \qquad (43)
\end{aligned}
$$

接下来，我们化简（43）中不同项的系数。

- $\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]$ 的系数可简化为

$$
\begin{aligned}
&(1-\tau_{\text{eff}}\gamma_y^s\mu)\left(1+\frac{11\tau_{\text{eff}}\gamma_x^s L_f^2}{2\mu}\right)\\
&\quad+4\tau_{\text{eff}}\gamma_y^s L_f^2\beta_G^2\left[\frac{\tau_{\text{eff}}\gamma_y^s n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)+4L_f M_{\mathbf{a}_{-1}}\left(2\kappa[\eta_x^c]^2+[\eta_y^c]^2\right)\right]\\
&\le 1-\frac{\tau_{\text{eff}}\gamma_y^s\mu}{4}.
\end{aligned}
$$

使用 $\gamma_x^s\le\frac{\gamma_y^s}{11\kappa^2}$、$\tau_{\text{eff}}\gamma_y^s\kappa L_f\beta_G^2\frac{n}{P}$最大值 $\left\{\beta_L^2\text{最大值}_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}、\frac{n-P}{n-1}\text{最大值}_i w_i\right\}\le\frac{1}{64}$、$\kappa L_f\beta_G\eta_x^c\le\frac{1}{16\sqrt{2M_{\mathbf{a}_{-1}}}}$和$L_f\beta_G\eta_y^c\le\frac{1}{16\sqrt{\kappa M_{\mathbf{a}_{-1}}}}$。

- $\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{(t)\,\mathbf{x}\,i}^{(t)}\right\|^2$ 的系数可以简化为

$$
\begin{aligned}
&\tau_{\text{eff}}^3[\gamma_x^s]^2\gamma_y^s L_f^2\frac{n(P-1)}{P(n-1)}+\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2}{2}(L_\Phi+L_f)\frac{n(P-1)}{P(n-1)}\\
&\le 2\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi\frac{n(P-1)}{P(n-1)}. \qquad (\because L_f\le L_\Phi,\tau_{\text{eff}}\gamma_y^s L_f\le 1)
\end{aligned}
$$

- $\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}_{(t)})\right\|^2$ 的系数可以简化为

$$
\begin{aligned}
&(1-\tau_{\text{eff}}\gamma_y^s\mu)\frac{13}{8}\tau_{\text{eff}}\gamma_x^s+4\tau_{\text{eff}}\gamma_y^s L_f^2\beta_G^2\left[\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)+4M_{\mathbf{a}_{-1}}[\eta_x^c]^2\right]\\
&\le\frac{\tau_{\text{eff}}\gamma_y^s}{48\kappa^2}.
\end{aligned}
$$

使用 $\gamma_x^s\le\frac{\gamma_y^s}{156\kappa^2}$、$\eta_x^c L_f\beta_G\le\frac{1}{64\kappa\sqrt{M_{\mathbf{a}_{-1}}}}$ 和 $\tau_{\text{eff}}\gamma_x^s L_f\beta_G\sqrt{\frac{n}{P}}$，$\max\left\{\frac{n-P}{n-1}\max_i w_i,\ \beta_L\sqrt{\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}}\right\}$ $(\frac{\gamma_x^s}{\gamma_y^s}\le\frac{1}{49\kappa^2})$ $\le 140\kappa$。

- σ_L^2 的系数可简化为

$$
\begin{aligned}
&2\tau_{\text{eff}}^2[\gamma_x^s]^2L_f\frac{n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2\\
&\quad+\tau_{\text{eff}}^2[\gamma_y^s]^2L_f\frac{n}{P}\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+\frac{\tau_{\text{eff}}^2[\gamma_x^s]^2(L_\Phi+L_f)}{2}\frac{n}{P}\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\\
&\quad+\tau_{\text{eff}}\left(4\gamma_y^s+\frac{5}{2}\gamma_x^s\right)\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2\\
&\le\frac{3}{2}\tau_{\text{eff}}^2[\gamma_y^s]^2L_f\frac{n}{P}\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+\frac{9}{2}\tau_{\text{eff}}\gamma_y^s\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2\\
&\quad+2\tau_{\text{eff}}^2[\gamma_x^s]^2L_f\frac{n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2.
\end{aligned}
$$

- σ_G^2 的系数可简化为

$$
\begin{aligned}
&\frac{2\tau_{\text{eff}}^2 n}{P}\left([\gamma_y^s]^2L_f+[\gamma_x^s]^2(L_f+L_\Phi)\right)\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\\
&\quad+\tau_{\text{eff}}L_f^2M_{\mathbf{a}_{-1}}\left([\eta_x^c]^2+[\eta_y^c]^2\right)\left(8\gamma_y^s+4\tau_{\text{eff}}[\gamma_x^s]^2L_f\frac{n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)+5\gamma_x^s\right)\\
&\le\frac{3\tau_{\text{eff}}^2[\gamma_y^s]^2L_f n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)+9\gamma_y^s\tau_{\text{eff}}L_f^2M_{\mathbf{a}_{-1}}\left([\eta_x^c]^2+[\eta_y^c]^2\right)\\
&\quad+4\tau_{\text{eff}}^2[\gamma_x^s]^2L_f^3M_{\mathbf{a}_{-1}}\left([\eta_x^c]^2+[\eta_y^c]^2\right)\frac{n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right).
\end{aligned}
$$

最后，将这些系数代入（43），对 $t=0$、.....、$T-1$ 求和并重新整理各项，可得

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\\
&\le\frac{4}{\tau_{\text{eff}}\gamma_y^s\mu}\left[\frac{\widetilde{\Phi}(\mathbf{x}^{(0)})-F(\mathbf{x}^{(0)},\mathbf{y}^{(0)})}{T}-\frac{\mathbb{E}\left(\widetilde{\Phi}(\mathbf{x}^{(T)})-F(\mathbf{x}^{(T)},\mathbf{y}^{(T)})\right)}{T}\right]+\frac{1}{12\mu\kappa^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\\
&\quad+\frac{8\tau_{\text{eff}}[\gamma_x^s]^2L_\Phi}{\gamma_y^s\mu}\frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2+18\kappa L_f\left([\eta_x^c]^2+[\eta_y^c]^2\right)\left[\sigma_L^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2+2\sigma_G^2M_{\mathbf{a}_{-1}}\right]
\end{aligned}
$$

$$+\frac{8\tau_{\text{eff}}[\gamma_x^s]^2\kappa}{\gamma_y^s}\frac{n}{P}\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\left\|\boldsymbol{a}_{i,-1}\right\|_2^2+2\sigma_G^2 M_{\mathbf{a}_{-1}}\right]$$
$$+6\tau_{\text{eff}}\gamma_y^s\kappa\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}+2\sigma_G^2\left(\frac{n-P}{n-1}\max_i w_i+\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right], \tag{44}$$

至此，证明完毕。 □

B.4 辅助引理

引理 B.6. 若本地客户端函数 $f_i(\mathbf{x}\ \cdot)$ 满足假设 *1*、*4*（L_f-光滑性以及 μ-强凹性（关于 $\mathbf{y}$），则函数 F 满足

$$\mathbb{E}\left\|\nabla_x\widetilde{F}(\mathbf{x},\mathbf{y}\right\|^2\leq 2\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x})\right\|^2+\frac{4L_f^2}{\mu}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x})-\widetilde{F}(\mathbf{x},\mathbf{y})\right]$$

证明。

$$\begin{aligned}\mathbb{E}\left\|\nabla_x\widetilde{F}(\mathbf{x},\mathbf{y}\right\|^2&\leq 2\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x})\right\|^2+2\mathbb{E}\left\|\nabla_x F\left(\mathbf{x},\mathbf{y}\right)-\nabla\widetilde{\Phi}(\mathbf{x})\right\|^2\\&\leq 2\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x})\right\|^2+2L_f^2\mathbb{E}\left\|\mathbf{y}^*(\mathbf{x})-\mathbf{y}\right\|^2 &&(L_f\text{-smoothness (Assumption 1)})\\&\leq 2\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x})\right\|^2+\frac{4L_f^2}{\mu}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x})-\widetilde{F}(\mathbf{x},\mathbf{y})\right]. &&(\text{Assumption 4})\end{aligned}$$

□

引理 B.7. 若本地客户端函数 $f_i(\mathbf{x}\ \cdot)$ 满足假设 *1*、*3* 和 *4*，则由算法 $\{\mathbf{x}_i^{(t,k)}$、$\mathbf{y}_i^{(t,k)}\}_{i,(t,k)}$生成的迭代序列*1* 满足

$$\begin{aligned}&\sum_{i=1}^n\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\mathbb{E}\left\|\nabla_x f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)})\right\|^2\\&\leq 2\sum_{i=1}^n\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2L_f^2\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)+2\sigma_G^2\left(\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\\&\quad+4\beta_G^2\left(\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left[\frac{2L_f^2}{\mu}\mathbb{E}\left(\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right)+\left\|\nabla_x\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\right].\end{aligned}$$

证明*i*

$$\begin{aligned}&\sum_{i=1}^n\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\mathbb{E}\left\|\nabla_x f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)})\pm\nabla_x f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2\\&\leq 2\sum_{i=1}^n\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2L_f^2\mathbb{E}\left[\left\|\mathbf{x}_i^{(t,k)}-\mathbf{x}^{(t)}\right\|^2+\left\|\mathbf{y}_i^{(t,k)}-\mathbf{y}^{(t)}\right\|^2\right]+2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\mathbb{E}\left\|\nabla_x f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2\\&\qquad\qquad(L_f\text{-smoothness})\\&\leq 2\sum_{i=1}^n\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2L_f^2\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)+2\left(\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left[\beta_G^2\left\|\nabla_x\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2+\sigma_G^2\right]\ (\text{Assumption 3})\end{aligned}$$

使用引理B.6 即可得出结果。 □

引理 B.8. 若本地客户端函数 $f_i(\mathbf{x}\ \cdot)$ 满足假设 *1*、*3* 和 *4*，则由算法 $\{\mathbf{x}_i^{(t,k)}$、$\mathbf{y}_i^{(t,k)}\}_{i,(t,k)}$*1* 生成的迭代序列满足

$$\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\mathbb{E}\left\|\nabla_y f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)})\right\|^2$$
$$\le 2\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2L_f^2\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)+2\left(\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left[\sigma_G^2+2\beta_G^2L_f\mathbb{E}\left(\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right)\right].$$

证明。 紧接着引理 B.7的证明，

$$\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\mathbb{E}\left\|\nabla_y f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)})\pm\nabla_y f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2$$
$$\le 2\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2L_f^2\mathbb{E}\left[\left\|\mathbf{x}_i^{(t,k)}-\mathbf{x}^{(t)}\right\|^2+\left\|\mathbf{y}_i^{(t,k)}-\mathbf{y}^{(t)}\right\|^2\right]+2\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\mathbb{E}\left\|\nabla_y f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2$$
(L_f-smoothness)
$$\le 2\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2L_f^2\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)+2\left(\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left[\beta_G^2\left\|\nabla_y\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2+\sigma_G^2\right]\ \text{(Assumption 3)}$$
$$\le 2\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2L_f^2\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)+2\left(\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left[\sigma_G^2+2\beta_G^2L_f\mathbb{E}\left(\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right)\right].$$

其中，最后一个不等式由 F 关于$\mathbf{y}$的光滑性和凹性得出。 □

B.5 Polyak–Łojasiewicz（PL）条件下的收敛

若全局函数满足假设 5，本节的结果经少量修改后仍然成立。关键区别在于，引理 A.6 不再成立。引理 B.2和引理 B.3 则完全照旧成立。引理 B.4 的表述需要做一些修改，因为证明中用到了引理A.6 。

引理 B.9。 假设局部损失函数 $\{f_i\}$ 满足假设*1*、*3*、*5*，且局部函数对应的随机预言机满足假设 *2*。在引理 *B.9*的条件下，由算法 *1*生成的迭代序列 $\{\mathbf{x}_i^{(t)}$，$\mathbf{y}_i^{(t)}\}$ 满足

$$L_f^2\sum_{i=1}^{n}\frac{p_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\le 2\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\sigma_L^2\sum_{i=1}^{n}p_i\left\|\boldsymbol{a}_{i,-1}\right\|_2^2+4L_f^2M_{\mathbf{a}_{-1}}\left([\eta_x^c]^2+[\eta_y^c]^2\right)\sigma_G^2$$
$$+8L_f^2M_{\mathbf{a}_{-1}}\beta_G^2[\eta_x^c]^2\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2+8\kappa L_f^3M_{\mathbf{a}_{-1}}\beta_G^2\left(2[\eta_x^c]^2+[\eta_y^c]^2\right)\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right],$$

其中 $M_{\mathbf{a}_{-1}}\triangleq$ 最大值$_i\left(\|\boldsymbol{a}_{i,-1}\|_1^2+\beta_L^2\|\boldsymbol{a}_{i,-1}\|_2^2\right)$。

引理 B.8 中的界也变为

$$\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\mathbb{E}\left\|\nabla_y f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)})\right\|^2$$
$$\le 2\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2L_f^2\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)+2\left(\max_i\frac{p_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left[\sigma_G^2+2\beta_G^2\kappa L_f\mathbb{E}\left(\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right)\right].$$

引理B.5 中的相同界仍然成立，但对学习率的条件更加严格，即 $\eta_y^cL_f\beta_G\le\frac{1}{16\kappa\sqrt{M_{\mathbf{a}_{-1}}}}$且 $\tau_{\text{eff}}\gamma_y^s\kappa L_f\beta_G^2\frac{1}{P}$最大值 $\left\{\kappa\beta_L^2\text{最大值}_i\frac{\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2},\ 1\right\}\le\frac{1}{64}$。因此，在学习率条件稍微更加严格的情况下，定理 1 中的界仍然成立。

C Fed-Norm-SGDA 的收敛性+针对非凸凹函数（定理 2）

本节安排如下。首先，在附录C.1中，我们给出一些中间结果，并在定理 2的证明中使用这些结果。接下来，在附录 C.2中，我们给出定理 2的证明，随后在附录 C.3中给出这些中间结果的证明。最后，我们在附录C.4中讨论将我们的结果推广至非凸单点凹函数。

我们所求解的问题为

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ \widetilde{F}(\mathbf{x},\mathbf{y}) \triangleq \sum_{i=1}^{n} w_i f_i(\mathbf{x},\mathbf{y}) \right\}.$$

我们定义 $\widetilde{\Phi}(\mathbf{x}) \triangleq \max_{\mathbf{y}} \widetilde{F}(\mathbf{x},\mathbf{y})$，以及 $\widetilde{\mathbf{y}}^*(\mathbf{x}) \in \arg\max_{\mathbf{y}} \widetilde{F}(\mathbf{x},\mathbf{y})$。由于 $\widetilde{F}(\mathbf{x},\cdot)$ 不再是强凹函数，$\mathbf{y}^*(\mathbf{x})$ 不一定唯一。在算法 1-Fed-Norm-SGDA+ 中，客户端更新如下

$$\begin{aligned}
\mathbf{x}_i^{(t,k)} &= \mathbf{x}^{(t)} - \eta_x^c \sum_{j=0}^{k-1} a_i^{(j)}(k) \nabla_x f_i(\mathbf{x}_i^{(t,j)}, \mathbf{y}_i^{(t,j)}; \xi_i^{(t,j)}), \\
\mathbf{y}_i^{(t,k)} &= \mathbf{y}^{(t)} + \eta_y^c \sum_{j=0}^{k-1} a_i^{(j)}(k) \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,j)}; \xi_i^{(t,j)}),
\end{aligned} \tag{45}$$

其中 $1 \le k \le \tau_i$。服务器更新如下

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \tau_{\text{eff}} \gamma_x^s \mathbf{g}_{\mathbf{x}}^{(t)}, \qquad \mathbf{y}^{(t+1)} = \mathbf{y}^{(t)} + \tau_{\text{eff}} \gamma_y^s \mathbf{g}_{\mathbf{y}}^{(t)}, \tag{46}$$

其中 $\mathbf{g}_{\mathbf{x}}^{(t)}$，$\mathbf{g}_{\mathbf{y}}^{(t)}$定义见（3）。归一化（随机）梯度向量定义如下

$$\begin{aligned}
\mathbf{g}_{\mathbf{x},i}^{(t)} &= \frac{1}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \nabla_x f_i\left(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}\right); \quad \mathbf{h}_{\mathbf{x},i}^{(t)} = \frac{1}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \nabla_x f_i\left(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}\right), \\
\mathbf{g}_{\mathbf{y},i}^{(t)} &= \frac{1}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \nabla_y f_i\left(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}\right); \quad \mathbf{h}_{\mathbf{y},i}^{(t)} = \frac{1}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \nabla_y f_i\left(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}\right).
\end{aligned} \tag{47}$$

C.1 中间引理

如第 5.2 节所述，我们分析平滑包络函数 $\Phi_{1/2L_f}$ 的收敛性。我们首先给出该函数单步衰减的界。

引理 C.1（平滑包络的单步衰减）。假设本地损失函数 $\{f_i\}$ 满足假设 *1*、*2*、*6* 和 *7*。那么，由算法 *1-Fed-Norm-SGDA+* 生成的迭代序列满足

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t+1)})\right] &\le \mathbb{E}\left[\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right] + \tau_{\mathit{eff}}^2 [\gamma_x^s]^2 L_f \frac{n}{P} \left[\sum_{i=1}^{n} \frac{w_i^2 \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} \left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2 \left(\frac{P-1}{n-1} + \frac{n-P}{n-1} \sum_{i=1}^{n} w_i^2\right)\right] \\
&\quad + 2\tau_{\mathit{eff}} \gamma_x^s \left\{ L_f^2 \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) + L_f \mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right] \right\} - \frac{\tau_{\mathit{eff}} \gamma_x^s}{8} \mathbb{E}\left\| \nabla \widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)}) \right\|^2,
\end{aligned}$$

其中 $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) = \mathbb{E}\left[\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\|^2 + \|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\|^2\right]$ 表示客户端 $i \in [n]$ 在训练轮次t的第 k 个本地步上的漂移。

在两个相邻的同步时刻（例如， t， $t+1$）之间，客户端会因本地下降/上升步骤而产生偏离，从而形成 $\{\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\}_{i,k}$ 项。此外， $\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right]$ 用于量化内层最大化的误差。在接下来的引理中，我们将对这两个误差项给出界。

引理 C.2（一致性误差）。 假设局部损失函数 $\{f_i\}$ 满足假设 *1*、*3*、*6*和 *7*。局部函数的随机预言机满足假设 *2*。此外，在算法*1-Fed-Norm-SGDA+*中，我们选择客户端学习率 η_y^c 使得 $\eta_y^c \leq \frac{1}{2L_f \max_i \|\boldsymbol{a}_i\|_1 \sqrt{2\max\{1,\beta_L^2\}}}$。那么，由算法 $\{\mathbf{x}_i^{(t)}、\mathbf{y}_i^{(t)}\}$ 生成的迭代序列满足算法 *1-Fed-Norm-SGDA+* 的条件

$$L_f^2 \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \leq 2\left([\eta_x^c]^2 + [\eta_y^c]^2\right) L_f^2 \sigma_L^2 \sum_{i=1}^{n} w_i \|\mathbf{a}_{i,-1}\|_2^2 + 4L_f^2 M_{\mathbf{a}_{-1}} \left([\eta_x^c]^2 G_{\mathbf{x}}^2 + [\eta_y^c]^2 \sigma_G^2\right)$$
$$+ 8[\eta_y^c]^2 L_f^3 M_{\mathbf{a}_{-1}} \beta_G^2 \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right],$$

其中 $M_{\mathbf{a}_{-1}} \triangleq \max_i(\|\mathbf{a}_{i,-1}\|_1^2 + \beta_L^2 \|\mathbf{a}_{i,-1}\|_2^2)$。

注意，一致性误差取决于差值 $\mathbb{E}[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})]$。这不同于引理 C.1 中的项$\mathbb{E}[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})]$。由于在算法 1-Fed-Norm-SGDA+ 中，更新 $\mathbf{y}_i^{(t,k)}$ 时，$\mathbf{x}$ 分量在 S 个通信轮次内保持不变于 $\widehat{\mathbf{x}}^{(s)}$，因此差值

$$\sum_{t=kS}^{(k+1)S-1} \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right]$$

可以解释为优化误差，即在 $F(\widehat{\mathbf{x}}^{(s)}, \cdot)$ 上经过 S 个通信轮次后最大化凹函数时的误差。接下来，我们对该误差进行界定。以下结果实质上将 FedNova（Wang 等人 (2020年)）的分析推广到了凹最大化（类似地，也包括凸最小化）问题。我们还将 Khaled 等人 (2020年)； Koloskova 等人 (2020年) 中相应的分析推广到了异构本地更新。

引理 C.3 (凹最大化的本地随机梯度更新)。 假设局部函数满足假设 *1*、*2*、*3*和 *6*。此外，令 $\left\|\mathbf{y}^{(t)}\right\|^2 \leq R$ 对所有 t 成立。我们运行算法 *1-Fed-Norm-SGDA+*，并采用客户端步长 η_y^c 使得 $64[\eta_y^c]^2 M_{\mathbf{a}_{-1}} L_f^2 \beta_G^2 \frac{n}{P} \leq 1$。此外，服务器步长 γ_y^s 满足

$$2\tau_{\mathit{eff}}\gamma_y^s L_f \frac{n}{P} \max\{\beta_G^2, 1\} \max\left\{\beta_L^2 \max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}, \frac{n-P}{n-1} \max_i w_i\right\} \leq \frac{1}{8},$$
$$2\tau_{\mathit{eff}}\gamma_y^s L_f \frac{n}{P} \max\left\{\frac{P-1}{n-1}, \beta_L^2 \max_{i,k} \frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right\} \leq \frac{1}{8}.$$

那么，算法 *1-Fed-Norm-SGDA+*生成的迭代序列满足

$$\frac{1}{S} \sum_{t=sS}^{(s+1)S-1} \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right]$$
$$\leq \frac{4R}{\tau_{\mathit{eff}}\gamma_y^s S} + \tau_{\mathit{eff}}\gamma_y^s \frac{n}{P}\left[\sigma_L^2 \sum_{i=1}^{n} \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + 2\sigma_G^2\left(\frac{n-P}{n-1} \max_i w_i + \beta_L^2 \max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]$$
$$+ 4L_f([\eta_x^c]^2 + [\eta_y^c]^2)\left[\sigma_L^2 \sum_{i=1}^{n} w_i \|\mathbf{a}_{i,-1}\|_2^2 + 2M_{\mathbf{a}_{-1}}(G_{\mathbf{x}}^2 + \sigma_G^2)\right],$$

其中 $M_{\mathbf{a}_{-1}} \triangleq$ 最大值$_i\left(\|\mathbf{a}_{i,\ -1}\|_1^2 + \beta_L^2 \|\mathbf{a}_{i,\ -1}\|_2^2\right)$。

备注 8。 值得注意的是，引理 C.3 的证明不需要局部函数具有全局凹性。相反，给定 $\mathbf{x}$，我们只需要局部函数在某个点 $\{f_i\}$ 处$\mathbf{y}^*(\mathbf{x})$具有凹性。这正是前文 8 中讨论的单点凹性假设（假设 Deng& Mahdavi(2021年);Sharma 等人 (2022年)提出的。因此，引理 C.3 适用于范围广得多的函数类。此外，引理 C.3中的界改进了现有工作中推导出的相应界。正如我们在附录 C.4 中所讨论的，这有助于我们为非凸单点凹函数（NC-1PC）取得更优的复杂度结果。

接下来，我们界定 $\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right]$ 之间的差异。

引理 C.4。假设局部函数满足假设 *1*、*2*、*3*和*7*。则由算法 *1-Fed-Norm-SGDA+* 生成的迭代序列满足

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right] \leq \frac{1}{T}\sum_{s=0}^{T/S-1}\sum_{t=sS}^{(s+1)S-1}\mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right]$$
$$+ 2\tau_{\mathit{eff}}\gamma_x^s G_{\mathbf{x}}(S-1)\sqrt{\frac{n}{P}}\sqrt{\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{P-1}{n-1} + \frac{n-P}{n-1}\sum_{i=1}^{n}w_i^2\right)}.$$

C.2 定理 2 的证明

为完整起见，我们先在此给出定理 2 的完整表述。

定理 4。假设局部损失函数 $\{f_i\}$ 满足假设 *1*、*2*、*3*、*6*和*7*。此外，令 $\left\|\mathbf{y}^{(t)}\right\|^2 \leq R$ 对所有 t成立。我们运行算法 *1-Fed-Norm-SGDA+*，客户端步长为 η_y^c，且满足 $64[\eta_y^c]^2 M_{\mathbf{a}_{-1}} L_f^2 \max\{\beta_G^2 \frac{n}{P}, 1\} \leq 1$。此外，服务器步长 γ_y^s 满足

$$2\tau_{\mathit{eff}}\gamma_y^s L_f \frac{n}{P}\max\{\beta_G^2, 1\}\max\left\{\beta_L^2 \max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}, \frac{n-P}{n-1}\max_i w_i\right\} \leq \frac{1}{8},$$
$$2\tau_{\mathit{eff}}\gamma_y^s L_f \frac{n}{P}\max\left\{\frac{P-1}{n-1}, \beta_L^2 \max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right\} \leq \frac{1}{8}.$$

然后，算法*1-Fed-Norm-SGDA+*生成的迭代序列满足

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \leq \mathcal{O}\left(\frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\mathit{eff}}\gamma_x^s T} + \tau_{\mathit{eff}}\gamma_x^s L_f\left[\frac{A_w}{P\tau_{\mathit{eff}}}\left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{n(P-1)}{P(n-1)} + F_w\right)\right]\right)$$
$$+ \mathcal{O}\left(\tau_{\mathit{eff}}\gamma_x^s L_f G_{\mathbf{x}}(S-1)\sqrt{\frac{A_w}{P\tau_{\mathit{eff}}}\left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{n(P-1)}{P(n-1)} + F_w\right)}\right)$$
$$+ \mathcal{O}\left(\frac{L_f R}{\tau_{\mathit{eff}}\gamma_y^s S} + \frac{\gamma_y^s L_f}{P}\left(A_w\sigma_L^2 + \sigma_G^2\left(\tau_{\mathit{eff}}\frac{n-P}{n-1}E_w + B_w\beta_L^2\right)\right)\right) + \mathcal{O}\left(\left([\eta_x^c]^2 + [\eta_y^c]^2\right)L_f^2\left[C_w\sigma_L^2 + D\left(G_{\mathbf{x}}^2 + \sigma_G^2\right)\right]\right), \tag{48}$$

其中，$\widetilde{\Phi}_{1/2L_f}(x) \triangleq \min_{\mathbf{x}'}\widetilde{\Phi}(x') + L_f\|x'-x\|^2$是包络函数，$\bar{\Delta}_{\widetilde{\Phi}} \triangleq \widetilde{\Phi}_{1/2L_f}(x_0) - \min_{\mathbf{x}}\widetilde{\Phi}_{1/2L_f}(x)$, $A_w \triangleq n\tau_{\mathit{eff}}\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$, $B_w \triangleq n\tau_{\mathit{eff}}\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$, $E_w \triangleq n\max_i w_i$, $C_w \triangleq \sum_{i=1}^{n}w_i(\|\boldsymbol{a}_i\|_2^2 - [\alpha_i^{(t,\tau_i-1)}]^2)$, 且$D \triangleq \max_i(\|\boldsymbol{a}_{i,-1}\|_1^2 + \beta_L^2\|\boldsymbol{a}_{i,-1}\|_2^2)$, $F_w \triangleq \frac{n}{(n-P)P(n-1)}\sum_{i=1}^{n}w_i^2$。参数取值如下：

$$\eta_x^c = \eta_y^c = \Theta\left(\frac{1}{L_f\bar{\tau}T^{3/8}}\right), \gamma_x^s = \Theta\left(\frac{P^{1/4}}{(\tau_{\mathit{eff}}T)^{3/4}}\right), \quad \gamma_y^s = \Theta\left(\frac{P^{3/4}}{(\tau_{\mathit{eff}}T)^{1/4}}\right), \quad S = \Theta\left(\sqrt{\frac{T}{\tau_{\mathit{eff}}P}}\right),$$

其中 $\bar{\tau} = \frac{1}{n}\sum_{i=1}^{n} \tau_i$，我们可以进一步简化为

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla \widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2$$
$$\leq \underbrace{\mathcal{O}\left(\frac{(\bar{\tau}/\tau_{\mathit{eff}})^{1/4}}{(\bar{\tau}PT)^{1/4}}\right) + \mathcal{O}\left(\frac{(\tau_{\mathit{eff}}P)^{1/4}}{T^{3/4}}\right)}_{\textit{Error with full synchronization}} + \underbrace{\mathcal{O}\left(\left(\frac{n-P}{n-1}\cdot\frac{E_w}{PT}\right)^{1/4}\right)}_{\textit{Partial participation error}} + \underbrace{\mathcal{O}\left(\frac{C_w\sigma_L^2 + D(G_{\mathbf{x}}^2+\sigma_G^2)}{\bar{\tau}^2T^{3/4}}\right)}_{\textit{Error due to local updates}}.$$

证明。 我们将引理 C.1 中的界从 $t=0$ 到 $T-1$ 求和，并重新排列各项，得到 t

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla \widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2$$
$$\leq \frac{8}{\tau_{\text{eff}}\gamma_x^s T}\sum_{t=0}^{T-1}\mathbb{E}\left[\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)}) - \widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t+1)})\right]$$
$$+ 8\tau_{\text{eff}}\gamma_x^s L_f \frac{n}{P}\left[\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\left(\sigma_L^2+\beta_L^2 G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{P-1}{n-1}+\frac{n-P}{n-1}\sum_{i=1}^{n}w_i^2\right)\right]$$
$$+ 16L_f\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right] + \frac{16}{T}\sum_{s=0}^{T/S-1}\sum_{t=sS}^{(s+1)S-1} L_f^2\sum_{i=1}^{n}\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)$$
$$\leq \frac{8\left[\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(0)}) - \widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(T)})\right]}{\tau_{\text{eff}}\gamma_x^s T} + 8\tau_{\text{eff}}\gamma_x^s L_f\frac{n}{P}\left[\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\left(\sigma_L^2+\beta_L^2G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{P-1}{n-1}+\frac{n-P}{n-1}\sum_{i=1}^{n}w_i^2\right)\right]$$
$$+ 32\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\sigma_L^2\sum_{i=1}^{n}w_i\|\mathbf{a}_{i,-1}\|_2^2 + 64L_f^2M_{\mathbf{a}_{-1}}\left([\eta_x^c]^2G_{\mathbf{x}}^2+[\eta_y^c]^2\sigma_G^2\right) \qquad \text{(From Lemma C.2)}$$
$$+ 128[\eta_y^c]^2L_f^3M_{\mathbf{a}_{-1}}\beta_G^2\frac{1}{T}\sum_{s=0}^{T/S-1}\sum_{t=sS}^{(s+1)S-1}\mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)},\mathbf{y}^{(t)})\right] + 16L_f\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]$$
$$\leq \frac{8\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_x^s T} + 8\tau_{\text{eff}}\gamma_x^s L_f\frac{n}{P}\left[\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\left(\sigma_L^2+\beta_L^2G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{P-1}{n-1}+\frac{n-P}{n-1}\sum_{i=1}^{n}w_i^2\right)\right]$$
(where $\bar{\Delta}_{\widetilde{\Phi}} \triangleq \widetilde{\Phi}_{1/2L_f}(\mathbf{x}_0) - \min_{\mathbf{x}}\widetilde{\Phi}_{1/2L_f}(\mathbf{x})$)
$$+ 32\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^{n}w_i\|\mathbf{a}_{i,-1}\|_2^2 + 2M_{\mathbf{a}_{-1}}\left(G_{\mathbf{x}}^2+\sigma_G^2\right)\right]$$
$$+ 32\tau_{\text{eff}}\gamma_x^s L_f G_{\mathbf{x}}(S-1)\sqrt{\frac{n}{P}}\sqrt{\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\left(\sigma_L^2+\beta_L^2G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{P-1}{n-1}+\frac{n-P}{n-1}\sum_{i=1}^{n}w_i^2\right)}$$
(From Lemma C.4)
$$+ 18L_f\left[\frac{4R}{\tau_{\text{eff}}\gamma_y^s S} + \tau_{\text{eff}}\gamma_y^s\frac{n}{P}\left(\sigma_L^2\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + 2\sigma_G^2\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right)\right]$$
(From Lemma C.3; using $A_m \leq \min\{\frac{1}{2}, \frac{1}{16\beta_G^2}\}$)
$$+ 72[\eta_y^c]^2L_f^2\left[\sigma_L^2\sum_{i=1}^{n}w_i\|\mathbf{a}_{i,-1}\|_2^2 + 2\sigma_G^2M_{\mathbf{a}_{-1}}\right]. \qquad (49)$$

我们可以使用常数 $A_w \triangleq n\tau_{\text{eff}}\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$, $B_w \triangleq \tau_{\text{eff}}n\left(\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)$, $E_w \triangleq$ 简化记号

$n\max_i w_i$, $C_w \triangleq \sum_{i=1}^n w_i \left\|\mathbf{a}_{i,-1}\right\|_2^2$, $D \triangleq M_{\mathbf{a}_{-1}}$, $F_w \triangleq \frac{n}{P}\sum_{i=1}^n w_i^2$并为简便起见省略数值常数，得到

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \\
&\lesssim \frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_x^s T} + \tau_{\text{eff}}\gamma_x^s L_f\left[\frac{A_w}{P\tau_{\text{eff}}}\left(\sigma_L^2+\beta_L^2 G_{\mathbf{x}}^2\right)+G_{\mathbf{x}}^2\left(\frac{n(P-1)}{P(n-1)}+\frac{(n-P)}{(n-1)}F_w\right)\right] \\
&\quad + \tau_{\text{eff}}\gamma_x^s L_f G_{\mathbf{x}}(S-1)\sqrt{\frac{A_w}{P\tau_{\text{eff}}}\left(\sigma_L^2+\beta_L^2 G_{\mathbf{x}}^2\right)+G_{\mathbf{x}}^2\left(\frac{n(P-1)}{P(n-1)}+\frac{(n-P)}{(n-1)}F_w\right)} \\
&\quad + \frac{L_f R}{\tau_{\text{eff}}\gamma_y^s S}+\frac{\gamma_y^s L_f}{P}\left(A_w\sigma_L^2+\sigma_G^2\left(\tau_{\text{eff}}\frac{n-P}{n-1}E_w+B_w\beta_L^2\right)\right)+\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[C_w\sigma_L^2+D\left(G_{\mathbf{x}}^2+\sigma_G^2\right)\right] \\
&= \frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_x^s T}+\gamma_x^s L_f\mathcal{I}_1^2+\frac{\gamma_y^s L_f\mathcal{I}_2}{P}+L_f\left[\tau_{\text{eff}}\gamma_x^s G_{\mathbf{x}}(S-1)\mathcal{I}_1+\frac{R}{\tau_{\text{eff}}\gamma_y^s S}\right] \\
&\quad + \left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[C_w\sigma_L^2+D(G_{\mathbf{x}}^2+\sigma_G^2)\right], \qquad (50)
\end{aligned}
$$

其中，在（50）中，为简化记号，我们定义了$\mathcal{I}_1 \triangleq \sqrt{\frac{A_w}{P\tau_{\text{eff}}}\left(\sigma_L^2+\beta_L^2 G_{\mathbf{x}}^2\right)+G_{\mathbf{x}}^2\left(\frac{n(P-1)}{P(n-1)}+\frac{(n-P)}{(n-1)}F_w\right)}$、$\mathcal{I}_2 \triangleq A_w\sigma_L^2+\left(B_w\beta_L^2+\tau_{\text{eff}}\frac{n-P}{n-1}E_w\right)\sigma_G^2$。

接下来，我们优化算法参数S、γ_x^s、γ_y^s、η_y^c、η_y^c，以在（50）上获得紧界。如果$R=0$，令$S=1$。否则，令$S=\sqrt{\frac{R}{\tau_{\text{eff}}^2\gamma_x^s\gamma_y^s G_{\mathbf{x}}\mathcal{I}_1}}$。将其代入（50），可得

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 &\lesssim \frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_x^s T}+\tau_{\text{eff}}\gamma_x^s L_f\mathcal{I}_1^2+\frac{\gamma_y^s L_f\mathcal{I}_2}{P}+L_f\sqrt{\frac{R\gamma_x^s G_{\mathbf{x}}\mathcal{I}_1}{\gamma_y^s}} \\
&\quad + \left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[C_w\sigma_L^2+D(G_{\mathbf{x}}^2+\sigma_G^2)\right], \qquad (51)
\end{aligned}
$$

接下来，我们关注（51）中包含γ_y^s的项：$L_f\left[\frac{\gamma_y^s\mathcal{I}_2}{P}+\sqrt{\frac{R\gamma_x^s G_{\mathbf{x}}\mathcal{I}_1}{\gamma_y^s}}\right]$。为优化这些项，我们选择$\gamma_y^s = \left(\frac{P}{2\mathcal{I}_2}\right)^{2/3}\left(R\gamma_x^s G_{\mathbf{x}}\mathcal{I}_1\right)^{1/3}$。将其代入（51），可得

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 &\lesssim \frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_x^s T}+\tau_{\text{eff}}\gamma_x^s L_f\mathcal{I}_1^2+L_f\left(\frac{\mathcal{I}_2}{P}R\gamma_x^s G_{\mathbf{x}}\mathcal{I}_1\right)^{1/3} \\
&\quad + \left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[C_w\sigma_L^2+D(G_{\mathbf{x}}^2+\sigma_G^2)\right], \qquad (52)
\end{aligned}
$$

最后，我们关注（52）中包含γ_x^s的项：$\frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_x^s T}+L_f\left(\frac{\mathcal{I}_2}{P}R\gamma_x^s G_{\mathbf{x}}\mathcal{I}_1\right)^{1/3}$。我们忽略高阶线性项。令$\gamma_x^s = \left(\frac{3\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}L_f T}\right)^{3/4}\left(\frac{\mathcal{I}_2}{P}RG_{\mathbf{x}}\mathcal{I}_1\right)^{-1/4}$，并将数值常数吸收到$\mathcal{O}(\cdot)$中，可得，

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \lesssim \frac{(\mathcal{I}_1\mathcal{I}_2 RG_{\mathbf{x}})^{1/4}}{(\tau_{\text{eff}}PT)^{1/4}}+\frac{(\tau_{\text{eff}}P)^{1/4}}{T^{3/4}}(\mathcal{I}_1\mathcal{I}_2 RG_{\mathbf{x}})^{-1/4}\mathcal{I}_1^2 \\
&\quad + \left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[C_w\sigma_L^2+D(G_{\mathbf{x}}^2+\sigma_G^2)\right], \\
&= \mathcal{O}\left(\frac{(\mathcal{I}_1\mathcal{I}_2)^{1/4}}{(\tau_{\text{eff}}PT)^{1/4}}\right)+\mathcal{O}\left(\frac{(\tau_{\text{eff}}P)^{1/4}}{T^{3/4}}\mathcal{I}_2^{3/4}\right)+\mathcal{O}\left(\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[C_w\sigma_L^2+D(G_{\mathbf{x}}^2+\sigma_G^2)\right]\right), \\
&\leq \mathcal{O}\left(\frac{(\bar{\tau}/\tau_{\text{eff}})^{1/4}}{(\bar{\tau}PT)^{1/4}}\right)+\mathcal{O}\left(\frac{\left(\frac{n(n-P)}{n-1}\max_i w_i\right)^{1/4}\sigma_G}{(PT)^{1/4}}\right)+\mathcal{O}\left(\frac{(\tau_{\text{eff}}P)^{1/4}}{T^{3/4}}\right)
\end{aligned}
$$

$$+\mathcal{O}\left(\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[C_w\sigma_L^2+D(G_{\mathbf{x}}^2+\sigma_G^2)\right]\right), \tag{53}$$

其中，在（53）中，我们仅展示了对 τ、n、P、T的依赖关系。最后，我们将算法参数表示为n、T、τ_{eff}、$\bar{\tau}$。

$$\gamma_x^s=\Theta\left(\frac{P^{1/4}}{(\tau_{\text{eff}}T)^{3/4}}\right),\quad \gamma_y^s=\Theta\left(\frac{P^{3/4}}{(\tau_{\text{eff}}T)^{1/4}}\right),\quad S=\Theta\left(\sqrt{\frac{T}{\tau_{\text{eff}}P}}\right).$$

最后，选择客户端学习率 $\eta_x^c=\eta_y^c=\frac{1}{L_f\bar{\tau}T^{3/8}}$，得到

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2\le\mathcal{O}\left(\frac{(\bar{\tau}/\tau_{\text{eff}})^{1/4}}{(\bar{\tau}PT)^{1/4}}\right)+\mathcal{O}\left(\frac{1}{(PT)^{1/4}}\left(\frac{n(n-P)}{(n-1)}\max_i w_i\right)^{1/4}\right)+\mathcal{O}\left(\frac{(\tau_{\text{eff}}P)^{1/4}}{T^{3/4}}\right)$$
$$+\mathcal{O}\left(\frac{C_w\sigma_L^2+D(G_{\mathbf{x}}^2+\sigma_G^2)}{\bar{\tau}^2T^{3/4}}\right).$$

□

关于 F的收敛

推论 2.1的证明。 按照 Lin 等人 (2020a) 的定义，

$$\begin{aligned}&\widetilde{\Phi}_{1/2L_f}(\mathbf{x})\triangleq\min_{\mathbf{x}'}\left\{\widetilde{\Phi}(\mathbf{x}')+L_f\left\|\mathbf{x}'-\mathbf{x}\right\|^2\right\};\qquad &&\widetilde{\mathbf{x}}\triangleq\arg\min_{\mathbf{x}'}\left\{\widetilde{\Phi}(\mathbf{x}')+L_f\left\|\mathbf{x}'-\mathbf{x}\right\|^2\right\},\\&\Phi_{1/2L_f}(\mathbf{x})\triangleq\min_{\mathbf{x}'}\left\{\Phi(\mathbf{x}')+L_f\left\|\mathbf{x}'-\mathbf{x}\right\|^2\right\};\qquad &&\bar{\mathbf{x}}\triangleq\arg\min_{\mathbf{x}'}\left\{\Phi(\mathbf{x}')+L_f\left\|\mathbf{x}'-\mathbf{x}\right\|^2\right\}.\end{aligned}\tag{54}$$

~此外，根据 Davis & Drusvyatskiy (2019年) 中的引理 2.2 可知 $\nabla\Phi_{1/2L_f}(\mathbf{x})=2L_f(\mathbf{x}-\widetilde{\mathbf{x}})$ 以及 $\nabla\Phi_{1/2L_f}(\mathbf{x})=2L_f(\mathbf{x}-\bar{\mathbf{x}})$。因此，

$$\begin{aligned}\left\|\nabla\Phi_{1/2L_f}(\mathbf{x})\right\|^2\le&2\left\|\nabla\Phi_{1/2L_f}(\mathbf{x})-\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x})\right\|^2+2\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x})\right\|^2\\=&\,8L_f^2\left\|\widetilde{\mathbf{x}}-\bar{\mathbf{x}}\right\|^2+2\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x})\right\|^2\end{aligned}$$

因此，我们得到

$$\begin{aligned}\min_{t\in[T]}\left\|\nabla\Phi_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2&\le\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\Phi_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2\\&\le\frac{2}{T}\sum_{t=0}^{T-1}\left[\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2+4L_f^2\left\|\widetilde{\mathbf{x}}^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|^2\right].\end{aligned}\tag{55}$$

其中 $\widetilde{\mathbf{x}}^{(t)}$，$\bar{\mathbf{x}}^{(t)}$ 遵循与（54）中相同的定义，只需将 $\mathbf{x}$ 替换为 $\mathbf{x}^{(t)}$。 □

推论的证明2.2。若所有客户端权重相等（$w_i=p_i=1/n$ 对所有 i），且每个客户端执行 τ 步本地SGDA+，则（7）可化简为

$$\min_{t\in[T]}\mathbb{E}\left\|\nabla\Phi_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2\le\mathcal{O}\left(\frac{1}{(\tau PT)^{1/4}}+\frac{(\tau P)^{1/4}}{T^{3/4}}\right)+\mathcal{O}\left(\frac{\sigma_L^2+\tau(G_{\mathbf{x}}^2+\sigma_G^2)}{\tau T^{3/4}}\right)+\mathcal{O}\left(\left(\frac{n-P}{n-1}\cdot\frac{1}{PT}\right)^{1/4}\right).$$

- 对于全客户端参与，这可化简为

$$\min_{t\in[T]}\mathbb{E}\left\|\nabla\Phi_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2\le\mathcal{O}\left(\frac{1}{(\tau nT)^{1/4}}+\frac{(\tau n)^{1/4}}{T^{3/4}}\right).$$

要达到一个 ϵ-驻点，假设 $n\tau\le T$，每客户端梯度复杂度为 $T\tau=\mathcal{O}\left(\frac{1}{n\epsilon^8}\right)$。由于$\tau\le T/n$，所需的最少通信轮次为 $T=\mathcal{O}\left(\frac{1}{\epsilon^4}\right)$。

• 对于部分参与，$\mathcal{O}\left(\left(\frac{n-P}{n-1}\cdot\frac{1}{PT}\right)^{1/4}\right)$是主导项，而且多次本地更新不会带来任何收敛收益。因此，对于 $\tau=\mathcal{O}(1)$，每梯度客户端复杂度和通信轮次数均为 $T\tau=\mathcal{O}\left(\frac{1}{P\epsilon^8}\right)$。不过，如果各客户端的数据来自相同分布（$\sigma_G=0$），那么每客户端梯度复杂度可恢复为 $\mathcal{O}\left(\frac{1}{P\epsilon^8}\right)$，通信轮次数为 $\mathcal{O}\left(\frac{1}{\epsilon^4}\right)$。

□

特殊情况

• 集中式确定性情形 ($\sigma_L=\sigma_G=0$，$\beta_G=1$，$\tau_{\text{eff}}=n=1$)：在此情形下 $A_w=B_w=1$，$C_w=D=0$。此外，$\mathcal{I}_1=G_{\mathbf{x}}\sqrt{\beta_L^2+1}$，$\mathcal{I}_2=0$。(48) 中的界简化为

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2\leq\mathcal{O}\left(\frac{\bar{\Delta}_{\widetilde{\Phi}}}{\gamma_x^sT}+\gamma_x^sL_fG_{\mathbf{x}}^2\left[(\beta_L^2+1)+(S-1)\sqrt{\beta_L^2+1}\right]+\frac{L_fR}{\gamma_y^sS}\right). \tag{56}$$

对于 $\beta_L=0$，(56) 给出了 Lin 等人 (2020a) 中的收敛结果。

• 单节点随机情形 ($\sigma_G=0$，$\beta_G=1$，$\tau_{\text{eff}}=n=1$)：在此情形下 $A_w=B_w=1$，$C_w=D=0$。此外，$\mathcal{I}_1=\sqrt{\sigma_L^2+(\beta_L^2+1)G_{\mathbf{x}}^2}$ $\mathcal{I}_2=\sigma_L^2$。(48) 中的界简化为

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2\leq&\,\Theta\left(\frac{\bar{\Delta}_{\widetilde{\Phi}}}{\gamma_x^sT}+\gamma_x^sL_f(\sigma_L^2+(\beta_L^2+1)G_{\mathbf{x}}^2)+\gamma_y^sL_f\sigma_L^2\right)\\&+\Theta\left(L_f\left[\gamma_x^sG_{\mathbf{x}}(S-1)\sqrt{\sigma_L^2+(\beta_L^2+1)G_{\mathbf{x}}^2}+\frac{R}{\gamma_y^sS}\right]\right).\end{aligned} \tag{57}$$

同样，对于 $\beta_L=0$，(57) 给出了 Lin 等人 (2020a) 中的收敛结果。

• 多个权重相同的($w_i=1/n$，$\forall\, i\in[n]$) 客户端、全客户端参与、客户端同步更新的随机情形($\tau_{\text{有效}}=1$)：在此情形下 $A_w=1$，$B_w=1$，$C_w=D=0$，0。（48）中的界可简化为

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2\leq&\,\Theta\left(\frac{\bar{\Delta}_{\widetilde{\Phi}}}{\gamma_x^sT}+\gamma_x^sL_f\left(G_{\mathbf{x}}^2+\frac{\sigma_L^2+\beta_L^2G_{\mathbf{x}}^2}{n}\right)\right)\\&+\Theta\left(\frac{\gamma_y^sL_f}{n}(\sigma_L^2+\sigma_G^2\beta_L^2)+\gamma_x^sL_fG_{\mathbf{x}}(S-1)\sqrt{G_{\mathbf{x}}^2+\frac{\sigma_L^2+\beta_L^2G_{\mathbf{x}}^2}{n}}+\frac{RL_f}{\gamma_y^sS}\right),\end{aligned} \tag{58}$$

请注意，与现有的同步更新算法分析不同，Woodworth 等人(2020年）；Yun 等人(2022年）；Sharma 等人(2022年），（58）中的界取决于客户端间异质性 σ_G^2。这是由于采用了更一般的噪声假设（假设2）。在现有研究中，β_L^2 假设为零，在这种情况下，（58）中的界也独立于 σ_G^2。更详细的说明请见附录A.2 。

• 多个权重相同的($w_i=1/n$，$\forall\, i\in[n]$) 客户端、全客户端参与、客户端更新次数多个但相同($\tau_i=\tau_{\text{有效}}=\tau$，$\forall\, i\in[n]$)。在此情形下 $A_w=B_w=1$，$C_w=\tau-1$，$D=(\tau-1)(\tau-1+\beta_L^2)$。（48）中的界随即简化为

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2\leq\Theta\left(\frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau\gamma_x^sT}+\tau\gamma_x^sL_f\left(G_{\mathbf{x}}^2+\frac{\sigma_L^2+\beta_L^2G_{\mathbf{x}}^2}{n\tau}\right)+\frac{\gamma_y^sL_f(\sigma_L^2+\beta_L^2\sigma_G^2)}{n}\right)\\&+\Theta\left(L_f\left[\tau\gamma_x^sG_{\mathbf{x}}(S-1)\sqrt{G_{\mathbf{x}}^2+\frac{\sigma_L^2+\beta_L^2G_{\mathbf{x}}^2}{n\tau}}+\frac{R}{\tau\gamma_y^sS}\right]+(\tau-1)\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[\sigma_L^2+(\tau-1+\beta_L^2)(G_{\mathbf{x}}^2+\sigma_G^2)\right]\right).\end{aligned} \tag{59}$$

对于 $\beta_L=\beta_G=0$，这一设定可简化为 Sharma 等人(2022年) 中考虑的情形。然而，正如前文所述，我们对局部更新误差的界更紧。

C.3 中间引理的证明

引理的证明 *ma C.1*。 利用（54）中的定义 $\bar{}\mathbf{x}^{(t)} = \operatorname{argmin}_{\mathbf{x}} \widetilde{\Phi}(\mathbf{x}) + L_f \left\|\mathbf{x} - \mathbf{x}^{(t)}\right\|^2$。另请注意，

$$\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t+1)}) \leq \widetilde{\Phi}(\bar{\mathbf{x}}^{(t)}) + L_f \left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t+1)}\right\|^2. \tag{60}$$

利用$\mathbf{x}$ 在（46）中的更新，

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t+1)}\right\|^2 = \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} + \tau_{\text{eff}}\gamma_x^s \sum_{i\in\mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t)}\right\|^2 \\
&= \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2 + \tau_{\text{eff}}^2[\gamma_x^s]^2\mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t)}\right\|^2 + 2\tau_{\text{eff}}\gamma_x^s\mathbb{E}\left\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}, \sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\rangle \\
&\leq \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2 + 2\tau_{\text{eff}}\gamma_x^s\mathbb{E}\left\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}, \nabla_x \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right\rangle + \tau_{\text{eff}}^2[\gamma_x^s]^2\mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t)}\right\|^2 \\
&\quad + \tau_{\text{eff}}\gamma_x^s\mathbb{E}\left[\frac{L_f}{2}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2 + \frac{2}{L_f}\left\|\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\left(\nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}) - \nabla_x f_i(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right)\right\|^2\right] \\
&\leq \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2 + \tau_{\text{eff}}^2[\gamma_x^s]^2\mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t)}\right\|^2 \qquad (61) \\
&\quad + \tau_{\text{eff}}\gamma_x^s\mathbb{E}\left[\frac{L_f}{2}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2 + 2L_f\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) + 2\left\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}, \nabla_x \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right\rangle\right], \qquad (62)
\end{aligned}$$

其中，(62)由 L_f-光滑性（假设1）和詹森不等式得出。根据(19)、(20)，我们可以将$\mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t')}} \tilde{w}_i \mathbf{g}^{(t')}{}_{\mathbf{x}\,i}\right\|^2$界定如下。

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t')}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t')}\right\|^2 &\leq \frac{n}{P}\sum_{i=1}^n \frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\left[\sigma_L^2 + \beta_L^2\mathbb{E}\left\|\nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)})\right\|^2\right] \\
&\quad + \frac{n}{P}\left(\frac{P-1}{n-1}\right)\mathbb{E}\left\|\sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 + \frac{n}{P}\frac{n-P}{n-1}\sum_{i=1}^n w_i^2\mathbb{E}\left\|\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \\
&\leq \frac{n}{P}\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + \frac{n(P-1)}{P(n-1)}G_{\mathbf{x}}^2 + \frac{n(n-P)}{P(n-1)}G_{\mathbf{x}}^2\sum_{i=1}^n w_i^2, \qquad (63)
\end{aligned}$$

其中，最后一个不等式利用了假设7。接下来，我们对（62）中的内积项进行界定。利用 L_f-光滑性 F（假设1）：

$$\begin{aligned}
&\mathbb{E}\left\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}, \nabla_x \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right\rangle \leq \mathbb{E}\left[\widetilde{F}(\bar{\mathbf{x}}^{(t)}, \mathbf{y}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) + \frac{L_f}{2}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2\right] \\
&\leq \mathbb{E}\left[\widetilde{\Phi}(\bar{\mathbf{x}}^{(t)}) + L_f\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2\right] - \mathbb{E}\widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \frac{L_f}{2}\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2 \\
&\leq \mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) + L_f\left\|\mathbf{x}^{(t)} - \mathbf{x}^{(t)}\right\|^2\right] - \mathbb{E}\widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \frac{L_f}{2}\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2 \qquad \text{(by definition of } \bar{\mathbf{x}}^{(t)})
\end{aligned}$$

$$\leq \mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \frac{L_f}{2}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2\right]. \tag{64}$$

将（62）和（64）中的界代入（60），得到

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t+1)})\right] &\leq \mathbb{E}\left[\widetilde{\Phi}(\bar{\mathbf{x}}^{(t)}) + L_f\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2\right] \\
&+ \tau_{\text{eff}}^2[\gamma_x^s]^2 L_f \frac{n}{P}\left[\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{P-1}{n-1} + \frac{n-P}{n-1}\sum_{i=1}^n w_i^2\right)\right] \\
&+ 2\tau_{\text{eff}}\gamma_x^s L_f^2 \sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) + 2\tau_{\text{eff}}\gamma_x^s L_f \mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right] - \frac{\tau_{\text{eff}}\gamma_x^s L_f^2}{2}\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\right\|^2 \\
&\leq \mathbb{E}\left[\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right] + \tau_{\text{eff}}^2[\gamma_x^s]^2 L_f \frac{n}{P}\left[\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{P-1}{n-1} + \frac{n-P}{n-1}\sum_{i=1}^n w_i^2\right)\right] \\
&+ 2\tau_{\text{eff}}\gamma_x^s\left\{L_f^2\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) + L_f\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right]\right\} - \frac{\tau_{\text{eff}}\gamma_x^s}{8}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2,
\end{aligned}$$

~其中我们使用 $\nabla\Phi_{1/2L_f}(\mathbf{x}) = 2L_f(\mathbf{x} - \bar{\mathbf{x}})$，来自式（54）。 □

引理C.2的证明。 我们使用式（45）中的个体迭代客户端更新方程。为了界定 $\Delta_{\mathbf{x}}^{(t,k)},_{\mathbf{y}}(i)$，我们首先界定 $\mathbf{x}$-误差 $\mathbb{E}\left\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\right\|^2$。从式（25）出发，利用假设 7，对于 $1 \leq k \leq \tau_i$，

$$\begin{aligned}
\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\mathbb{E}\left\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\right\|^2 &\leq \frac{[\eta_x^c]^2}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\left[\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2\left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + \left(\sum_{j=0}^{k-1} a_i^{(j)}(k)\right)\sum_{j=0}^{k-1} a_i^{(j)}(k) G_{\mathbf{x}}^2\right] \\
&\leq [\eta_x^c]^2\left[\sigma_L^2\left\|\mathbf{a}_{i,-1}\right\|_2^2 + G_{\mathbf{x}}^2\left(\left\|\mathbf{a}_{i,-1}\right\|_1^2 + \beta_L^2\left\|\mathbf{a}_{i,-1}\right\|_2^2\right)\right],
\end{aligned} \tag{65}$$

其中我们使用式（26）。接下来，我们利用式（$\mathbb{E}\left\|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\right\|^2$）中的界进行界定，得到

$$\begin{aligned}
\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\mathbb{E}\left\|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\right\|^2 &\leq [\eta_y^c]^2\sigma_L^2\left\|\mathbf{a}_{i,-1}\right\|_2^2 + 2[\eta_y^c]^2 L_f^2\left(\left\|\mathbf{a}_{i,-1}\right\|_1 + \beta_L^2\alpha\right)\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{y}}^{(t,k)}(i) \\
&+ 2[\eta_y^c]^2\left(\left\|\mathbf{a}_{i,-1}\right\|_1^2 + \beta_L^2\left\|\mathbf{a}_{i,-1}\right\|_2^2\right)\mathbb{E}\left\|\nabla_y f_i(\hat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right\|^2.
\end{aligned} \tag{66}$$

与式（30）相比，区别在于式（$\Delta_{\mathbf{y}}^{(t,k)}(i)$ 66）中出现了 $\Delta_{\mathbf{x}}^{(t,k)},_{\mathbf{y}}(i)$，而非如此。对式（66）中的智能体求加权和，得到

$$L_f^2\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{y}}^{(t,k)}(i) \leq 2[\eta_y^c]^2 L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\left\|\mathbf{a}_{i,-1}\right\|_2^2 + 2M_{\mathbf{a}_{-1}}\left(\beta_G^2\mathbb{E}\left\|\nabla_y\widetilde{F}(\hat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right\|^2 + \sigma_G^2\right)\right]. \tag{67}$$

其中，我们选择 η_y^c使得 $A_m \triangleq 2L_f^2[\eta_y^c]^2\max_i\|\boldsymbol{a}_i\|_1\left(\|\mathbf{a}_{i,-1}\|_1 + \beta_L^2\alpha\right) \leq \frac{1}{2}$，并定义 $M_{\mathbf{a}_{-1}} \triangleq \max_i\left(\|\mathbf{a}_{i,-1}\|_1^2 + \beta_L^2\|\mathbf{a}_{i,-1}\|_2^2\right.$

接下来，由 L_f-光滑性（假设1）和引理A.7 可得

$$\mathbb{E}\left\|\nabla_y\widetilde{F}\left(\hat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)}\right)\right\|^2 \leq 2L_f\mathbb{E}\left[\widetilde{\Phi}(\hat{\mathbf{x}}^{(s)}) - \widetilde{F}(\hat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right].$$

随后，结合（65）和（67），可得

$$L_f^2 \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \le 2\left([\eta_x^c]^2 + [\eta_y^c]^2\right) L_f^2 \sigma_L^2 \sum_{i=1}^{n} w_i \left\|\mathbf{a}_{i,-1}\right\|_2^2 + 4L_f^2 M_{\mathbf{a}_{-1}} \left([\eta_x^c]^2 G_{\mathbf{x}}^2 + [\eta_y^c]^2 \sigma_G^2\right) + 8[\eta_y^c]^2 L_f^3 M_{\mathbf{a}_{-1}} \beta_G^2 \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right].$$

至此，证明完成。 □

引理 *C.3* 的证明。我们定义 $\mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}) \in$ 最大值点集合$_{\mathbf{y}}\, \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y})$。于是，

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{y}^{(t+1)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})\right\|^2 \overset{(46)}{=} \mathbb{E}\left\|\mathbf{y}^{(t)} + \tau_{\text{eff}}\gamma_y^s \mathbf{g}_{\mathbf{y}}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})\right\|^2 \\
&= \mathbb{E}\left\|\mathbf{y}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})\right\|^2 + \tau_{\text{eff}}^2[\gamma_y^s]^2 \mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^{(t)}\right\|^2 + 2\tau_{\text{eff}}\gamma_y^s \mathbb{E}\left\langle \mathbf{y}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}), \sum_{i=1}^{n} w_i \mathbf{h}_{\mathbf{y},i}^{(t)} \right\rangle. \qquad (68)
\end{aligned}$$

$\mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^{(t)}\right\|^2$ 在（37）中有界。我们只需进一步给（$\mathbb{E}\left\|\sum_{i=1}^n w_i \mathbf{h}^{(t)}{}_{\mathbf{y},i}\right\|^2$）中的 37 作界。

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^{n} w_i \mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2 &\le \mathbb{E}\left\|\sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \left(\nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}) - \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)}) + \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right)\right\|^2 \\
&\le 2L_f^2 \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \Delta_{\mathbf{y}}^{(t,k)}(i) + 2\left\|\nabla_y \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right\|^2 \qquad \text{(Jensen's inequality)} \\
&\le 2L_f^2 \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \Delta_{\mathbf{y}}^{(t,k)}(i) + 4L_f \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right]. \qquad (69)
\end{aligned}$$

接下来，我们对（68）中的第三项作界。

$$\begin{aligned}
&\mathbb{E}\left\langle \mathbf{y}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}), \sum_{i=1}^{n} w_i \mathbf{h}_{\mathbf{y},i}^{(t)} \right\rangle = \mathbb{E}\left\langle \mathbf{y}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}), \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}) \right\rangle \\
&= \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \mathbb{E}\left[\left\langle \mathbf{y}^{(t)} - \mathbf{y}_i^{(t,k)}, \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}) \right\rangle + \left\langle \mathbf{y}_i^{(t,k)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}), \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}) \right\rangle\right] \\
&\le \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \mathbb{E}\Bigg[f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)}) - f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}) + \frac{L_f}{2}\left\|\mathbf{y}^{(t)} - \mathbf{y}_i^{(t,k)}\right\|^2 \qquad (L_f\text{-smoothness}) \\
&\qquad + f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}) - f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}))\Bigg] \qquad \text{(Concavity in } \mathbf{y}) \\
&= \frac{L_f}{2} \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \Delta_{\mathbf{y}}^{(t,k)}(i) - \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right]. \qquad (70)
\end{aligned}$$

将 (37)、(69)、(70) 代入 (68)，得到

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{y}^{(t+1)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})\right\|^2 \\
&\le \mathbb{E}\left\|\mathbf{y}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})\right\|^2 + \tau_{\text{eff}}^2[\gamma_y^s]^2 \left[\frac{\sigma_L^2 n}{P} \sum_{i=1}^{n} \frac{w_i^2 \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \frac{2\sigma_G^2 n}{P}\left(\frac{n-P}{n-1} \max_i w_i + \beta_L^2 \max_i \frac{w_i \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]
\end{aligned}$$

$$- 2\tau_{\text{eff}}\gamma_y^s\left(1 - 2\tau_{\text{eff}}\gamma_y^s L_f \frac{n}{P}\left[\frac{P-1}{n-1} + \beta_G^2\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2 \max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]\right)\mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right]$$
$$+ \tau_{\text{eff}}\gamma_y^s L_f\left[1 + 2\tau_{\text{eff}}\gamma_y^s L_f \frac{n}{P}\left(\frac{P-1}{n-1} + \frac{n-P}{n-1}\max_i w_i + \beta_L^2 \max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\right]\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i), \tag{71}$$

由于 $\Delta_{\mathbf{y}}^{(t,k)} \le \Delta_{\mathbf{x}\,\mathbf{y}}^{(t,k)}$。利用引理 $\Delta_{\mathbf{x}}^{(t,k)}$、$\mathbf{y}$ 中的界 C.2，

$$\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \le 2\left([\eta_x^c]^2 + [\eta_y^c]^2\right)\sigma_L^2\sum_{i=1}^n w_i\|\mathbf{a}_{i,-1}\|_2^2 + 4M_{\mathbf{a}_{-1}}\left([\eta_x^c]^2 G_{\mathbf{x}}^2 + [\eta_y^c]^2\sigma_G^2\right)$$
$$+ 8[\eta_y^c]^2 L_f M_{\mathbf{a}_{-1}}\beta_G^2\mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right]. \tag{72}$$

将 (72) 代入 (71)，并根据 γ_y^s、η_y^c 的选择化简各项，得到

$$\mathbb{E}\left\|\mathbf{y}^{(t+1)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})\right\|^2$$
$$\le \mathbb{E}\left\|\mathbf{y}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})\right\|^2 + \tau_{\text{eff}}^2[\gamma_y^s]^2\left[\frac{\sigma_L^2 n}{P}\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \frac{2\sigma_G^2 n}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]$$
$$- \tau_{\text{eff}}\gamma_y^s\mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right] + 4\tau_{\text{eff}}\gamma_y^s L_f([\eta_x^c]^2 + [\eta_y^c]^2)\left[\sigma_L^2\sum_{i=1}^n w_i\|\mathbf{a}_{i,-1}\|_2^2 + 2M_{\mathbf{a}_{-1}}(G_{\mathbf{x}}^2 + \sigma_G^2)\right].$$

其中 γ_y^s、η_y^c 满足

$$2\tau_{\text{eff}}\gamma_y^s L_f\frac{n}{P}\left(\frac{P-1}{n-1} + \frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right) \le 1,$$
$$2\tau_{\text{eff}}\gamma_y^s L_f\frac{n}{P}\left[\frac{P-1}{n-1} + \beta_G^2\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right] \le \frac{1}{4},$$
$$2L_f\frac{n}{P}\left[8[\eta_y^c]^2 M_{\mathbf{a}_{-1}} L_f\beta_G^2\right] \le \frac{1}{4}$$

那么，$\mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right]$ 的系数可以界定为 $-\tau_{\text{eff}}\gamma_y^s$。因此，通过重新排列各项并对 t 求和，可得所需结果。

$$\frac{1}{S}\sum_{t=sS}^{(s+1)S-1}\mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right]$$
$$\le \frac{\mathbb{E}\left\|\mathbf{y}^{sS} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})\right\|^2}{\tau_{\text{eff}}\gamma_y^s S} + \tau_{\text{eff}}\gamma_y^s\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + 2\sigma_G^2\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]$$
$$+ 4L_f([\eta_x^c]^2 + [\eta_y^c]^2)\left[\sigma_L^2\sum_{i=1}^n w_i\|\mathbf{a}_{i,-1}\|_2^2 + 2M_{\mathbf{a}_{-1}}(G_{\mathbf{x}}^2 + \sigma_G^2)\right].$$

□

引理 C.4 的证明。设 $t = sS,\ sS+1,\ \ldots,\ (s+1)S-1$，其中 k 是一个正整数。设 $\widehat{\mathbf{x}}^{(s)}$ 是算法 $\mathbf{y}$-更新对应的最新快照迭代点，其中算法为 1-Fed-Norm-SGDA+。则

$$\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^*(\mathbf{x}^{(t)})) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})) + \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)}) + \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right] \\
&\leq \mathbb{E}\left[\widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^*(\mathbf{x}^{(t)})) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^*(\mathbf{x}^{(t)}))\right] + \mathbb{E}\left[\widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right] + G_{\mathbf{x}}\mathbb{E}\left\|\mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(s)}\right\| \\
&\leq 2G_{\mathbf{x}}\mathbb{E}\left\|\mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(s)}\right\| + \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right]. \qquad (73)
\end{aligned}
$$

其中，$\mathbf{y}^*(\cdot) \in$ 最大值点集合$_{\mathbf{y}}$ $F(\cdot,\ \mathbf{y})$，而（73）由 $G_{\mathbf{x}}$-Lipschitz 连续性可得 $F(\cdot,\ \mathbf{y})$（假设 7）。接下来，我们看到

$$
\begin{aligned}
\mathbb{E}\left\|\widehat{\mathbf{x}}^{(s)} - \mathbf{x}^{(t)}\right\| &\leq \sqrt{\mathbb{E}\left\|\widehat{\mathbf{x}}^{(s)} - \mathbf{x}^{(t)}\right\|^2} && \text{（詹森不等式）} \\
&\overset{(46)}{=} \sqrt{\mathbb{E}\left\|\tau_{\text{eff}}\gamma_x^s \sum_{t'=sS}^{t-1} \sum_{i\in\mathcal{C}^{(t')}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t')}\right\|^2} \\
&\leq \tau_{\text{eff}}\gamma_x^s \sqrt{(S-1)\sum_{t'=sS}^{t-1} \mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t')}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t')}\right\|^2} \\
&\leq \tau_{\text{eff}}\gamma_x^s (S-1)\sqrt{\frac{n}{P}}\sqrt{\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2) + G_{\mathbf{x}}^2\left(\frac{P-1}{n-1} + \frac{n-P}{n-1}\sum_{i=1}^n w_i^2\right)}. \\
& && \text{(from (63))}
\end{aligned}
$$

在（73）中使用此界，并对 t 求和，可得

$$
\begin{aligned}
\frac{1}{S}\sum_{t=sS}^{(s+1)S-1} \mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right] &\leq \frac{1}{S}\sum_{t=sS}^{(s+1)S-1} \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right] \\
&+ 2\tau_{\text{eff}}\gamma_x^s G_{\mathbf{x}}(S-1)\sqrt{\frac{n}{P}}\sqrt{\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2) + G_{\mathbf{x}}^2\left(\frac{P-1}{n-1} + \frac{n-P}{n-1}\sum_{i=1}^n w_i^2\right)}.
\end{aligned}
$$

最后，对 $s = 0$ 到 $T/S - 1$ 求和可得结果。 □

C.4 将结果扩展到非凸单点凹（NC-1PC）函数（定理 3）

仔细重新审视定理 2 的证明后，我们注意到引理 C.1 和引理 C.2 并不依赖于凹性假设。引理 C.3 确实使用了局部函数的凹性 $\{f_i\}$。然而，这仅用于推导（70）。此外，这只要求局部函数在一个全局点 $\mathbf{y}^*$（$\widehat{\mathbf{x}}^{(s)}$）处满足凹性。因此，如备注 8 中所述，即使对于 NC-1PC 函数，该结论仍然成立。这本身就是一个独立的结果，因为我们将局部随机梯度方法在凸最小化（凹最大化）问题上的现有收敛结果，扩展到了更一般的单点凸最小化（或单点凸最大化）问题。因此，我们在此针对更一般的情形重新陈述该结果。

引理*C.5*（单点凹最大化的局部随机梯度更新）。假设局部损失函数 $\{f_i\}$ 满足假设 *1*、*2*、*3*、*7*。假设对于所有 *x*，所有 f_i 都在同一个全局最小值点 *y*(x)* 处满足假设 *8*，并且对于所有 t，$\left\|\mathbf{y}^{(t)}\right\|^2 \leq R$。如果我们运行算法 *1-Fed-Norm-SGDA+*，并采用与引理 *C.3* 中相同的客户端和服务器步长条件 η_y^c、γ_y^s，那么算法 *1-Fed-Norm-SGDA+* 生成的迭代序列也满足引理 *C.3* 中的界。

接下来，引理 C.4 无论是否满足凹性，也都成立。因此，定理2中针对非凸-凹极小极大问题得出的收敛结果适用于范围更广的函数类。我们简要重述修改后的定理。

定理。假设局部损失函数 $\{f_i\}$ 满足假设 *1*、*2*、*3*、*7*。假设对于所有 *x*，所有 f_i 都满足假设 *8*，且具有共同的全局极小点 *y*(x)*，并且 $\left\|y^{(t)}\right\|^2 \leq R$ 对所有 t 均成立。若我们运行算法 *1-Fed-Norm-SGDA+*，并采用与定理 *4* 中相同的客户端和服务器步长 η_y^c、γ_y^s，则算法 *1-Fed-Norm-SGDA+*生成的迭代序列也满足定理 *4* 中的界。

备注 9。 同样，选择与原始全局目标函数相同的客户端权重 $\{w_i\}$，我们可证明原始目标函数 F 的收敛性。

D 补充实验

客户端之间通过以太网连接来传递参数及相关信息。我们的算法使用 PyTorch 1.0.0 和 Python 3.6.3 中的并行训练工具实现。

在鲁棒神经网络训练和公平分类实验中，所有算法均使用批量大小为 32。只有在 Momentum 本地SGDA(+) 中使用动量参数0.9。

鲁棒神经网络训练 这里我们进一步探究 Fed-Norm-SGDA 在+ 鲁棒神经网络训练问题上的性能。我们使用 VGG-11 模型对 CIFAR10 数据集进行分类。在图6中，我们展示了客户端之间数据异质性增加所产生的影响；图7 则展示了在联邦极小极大问题中使用多个客户端的优势。随着 k倍的增加 n，我们观察到，达到目标测试准确率（此处为 70%）所需的通信轮次几乎减少了 k倍。

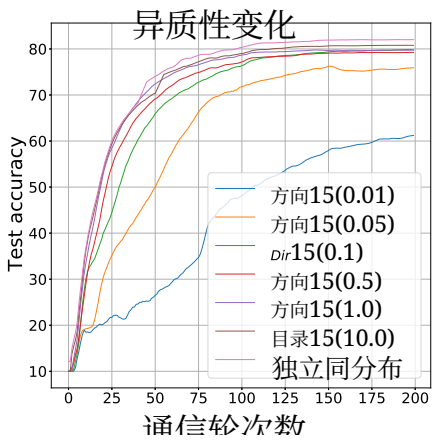

图 6：客户端间数据异质性（由 α量化）对 Fed-Norm-SGDA+性能的影响。

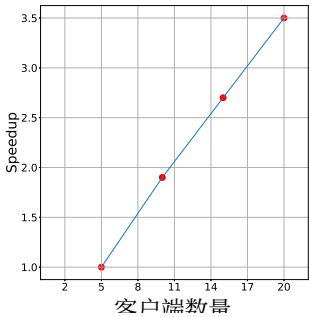

图 7：增加　客户端集合对鲁棒神经网络训练任务中 Fed-Norm-SGDA+ 性能的影响　.

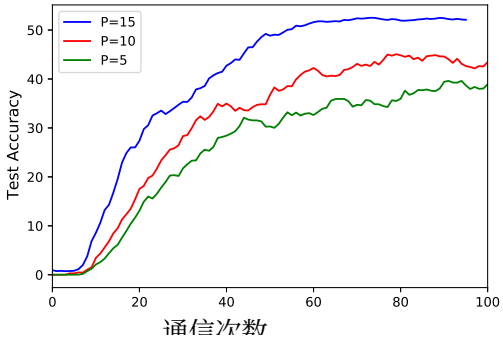

图8：部分　　客户端参与对 Fed-Norm-SGDA 在公平图像分类任务中的性能影响　　k.

公平分类 我们还展示了客户端部分参与对公平分类问题的影响。图8 补充了正文中的图8，评估 VGG11 模型在 CIFAR10 数据集上的公平性。我们绘制了模型在最差分布上的测试准确率。随着参与客户端数量的增加，性能持续提升。